UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

USING MACHINE LEARNING APPLICATIONS AND HREFV2 TO ENHANCE HAIL PREDICTION FOR OPERATIONS

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
MASTER OF SCIENCE IN METEOROLOGY

By

AMANDA L. BURKE
Norman, Oklahoma
2019

USING MACHINE LEARNING APPLICATIONS AND HREFV2 TO ENHANCE HAIL PREDICTION FOR OPERATIONS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. Amy McGovern (Chair)

Dr. Nathan Snook

Dr. Cameron Homeyer

Dr. Jason Furtado

Acknowledgments

This thesis would not have been possible without the guidance from my advisors, Amy McGovern and Nathan Snook, and suggestions from my entire committee, including Cameron Homeyer and Jason Furtado. The feedback and advice I received from my committee was priceless and crucial for the development of my thesis. I especially want to thank David John Gagne II for the continued support and answers to the multitude of questions that came up throughout my thesis. I also wish to thank Steven Weiss, Israel Jirak, Adam Clark and others for giving me the opportunity to participate in the 2018/2019 HWT SFEs. These experiments not only provided feedback that partially shaped the direction of this thesis, but also expanded my passion for weather and forecasting! Regarding data support, I want to thank Chris Cook and Kevin Thomas for providing computational and archival support, especially during the SFEs.

Last, but by far not least, I cannot thank my family and friends enough for the continued support and encouragement throughout my master's. From both New York and Oklahoma, my support system really helped me get through the tough times and I am very appreciative. I especially want to thank my Parents, Aunts, Emily Thurnherr, Morgan Seeley, Joel McAuliffe, Briana Lynch, Melanie Schroers, Dakota Hahn and many others!

This thesis was primarily supported by the Joint Technology Transfer Initiative (JTTI) Grant NA16OAR4590239 provided by NOAA. Computing was executed on the University of Texas Advanced Computing Center (TACC) Stampede supercomputer.

Table of Contents

List Of Tables

List Of Figures

Abstract

In this thesis, I demonstrate how hail prediction can be improved through post-processing numerical weather prediction (NWP) data from the new High-Resolution Ensemble Forecast system version 2 (HREFv2) with machine learning (ML) models. Multiple operational models and ensembles currently predict hail, however ML models obtain optimal predictions that are computationally efficient and do not explicitly predict hail. Additionally, ML models can synthesize multiple datasets, which helps address the abundance of ensemble data that can lead to cognitive forecaster overload. The observational dataset used to train all of the ML models is the maximum expected size of hail (MESH), a Multi-Radar Multi-Sensor (MRMS) product.

Random forest models, along with a calibration step using isotonic regression, produce severe hail predictions over the contiguous United States (CONUS) with probability magnitudes similar to operational forecasts. Calibrating the ML-based predictions toward familiar forecaster output combines higher skill from the ML predictions with forecaster trust. Verification suggests that the calibrated ML models maintain spatially similar regions of severe hail while providing similar hail probability magnitudes, as compared to the SPC day 1 hail outlook and practically perfect data. The ML output calibrated towards the local storm reports exhibited better or similar skill to the uncalibrated predictions, while decreasing model bias.

In addition to CONUS hail forecasting, I investigate the superiority of localized forecasting by regionally training ML models in high-impact hail environments. Objective, subjective, and statistical verification indicates the southern plains ML model produces superior hail forecasts over the CONUS-trained ML model. Moisture fields are emphasized in the southern plains region, with additional instability and shear variables displaying importance, similar to the CONUS-trained model. A larger training dataset spanning multiple years, as well as exploration into different regional weighting functions, could improve the other locally trained ML models.

Chapter 1

Introduction

Hail is a high-impact severe weather hazard, annually causing in excess of $1 billion of property damage and $1 billion of crop damage (Jewell and Brimelow, 2009). Often, isolated hail events, especially those impacting large-scale urban areas, are particularly damaging. For example, a single hailstorm during the afternoon rush-hour in the Denver, CO metropolitan area on 8 May 2017 resulted in $2.3 billion of insurance claims (Svaldi, 2018). The economic impacts of severe hail underscore the need for accurate and timely predictions, which allow individuals and businesses to take action towards mitigating risk to their property and safety. However, accurate predictions of hail remain a challenge given the rapid evolution of hail-producing convective storms. Additionally, challenges arise from the uncertainties of, and limitations in, atmospheric observation data needed to properly resolve the small-scale environment.

To help address resolution issues, convection allowing models (CAMs) were developed that are able to partially resolve convection. Multiple methods have focused on explicit hail prediction using storm-scale models (e.g., Adams-Selin and Ziegler, 2016; Snook et al., 2016; Labriola et al., 2017, 2019). However, explicitly predicting hail is limited by physical assumptions as well as sensitivities in model initial conditions, boundary conditions, and other parameter choices. Additionally, the abundance of hail forecasts produced from storm-scale models can quickly lead to forecaster overload. Recently, studies have focused on using machine learning (ML) to synthesize large amounts of atmospheric data to produce skillful forecast products, without the need for explicit prediction (e.g., Gagne, 2016; Gagne et al., 2017; McGovern et al., 2017; Lagerquist et al., 2017; Herman and Schumacher, 2018a,b). Instead, ML models map a set of inputs to a given output by optimizing the model's structure, such that the differences between the ML predictions and the output observations, or ”ground truth” are minimized. Using these learned structures, ML models are able to make predictions on new sets of model data with relatively minimal computational expense, another major advantage compared to running a gridded NWP model.

Despite the advantages of using ML models for hail prediction, during the 2014 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE), forecasters expressed distrust in automated guidance when proper training or knowledge of a new product's skill and reliability are not provided (Karstens et al., 2015). To address forecast trust, the first part of this thesis explores calibration of real-time ML output to resemble existing operational forecasts. Specifically, calibrated predictions over the Contiguous United States (CONUS) are aimed to mirror Storm Prediction Center (SPC) forecasts, to increase trust of automated guidance in an operational setting.

In addition to full CONUS hail forecasting, the second part of this thesis explores regional models for localized environments resulting in large hail. We hypothesize that regional ML models tested over their respective sector will produce hail forecasts with superior performance than forecasts created by a CONUS-trained model. If a first iteration of regionalization proves feasible, further research could produce highly skillful, tailored forecasts for areas experiencing prevalent severe hail.

Within this thesis, previous studies detailing historical and present-day hail forecasting techniques, ML applications for hail forecasting, and regionalization methods across multiple meteorological disciplines are described in chapter 2. Chapter 3 explores pre-processing procedures for data input to the statistical ML models. The framework for creating and evaluating the ML hail forecasts is described in chapter 4.

Hail forecasting results on the CONUS and regional scales are detailed in chapters 5 and 6, respectively. Finally, a summary of results from both hail forecasting regimes and future changes to improve the hail forecasts are described in chapter 7.

Chapter 2

Background

Examination into hail formation, including the environmental setup responsible for severe hail production, is detailed in section 2.1. Section 2.2 describes the challenges of explicit hail prediction and the use of storm-scale datasets to help address some of the issues with explicit severe weather forecasting. Finally, section 2.3 discusses the advantages associated with ML models for hail prediction, and regionalization background for further analysis with ML-based forecasting.

2.1 Hail Formation

Multiple studies have investigated the features needed for hail production, including the dependence of severe hail growth on strong updraft development. Dennis and Kumjian (2017) found that optimal hail growth results from sustained balance between updraft speed and hailstone terminal fall velocity. In discriminating between the convective development of supercells, which largely result in severe hail, and quasi-linear convective systems, Thompson et al. (2012) established that available buoyancy is a major contributor. McCaul and Weisman (2001) also studied the affects of buoyancy on simulated convective storms, with the addition of various shear profiles, finding that updraft intensity is related to changes in said environmental fields. In small convective available potential energy (CAPE) regimes with moderate shear, concentrations of low-level buoyancy are crucial for updraft development. Greater concentrations of low-level buoyancy and shear are also important, although not as critical as in small CAPE settings, for peak updraft development in large CAPE environments. Additionally, Weisman and Klemp (1982) discovered that increasing shear under fixed CAPE environments decreases updraft intensity and increases tilt. Finally, increases in midlevel storm-relative winds were identified for mesocyclone development and persistence within supercells, which again result in large hail (Brooks et al., 1994).

High-speed, especially rotating, updrafts are important for hail formation, however multiple other factors affect optimal hail development. In a study comparing supercell precipitation types, lower midlevel relative humidity values resulted in small hailstone development, despite no substantial updraft intensity differences between types (Grant and van den Heever, 2014). In fact, hail growth is most efficient when embryos develop outside the main updraft core. Embryos that grow to be millimeters in diameter before injection to the main growth region are more likely to achieve the needed balance between updraft speed and particle fall speed (e.g., Heymsfield, 1982; Foote, 1984). Heymsfield (1982) found that frozen drops are the most likely embryo type found within convective storms, as they can develop rapidly into hailstones. Starting as small frozen aggregates, the particles melt after being injected into the main growth region, creating frozen drops with large diameters that rapid develop into hail. The main growth region occurs in high-moisture areas with temperatures between -10 and -25 °C, where supercooled liquid is able to accrete onto particles (e.g., Nelson, 1983; Foote, 1984). However, if the number of hail embryos exceeds the available supercooled liquid, competition can limit possible hail growth (Heymsfield, 1982).

In addition to sufficient updraft speeds and available moisture, the broadness of an updraft impacts hail development, especially severe hail. Multiple studies have highlighted the importance of horizontal trajectories that allow for severe hail development within large updrafts (e.g., Nelson, 1983; Foote, 1984; Miller et al., 1990; Dennis and Kumjian, 2017). As mentioned above, Weisman and Klemp (1982) found that increasing shear under fixed CAPE environments led to greater updraft tilt and decreased updraft intensity. A tilted, or broad, updraft allows the hail embryos to stay suspended in areas of supercooled liquid longer, leading to larger growth. Similarly, Dennis and Kumjian (2017) found that increases in deep-layer shear (east-west direction) led to significantly larger hail size development. On the occurrence of large hail, the southern and central plains experience the greatest probabilities of severe hail climatologically.[1] Cintineo et al. (2012) identified the Great Plains area as a maxima for annual severe hail, with a secondary maxima in the Southeastern United States. Finally, a monthly occurrence model associated large-scale environments with incidences of large hail over the contiguous United States, identifying similar areas of hail maxima as the previous climatologies (Allen et al., 2015).

2.2 Explicit Hail Prediction

To produce skillful hail forecasts through explicit hail prediction, numerical weather prediction (NWP) models must accurately predict the development of convective storms, as well as produce reasonably accurate representations of hail within the model's microphysical scheme (Labriola et al., 2017). At these small scales, model forecast errors can lead to large uncertainties in the timing and location of hail-producing convective storms, decreasing skill.

For hail prediction on longer timescales (up to 48 hours), most methods rely on approximating environmental data at the convective scale (e.g., Johns and Doswell, 1992). Environmental fields, such as temperature, dew point temperature, and CAPE, can be extracted from atmospheric soundings launched twice daily across the United

[1] https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=sigHail

States. However, the spatial and temporal coverage of atmospheric soundings are generally insufficient to provide accurate initial conditions for explicit prediction of storms on the convective scale. Previous work examined approximating the maximum diameter of hail from sounding data by adjusting for diurnal temperature changes after the 1200 UTC launch (Moore and Pino, 1990). However, even the adjusted sounding data may not be completely representative of the conditions leading to hail-producing storm updrafts.

Where limitations in scale cause uncertainties in local storm characteristics, CAMs have shown skill in predicting convective morphologies (e.g., Weisman et al., 2008). Although CAMs are able to partially resolve convective-scale storms and relevant attributes of the local storm environment, errors can still persist in the timing and placement of convection. In recent years, CAMs have been employed in the National Oceanic and Atmospheric Administration's (NOAA's) HWT SFE. For example, during the 2010 HWT SFE, operational forecasters subjectively indicated the CAM guidance improved upon forecasting convection, compared to traditional convective-parameterizing schemes (Clark et al., 2012). Also, Gallo et al. (2017) noted that CAMs played an important role in reliable short-term forecasts, especially hourly forecasts, during the 2015 HWT SFE.

Generally, CAMs cannot skillfully predict severe hazards (such as severe wind, severe hail, or tornadoes) within individual storms on time-scales of longer than a couple of hours. Even when high-resolution, small-scale environmental data is available, CAMs still face the problem of rapid storm development and rapidly growing model errors. To address this issue, severe weather forecasts derived from CAMs often rely upon hourly predictions of proxy variables, such as updraft speed or column-integrated graupel. In this way, the hourly maximum of a given atmospheric field is output at each grid point to approximate storm intensity and location, as well as the potential for severe weather threats, such as wind, hail, and tornadoes. Hourly maxima have been found skillful as guidance when forecasting severe weather, with minimal calibration needed (Sobash et al., 2011). In addition, Kain et al. (2010) found that hourly maximum values provide skill for severe weather forecasting, particularly in determining hail threats in nonsupercellular storms. Explicit prediction of hail using ensembles of storm-scale models has been proven feasible on 0-3 hour time scales (Snook et al., 2016; Labriola et al., 2017), however they are not produced real-time. For day-ahead forecasts (12-36 hr lead time), CAM ensemble forecasts have shown improved skill compared to individual deterministic CAM forecasts (Loken et al., 2017). The skill CAM data provide in predicting severe weather threats, along with the high spatio-temporal resolution of said data, underscores the usefulness of CAMs as input to more advanced hail prediction models.

2.3 Machine Learning for Hail Forecasting

One such advanced method predicts probabilities and sizes of hail using ML models over the contiguous United States (CONUS). Previous studies have demonstrated an increase in hail forecasting performance associated with ML-based predictions, using storm-scale ensembles, over direct prediction of hail from NWP model output or proxy variables (Gagne, 2016; Gagne et al., 2017). Gagne et al. (2017) found that the ML-based method produced skillful hail forecasts when tested with differing model configurations. Although the ML model objectively increased forecasting performance over the CONUS, forecaster trust is an important aspect of an operationally-focused product. Subjective commentary from the 2018 HWT SFE indicated that forecasters are less likely to trust model guidance with unfamiliar or dissimilar output compared to human-produced forecasts.[2] To increase trust, the full CONUS-trained model output is calibrated, as mentioned in chapter 1. In addition to increases in forecaster trust, calibration has shown increases in skill and reliability of probabilistic forecasts (e.g., Raftery et al., 2005; Hagedorn et al., 2008; Hamill et al., 2008).

While previous ML methods for hail forecasting have proven successful over the full CONUS (Gagne et al., 2017), local environmental influences may impact forecasting performance such that regional models are superior in areas prone to high-impact hail. Regionalization studies appear in multiple meteorological disciplines, including heavy precipitation (e.g., Yang and Smith, 2006; Charba and Samplatsky, 2011; Zhang et al., 2016), climate (e.g., Charba and Samplatsky, 2011; Netzel and Stepinski, 2016; Lai and Dzombak, 2019), and tornadic environment studies (e.g., Brown, 2002; Brooks et al., 2003; Trapp and Brooks, 2013; Moore, 2018). Regarding hail research, regional studies have examined the spatial and temporal distributions of large hail within non-tornadic thunderstorms (Kelly et al., 1985), areal hail risk with regards to insurance costs (Changnon, 1999), and historical hail occurrence on both the national and regional scales (Changnon and Changnon, 2000). More recently, localized hail studies have focused on severe hailstorm development using satellite data (Cecil and Blankenship, 2012), identification of hail maxima from comparisons of maximum expected size of hail (MESH) data and NOAA *Storm Data* reports (Cintineo et al., 2012), and also extreme value theory applied to spatial distributions of observed hail sizes within the CONUS (Allen et al., 2017). Focusing on the regional scale for ML-based hail forecasting could highlight the local dynamical and physical attributes responsible for hail production that may be broadly ignored in a full CONUS analysis. Therefore, leveraging areas previously identified for severe hail occurrence, based on differing environmental

[2] Forecasters from 2018 HWT SFE, Week 3 (May 14-18)

characteristics leading to hail formation, with statistical ML methods could develop a model with superior performance in areas experiencing greater hail incidences.

Chapter 3

Data

Chapter 3 describes the input data to the ML models (Section 3.1), and a pre-processing step applied to said data (Section 3.2).

3.1 Data Sources

For this research, all of the ML-based hail forecasts are generated from the High Resolution Ensemble Forecast system version 2 (HREFv2) (Jirak et al., 2018). Starting in April 2017, the SPC began running the HREFv2, an ensemble based off the Storm Scale Ensemble of Opportunity (SSEO). The SSEO, a ”poor man’s ensemble” (Ebert, 2001), is computationally efficient and made up of operational CAMs produced by NOAA (Jirak et al., 2012). The success of the SSEO (Clark et al., 2016) in probabilistic severe weather prediction brought attention to the HREFv2 dataset for use in skillful weather prediction.

Developed by National Centers for Environmental Prediction (NCEP)/ Environmental Modeling Center (EMC), the HREFv2 is run daily by NCEP Central Operations (NCO)[1], and consists of an eight-member ensemble with time-lagged members initialized at 0000 UTC, 0600 UTC, 1200 UTC, and 1800 UTC. Only the members initialized at 0000 UTC and 1200 UTC were provided by the SPC for this thesis. The HREFv2 is a diverse ensemble that includes multiple initial conditions and microphysics

[1] `https://www.spc.noaa.gov/exper/href/#`

schemes. The microphysical schemes include the Weather Research and Forecasting (WRF) single-moment 6-Class (Hong et al., 2010) and Ferrier-Aligo (Aligo et al., 2014) schemes. The planetary boundary layer (PBL) schemes consist of the Yonsei University (YSU) (Hong et al., 2006) and local Mellor-Yamada (MYJ) schemes (Janjić, 1990, 1994). The members output nearly identical horizontal grid spacing, approximately 3 km. The number of vertical levels at which data are produced differs between ensemble members; the four HiresW members, including the time-lagged members, have 50 vertical levels, the two NSSL-ARW members have 40 vertical levels, and the two NAM-NEST members have 60 vertical levels (Table 3.1). Forecast products are valid 1200 UTC to 1200 UTC the next day, the same period as a day 1 SPC convective outlook. A detailed description of the HREFv2 members is provided in Table 3.1.

Table 3.1: Configuration of the eight member High Resolution Ensemble Forecast system version 2 (HREFv2), similar to www.spc.noaa.gov/exper/href/

Members	Dynamical Core	Initializations	PBL	Microphysics	Vertical Levels	Grid Spacing
HiresW ARW	WRF-ARW	0000 UTC	YSU	WSM6	50	3.2km
HiresW ARW	WRF-ARW	1200 UTC	YSU	WSM6	50	3.2km
HiresW NMMB	NMMB	0000 UTC	MYJ	Ferrier-Aligo	50	3.2km
HiresW NMMB	NMMB	1200 UTC	MYJ	Ferrier-Aligo	50	3.2km
HiresW NSSL	NSSL-WRF	0000 UTC	MYJ	WSM6	40	3.2km
HiresW NSSL	NSSL-WRF	1200 UTC	MYJ	WSM6	40	3.2km
NAM Nest	NMMB	0000 UTC	MYJ	Ferrier-Aligo	60	3km
NAM Nest	NMMB	1200 UTC	MYJ	Ferrier-Aligo	60	3km

For training purposes, attributes of MESH (Witt et al., 1998) derived from NOAA/ NSSL Multi-Radar Multi-Sensor (MRMS) radar data (Zhang et al., 2011) are used by the ML models as observations. Because MESH outputs have exhibited greater skill for values exceeding 19 mm (Wilson et al., 2009), only MESH values greater than 19mm are considered. Despite known biases, such as over-prediction of higher values (e.g., Wilson et al., 2009; Cintineo et al., 2012; Ortega, 2018; Murillo and Homeyer, 2019), MESH was chosen over the local storm reports (LSRs) as the observational dataset for training the ML models because of the known population and size biases with LSRs (e.g., Schaefer et al., 2004; Cintineo et al., 2012). Also, multi-radar MESH allows for over-sampling of storms to provide a more accurate measurement of the expected hailstone diameter, and has performed well compared to single radar MESH observations (Ortega et al., 2005, 2006). Finally, Melick et al. (2014) found that MESH values, filtered using a Gaussian smoother, observed more hail objects than filtered LSRs and act as a useful independent dataset in low population areas.

3.2 Object-tracking

Before the HREFv2 and MESH data are input into the ML models, the two datasets are pre-processed using an object-tracking algorithm to address the relative rarity of hail (and severe weather in general) at any given location. The object tracking method and ML models evaluated for hail prediction are based upon those used in Gagne et al. (2017). Figure 3.1 includes the different steps of the object tracking method. This algorithm identifies potential storm objects where a chosen variable field exceeds a user-specified threshold for storm identification. For the HREFv2 dataset, storm objects were identified using maximum hourly upwards vertical velocity (MAXUVV) values greater than 8 ms^{-1} (based on updraft speed resulting in near-pea sized hail[2]), rather than column total graupel of at least 3 kg m^{-2} from Gagne et al. (2017). The selection of updraft speed rather than updraft helicity or column total graupel, and the use of a relatively low minimum threshold value, were designed to capture all possible hail storms rather than only high-end supercells. Although supercells are typically

[2] https://www.nssl.noaa.gov/education/svrwx101/hail/

responsible for the most severe hail events, marginal hail is important to the public, the insurance industry, and agriculture. In the observations, potential MESH storm objects are generated for values exceeding 19 mm, differing from the 12 mm threshold used in Gagne et al. (2017), for reasons outlined above. Using MESH >19mm could potentially impact the number of marginal hail events captured, however the main focus of this research is to provide severe hail predictions to the SPC. Although all types of potential hail storms are considered by the HREFv2 data, severe hailstorms are more likely to be evaluated based on the MESH threshold. Although threshold parameters identify potential storm objects, areal limits are needed to determine the size of said objects. The objects are created using a watershed method developed by Lakshmanan et al. (2009), where object areas must be between 36 km^2 (12 grid points) and 300 km^2 (100 grid points).

After identification, the potential storm objects are matched in time and space to create storm tracks. Specifically, at each hour the storm objects less than 24 km apart are combined together. This distance allows for uncertainties in object placement, while reducing the risk of track misrepresentation. Using the identified model tracks, HREFv2 environmental and storm variables are evaluated by multiple statistical functions, to be input into the ML models. Table 3.2 includes a full list of the input HREFv2 variables. Storm variables are defined as those directly impacting the storm, such as CAPE, convective initiation (CIN), storm relative helicity (SRH), etc. Environmental variables affect the near-storm fields, such as temperature, dew point temperature, geopotential height, etc. The environmental variables are extracted from the previous forecast hour, to mitigate contamination of storms on environmental conditions, with storm variables extracted at the current forecast hour. Statistical evaluations of the model track data include the mean, maximum, minimum, standard deviation, and 10th, 50th, and 90th percentiles.

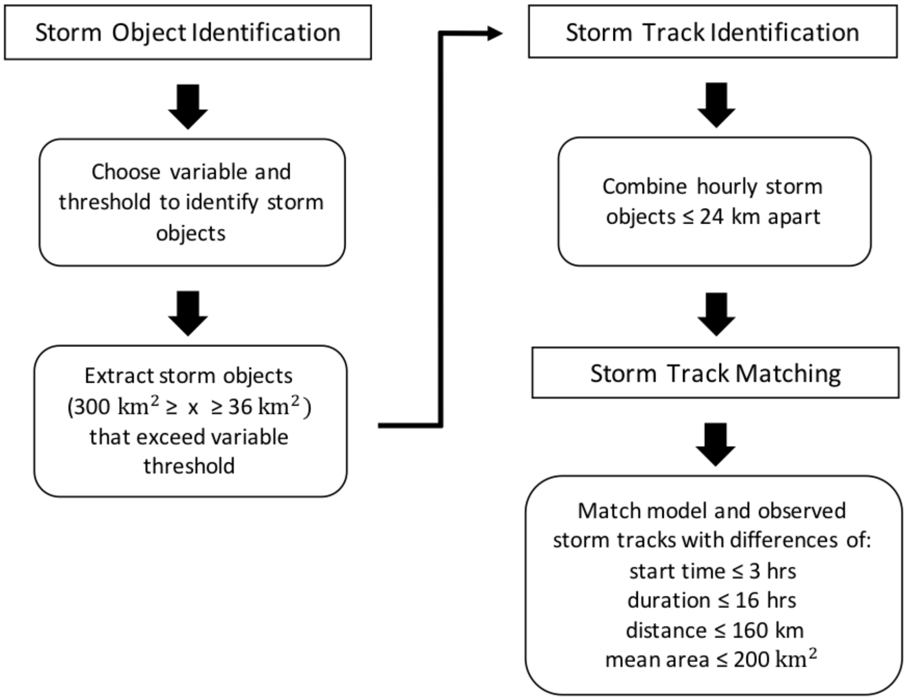

Storm Object Identification
Choose variable and threshold to identify storm objects
Extract storm objects (300 km^2 ≥ x ≥ 36 km^2) that exceed variable threshold
Storm Track Identification
Combine hourly storm objects ≤ 24 km apart
Storm Track Matching
Match model and observed storm tracks with differences of:
start time ≤ 3 hrs
duration ≤ 16 hrs
distance ≤ 160 km
mean area ≤ 200 km^2

Figure 3.1: Object-tracking method from Gagne et al. (2017) that first identifies storm objects, then combines the objects in time/space, and finally matches the observed and model storm tracks.

In addition to the input HREFv2 data, observations derived from the MESH storm tracks are needed to train the ML models for hail prediction. One observational dataset includes HREFv2 and MESH track pairings. Tracks within a distance of 80 km are paired together, where the distances between the model and observed hail tracks are calculated. The calculated distances take into account differences in track starting time and location, as well as duration and size differences. If any of the track distance parameters exceed maximum values defined in Gagne et al. (2017), the model and observed tracks are not paired. Setting higher maximum values for the different parameters allows for the uncertainty in track placement from either dataset. Additionally, hail sizes from the paired MESH tracks are associated with gamma distribution parameters and comprise another observational dataset.

For training the ML models, HREFv2 and MESH data are evaluated between 1 April to 31 July 2017, while the test set includes data from 1 May through 31 August 2018. Different years are used for training and testing to reduce the chance of overfitting. The duration of the training period (April-July) is selected based on greater hail potential and number of observations over the CONUS in the spring and early summer. The testing period includes the 2018 HWT SFE, from 30 April through 1 June 2018, during which forecasts were provided to HWT SFE participants for evaluation and feedback. April 2018 is not included for testing because of incomplete HREFv2 data. A description of the ML-based hail prediction method, using HREFv2 and MESH observations as inputs, is included in chapter 4.

Table 3.2: The twenty-nine HREFV2 storm and environment variables extracted during object-tracking. Multiple levels indicate a variable was extracted at separate elevations. CAPE is convective available potential energy, CIN is convective inhibition, MAXUVV is the Maximum Hourly Upward Vertical Velocity, and MAXDVV is the Maximum Hourly Downward Vertical Velocity.

Storm		**Environmental**	
Variable	**Level(s)**	**Variable**	**Level(s)**
MAXUVV	-	Precipitable Water	-
Storm Relative Helicity	1 and 3km	Temp	500, 700, 850, and 1000 hpa
Hourly Max Reflectivity	1km	Dew Point Temp	
MAXDVV	-	Geopotential Height	500, 700, and 850 hpa
Hourly Max UH	2-5 km	U Wind	
		V Wind	
		Hourly Max U Wind	
		Hourly Max V Wind	
		Surface Lifted Index	-
		CAPE	-
		CIN	-

Chapter 4

ML Hail Prediction and Evaluation Methods

To create 24-hour (1200 UTC to 1200 UTC the next day) ML hail forecasts, the HREFv2 model output and MESH observations are first pre-processed, as described in chapter 3. After object detection, the model storm tracks are input into ML models for predicting hail at the severe (>25 mm) and significant-severe (sig-severe, >50 mm) hail thresholds, to resemble a SPC hail outlook. The training and testing process for ML-based hail predictions is provided in section 4.1.

After producing 24-hour hail forecasts, the predictions are calibrated (section 4.2). Calibration towards familiar forecaster output, such as LSRs, produces output resembling human-made hail forecasts and therefore may increase trust in the ML models. Finally, forecast verification methods are detailed in section 4.3, with a model interpretation method described in section 4.4. The process for training and testing regional ML models is outlined in chapter 6.

4.1 Machine Learning Process

The ML model chosen for hail prediction in this thesis is the random forest (RF), or an ensemble of decision trees. Decision trees can be used for classification or regression tasks, to predict classes (hail or no hail) or specific numbers (hail size). A RF combines multiple decision trees together that each evaluate a random shuffling, with replacement, of variables. This randomness reduces the bias and variance of the model and generalizes predictions better when compared to a single decision tree. The number of variables, decision trees, splits the trees can make, and so on dictate the RF's performance. For this research, both the classification and regression RF models use 500 trees, where each tree analyzes a square root number of variables (14) and requires at least one sample per leaf node.

The multi-step process for ML-based hail prediction includes associating HREFv2 storm tracks with hail observations, then predicting hail sizes, and finally calibrating the ML probabilities of hail (Fig. 4.1). During the first step, each ensemble member predicts the probability of a HREFv2 storm track being matched with a MESH track. Using the input HREFv2 data and paired observational tracks described in section 3.2, a conditional threshold is determined through training a RF classification model with cross-validation (Breiman and Spector, 1992). If the predicted matching probability is greater than the conditional threshold, the HREFv2 storm track is associated with hail.

In the second step, matched HREFv2 storm tracks are evaluated using RF regression models to predict hail size. Two regression models individually determine the scale and shape parameters of a predicted MESH gamma distribution. To predict the parameters, a single regression model is trained, using the input and observational data from section 3.2, to output the log-normalized principal component of each parameter. Principal component predictions of the scale and shape parameters are returned to the original data space and resolve the MESH gamma distribution for each predicted matched storm object.

Predicted hail sizes for each HREFv2 track are extracted from the gamma distribution such that the highest storm object values, MAXUVV in this case, are associated with the largest MESH values. Unlike in Gagne et al. (2017), the scale and shape parameters of the hail size gamma distribution are predicted separately using independent RF regression models, instead of a single model to predict both. Also differing from Gagne et al. (2017), the distribution of storm objects created for each member are based off data from the entire training period rather than only daily values. Preliminary testing identified a subjective high MESH bias on marginal days when using the daily values.

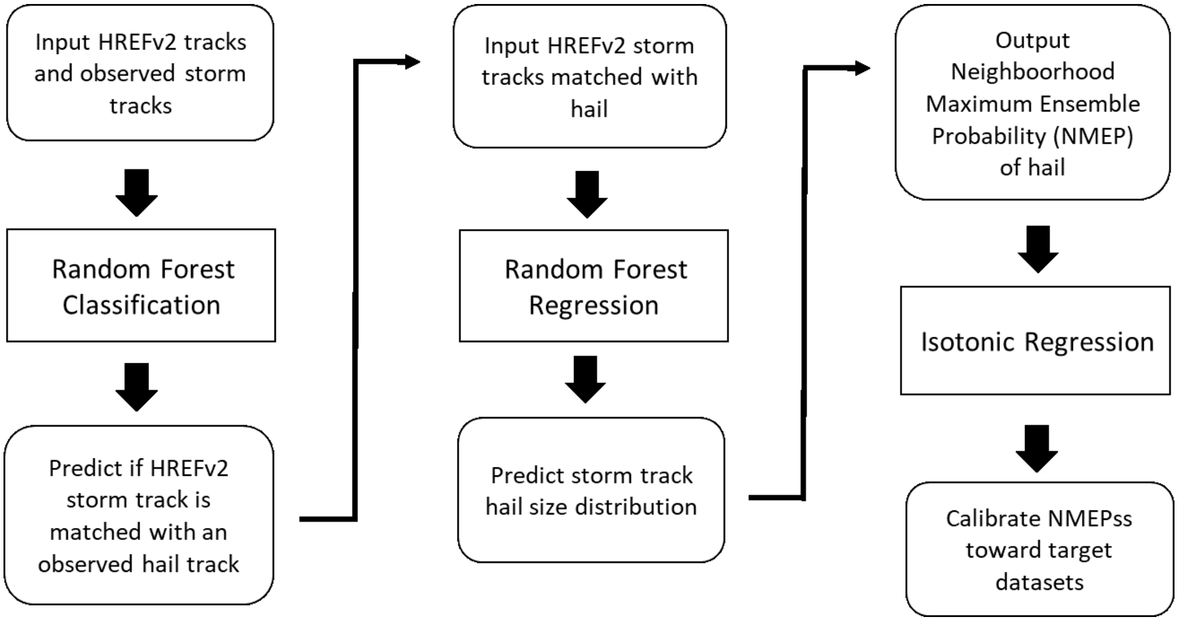

Input HREFv2 tracks and observed storm tracks
Random Forest Classification
Predict if HREFv2 storm track is matched with an observed hail track
Input HREFv2 storm tracks matched with hail
Random Forest Regression
Predict storm track hail size distribution
Output Neighboorhood Maximum Ensemble Probability (NMEP) of hail
Isotonic Regression
Calibrate NMEPss toward target datasets

Figure 4.1: Process for producing and calibrating machine learning hail predictions.

After each member predicts hail size, the neighborhood maximum ensemble probability (NMEP) of severe and significant severe hail within 42 km of a point is calculated over a 24-hour period. The NMEP hail predictions, based off the definition of ensemble probabilities in Schwartz and Sobash (2017), are evaluated on the 3 km HREFv2 grid. The grid is further smoothed with a 2D Gaussian filter (σ = 42 km), to manage the uncertainty in weather and allow for verification comparison to SPC products. For a complete description of the data pre-processing and ML methods, I refer the reader to Gagne et al. (2017).

4.2 Calibration

The uncalibrated ML NMEP predictions are capable of producing probabilities up to 100%, while SPC hail outlooks never predict probabilities exceeding 60%. Ideally, to build forecaster trust in the model, NMEP predictions from the HREFv2 ML model should produce comparable probability values with operational products, such as the SPC hail outlook. During the 2018 HWT SFE, forecasters were reluctant to trust the ML NMEP hail predictions due to the perceived high bias of the ML forecasts (Personal Communication, May 2018). To address this problem, the last step of the ML-based hail prediction process calibrates the NMEP hail predictions toward output similar to SPC forecasts.

The ML predictions are calibrated using isotonic regression, chosen because the model is computationally efficient and non-parametric. It should be noted that the overall area of non-zero hail probability does not change with isotonic regression, but only probability magnitudes. LSRs and SPC practically perfect (PP) (Davis and Carr, 2000; Hitchens et al., 2013) probabilities of severe hail occurrence were chosen as the two target datasets for calibration. The LSR target data was calculated as a binary field such that any grid point within 40 km of at least one severe or significant severe hail report was associated with hail. The SPC PP probability field is a dataset that serves, during the HWT SFE, as an estimate of the optimal outlook a forecaster would issue if all LSR locations were known beforehand. An additional Gaussian smoothing filter ($\sigma = 42$ km) applied to the PP probabilities accounts for the previously mentioned severe weather uncertainties. The input and target data for isotonic regression were extracted for the 24-hour forecasts, as well as three 4-hour periods (1700 to 2100, 1900 to 2300, and 2100-0100 UTC) used by 2018 HWT SFE forecasters. The calibration algorithm was analyzed with split training (70% of data) and testing (30% of data) sets over the original ML model predictions (1 May to 31 August 2018). The training and testing days were randomly selected to limit biases resulting from synoptic or seasonal patterns. Randomly choosing days to train/test could impact the calibrated forecasts based on the relative representation of severe hail within each dataset.

4.3 Forecast Evaluation

The 24-hour NMEP hail predictions are quantitatively and qualitatively verified over the full CONUS and regional subdomains, with added verification of the calibrated output over the full CONUS. The CONUS ML NMEP forecasts are subjectively verified against the SPC daily hail outlook and the PP output to examine similarities in spatial extent and magnitude. Both the severe and significant severe thresholds are mapped for consistency with the Gagne et al. (2017) study. The uncalibrated regional ML hail forecasts are also subjectively verified against the regional PP output.

Next, both domains of ML NMEP hail predictions are quantitatively verified through reliability, equitable threat score (ETS), and bias calculations. ETS and bias (Table 4.1) are calculated rather than the traditional performance diagram metrics, because of the PP's varying observational probabilities. Measuring ETS allows for observationally varying forecast probabilities and the ability to use the PP data, not just the LSRs, for verification.

All three metrics evaluate the CONUS ML model output over the isotonic regression test set (for equal comparisons across calibrated and uncalibrated datasets). Only the uncalibrated regional ML forecasts are verified, and therefore examined over the full testing set. Verification of the 24-hour CONUS forecasts includes an additional hail proxy field, updraft helicity (UH), to provide another non-ML baseline. The 2-5km HREFv2 NMEP UH predictions are calculated at the >75 m^2s^{-1} and >150 m^2s^{-1}

Table 4.1: Two of the forecast evaluation metrics used to examine the machine learning-based hail predictions.

Metric	**Equation(s)**
Bias	Bias $= \frac{Hits+FalseAlarms}{Hits+Misses}$
Equitable Threat Score (ETS)	ETS $= \frac{Hits-RandomHits}{Hits+Misses+FalseAlarm-RandomHits}$ Random Hits $= \frac{(Hits+Misses)(Hits+FalseAlarms)}{Total}$

thresholds, which are related to severe and sig-severe hail, respectively (Gagne et al., 2017). Reliability diagrams, examining both the regional and CONUS ML models, are calculated with the LSRs as observational truth, while the ETS and bias diagrams separately verify against the LSRs and PP datasets as truth. The subjectively best CONUS ML model is verified over different forecast periods used during the 2018 HWT SFE (1700-2100, 1900-2300, 2100-0100, 1200-1200 UTC). Rather than evaluate the regional ML models over different forecast periods, which would limit the already small dataset, statistical significance tests are conducted to identify differences between the regional and CONUS ML models over each subdomain. A greater description of the regional models is provided in chapter 6.

4.4 Model Interpretation

In addition to examination of the ML hail predictions, the ML framework for creating said predictions is explored through permutation variable importance. Biases towards correlated variables exist within the scikit-learn variable importance algorithm [1] (e.g., Strobl et al., 2008), and also permutation variable algorithms that create unrealistic circumstances (Molnar, 2019). A revised version of the Breiman (2001) method for permutation variable importance involves an iterative approach to better handle correlated variables (Lakshmanan et al., 2015).

For both importance methods, the ML models are trained once, after-which performance is calculated from an unpermuted model. The Brier Skill Score (BSS) determines the unpermuted ML model's performance in this thesis, although other skill metrics are also valid. After calculation of the unpermuted model's skill, each variable is shuffled successively amongst it's cases. For each permuted variable, a new skill score is calculated to determine the model's performance after shuffling. If performance decreases relative to the original skill score, then the shuffled variable was considered important for prediction. The original variable importance method shuffles each variable independently, leaving the other variables unpermuted, and calculates changes in skill over multiple iterations (Breiman, 2001). The first pass of the revised method resembles one iteration of the original algorithm, where the variable with the greatest decrease in skill is most important (rank 1). After the first important variable is discovered, another pass of the original method determines the second most important variable, however the first variable remains shuffled. The third relevant variable is found with the first two important variables remaining shuffled, and so on and so forth until the desired number of variables are ranked for importance (Lakshmanan et al., 2015).

An application of the revised method was developed by Eli Jergensen.[2] While bootstrapping the skill calculations for each shuffled variable would produce more accurate results, the HREFv2 variable scores were not resampled. Calculating bootstrapped

[1] `https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html`

[2] `https://github.com/gelijergensen/PermutationImportance`

skill for each shuffled variable (208) across the HREFv2 members (8), repeatedly, is very time consuming. Ranking a subset of variables for importance, such as 20, with bootstrapped skill distributions would be more computationally feasible. Although the HREFv2 variable skill scores are not resampled, some bias is accounted for when averaging the variable ranks across the ensemble. The ensemble rank results are presented on a log scale for less cluttered interpretation.

Chapter 5

CONUS ML Hail Forecasts

After the multi-step process for producing ML-based hail forecasts, verification of the NMEP hail predictions identifies areas of success within the ML models, as well as areas of improvement. Mentioned in chapter 1, forecasters are less likely to trust guidance without reliability and skill evaluation (Karstens et al., 2015). Towards this end, objective and subjective verification of the CONUS-trained ML model output is evaluated. Case studies of marginal (section 5.1) and high-end (section 5.2) hail events are investigated to determine the robustness of the ML model framework over differing hail severity regimes. Storm reports are overlaid on the NMEP hail predictions for further visual inspection of performance. The ML output is examined objectively through calculations of reliability (section 5.3.1), equitable threat score (ETS), and bias (section 5.3.2). Lastly, the CONUS-trained model framework is evaluated using permutation variable importance, to determine which fields are important for CONUS ML hail prediction (section 5.4). A discussion of the different evaluation results is included in section 5.5.

5.1 Marginal Hail Case Study

On 8 May 2018, the SPC hail outlook valid 1200 UTC included two regions with 5% probability of hail (Fig. 5.1). On this day, a trough and surface cold front moved into Oregon. Although CAPE was relatively weak, a deep mixed layer provided enough support for a few non-severe hail-producing storms in the western US. Storms over the mid to lower Missouri Valley were associated with ample CAPE, but were restricted by a dry boundary layer ahead of a surface trough. The PP output showing severe hail probabilities on 8 May 2018 includes values between 5% and 15% primarily over South Dakota, Minnesota, and northwestern Iowa (Fig. 5.1b). There were no areas of significant severe (sig-severe) hail probability (Fig. 5.1c), as well as no severe hail reports over the western United States.

The uncalibrated ML NMEP prediction (Fig. 5.2a), valid 1200 UTC on 8 May 2018 contains hail probabilities up to 35% over portions of Iowa, South Dakota, Nebraska, and Missouri. Additionally, there were probabilities up to 22% over Oklahoma and Kansas, and a separate region over Oregon. These predicted probabilities are substantially higher than those in the corresponding SPC outlook (Fig. 5.1a) or the PP output (Fig. 5.1b). The region of highest probability in the uncalibrated severe hail prediction (Fig. 5.2a) is displaced southeast of the observed severe hail reports (black dots). At the sig-severe threshold, the uncalibrated hail prediction indicates probabilities up to 4% over eastern Iowa. No significant severe hail was reported (Fig. 5.2b). Overall, the uncalibrated output over-forecasts both severe and sig-severe hail.

Compared to the uncalibrated hail prediction (Fig. 5.2a), the severe NMEP output calibrated to the LSRs (Fig. 5.2c) more closely resembles the SPC hail outlook (Fig. 5.1a) and the severe hail PP output (Fig. 5.1b) in terms of maximum probability. The LSR calibrated severe hail output has the same spatial coverage of non-zero hail probabilities as the uncalibrated prediction, but with probabilities not exceeding 14%. At the sig-severe threshold, the LSR calibrated prediction (Fig. 5.2d) exhibits a similar area of probabilities as the uncalibrated (Fig. 5.2b), but with a slightly larger extent of the lower probabilities. As no sig-severe hail was reported on this day, the sig-severe

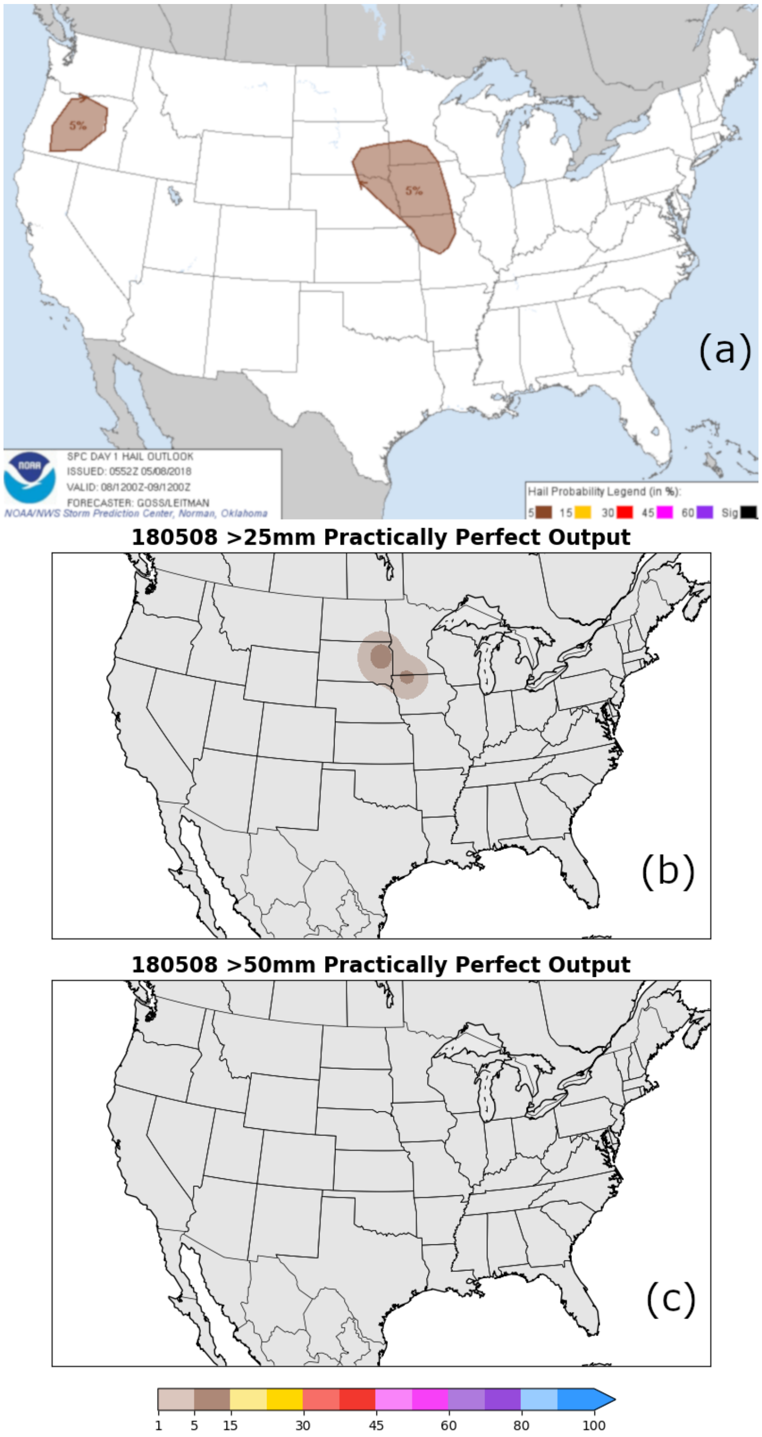

5%
5%
(a)
NOAA
SPC DAY 1 HAIL OUTLOOK
ISSUED: 0552Z 05/08/2018
VALID: 08/1200Z-09/1200Z
FORECASTER: GOSS/LEITMAN
NOAA/NWS Storm Prediction Center, Norman, Oklahoma
Hail Probability Legend (in %):
5 15 30 45 60 Sig
180508 >25mm Practically Perfect Output
(b)
180508 >50mm Practically Perfect Output
(c)
1 5 15 30 45 60 80 100

Figure 5.1: SPC forecasts from 8 May 2018 including (a) the day 1 hail outlook valid 1200 UTC, (b) practically perfect output at the severe hail threshold,and (c) significant severe threshold.

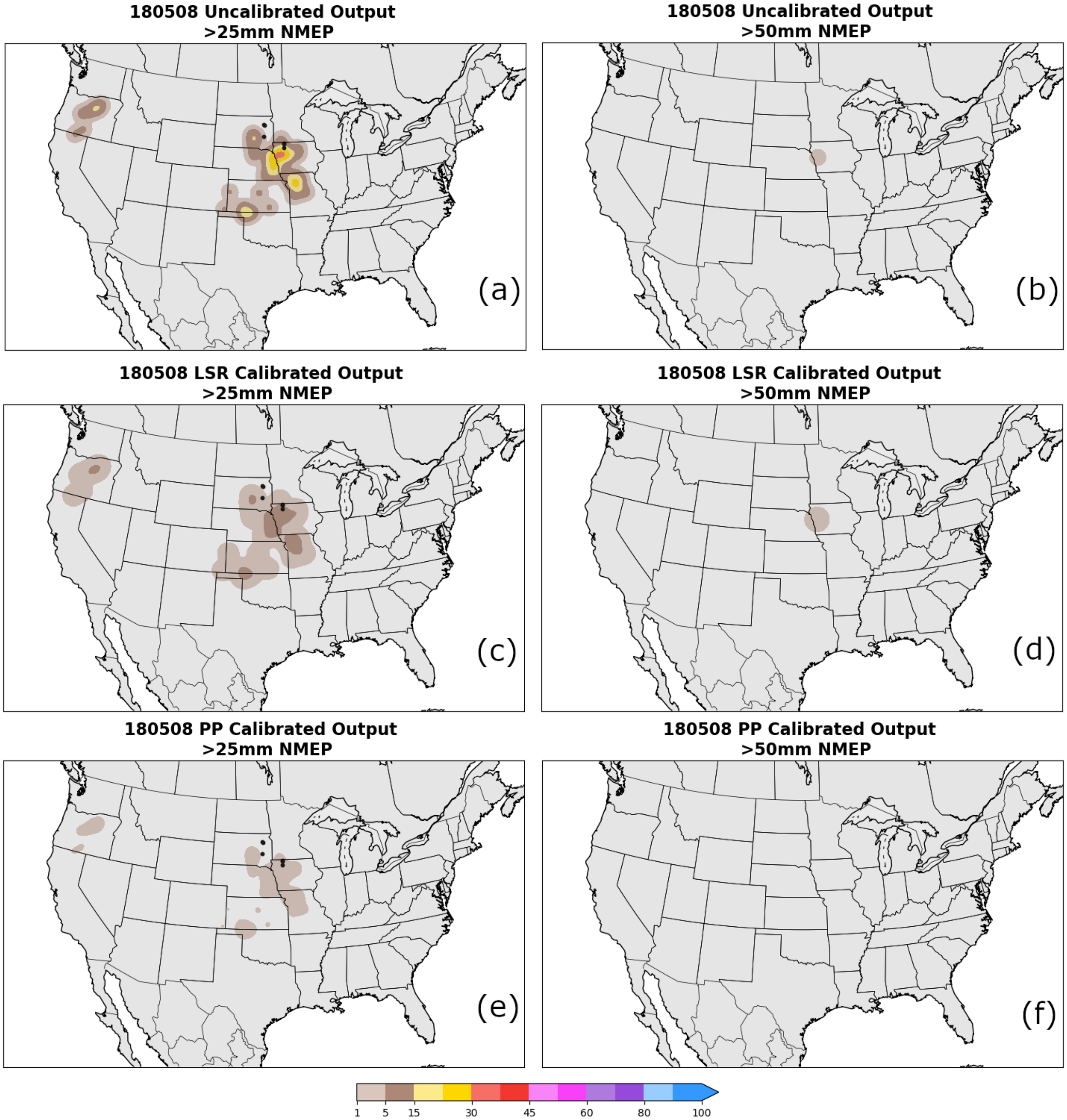

180508 Uncalibrated Output
>25mm NMEP
(a)
180508 Uncalibrated Output
>50mm NMEP
(b)
180508 LSR Calibrated Output
>25mm NMEP
(c)
180508 LSR Calibrated Output
>50mm NMEP
(d)
180508 PP Calibrated Output
>25mm NMEP
(e)
180508 PP Calibrated Output
>50mm NMEP
(f)
1
5
15
30
45
60
80
100

Figure 5.2: ML neighborhood maximum ensemble probability of hail predictions for 8 May 2018. Both calibrated and uncalibrated predictions are produced at the (a,c,e) severe and (b,d,f) significant severe hail thresholds. Predictions are calibrated to the local storm reports (LSRs) and practically perfect (PP) probabilities. The black dots are severe and significant severe hail reports.

LSR calibrated prediction displays a slight high bias. Overall, the LSR calibrated model for this case demonstrates how calibration using isotonic regression can help the magnitude over-prediction bias present in the uncalibrated hail forecasts. More data at the sig-severe threshold could impact the slight spatial bias associated with the LSR calibrated model.

Next, the NMEP hail predictions are also calibrated to the SPC PP dataset. Overall, calibration to the PP data yields lower probabilities of severe (Fig. 5.2e) and sig-severe (Fig. 5.2f) hail than the LSR calibrated (Fig. 5.2c, d). The severe hail probabilities predicted by the PP calibrated model are lower than the SPC PP dataset, though the overall area exceeding 1% probability is similar to that of the LSR calibrated (Fig. 5.2c). At the sig-severe threshold, the PP calibrated model correctly predicts no areas exceeding a 1% chance of sig-severe hail (Fig. 5.2f).

Overall, the severe calibrated predictions exhibit probability values more comparable to the SPC hail outlook and PP data than the uncalibrated output. The overall spatial extent of probabilities exceeding 5% for all three severe ML predictions was also similar to the area of 5% hail risk identified in the SPC Day 1 hail outlook for this case. At the sig-severe threshold, more data could further decrease the over-forecasting bias in the uncalibrated prediction without increasing the areal extent as was seen in the LSR calibrated forecast.

5.2 High-end Hail Case Study

A high-end severe hail event occurred on 29 July 2018, with multiple hail-producing storms resulting in 106 severe and 30 significant severe hail reports. Most of the reports concentrated over Colorado, with some extending into Wyoming, Kansas, South Dakota, and Nebraska. On this day, a strengthening upper-level trough and mid-level jet over the Mid-west, strong diurnal heating, and a moist boundary layer set the stage for severe storms over the Central High Plains. The day 1 SPC hail outlook valid 1200 UTC (Fig. 5.3a) featured a 15% chance of severe hail over eastern Colorado and a 5% chance of severe hail over a larger region extending from Montana to Arkansas. Sig-severe hail was not anticipated until the hail outlook valid 1300 UTC. The severe hail PP output (Fig. 5.3b) indicates maximum probabilities up to 38% over northeastern Colorado. At the sig-severe threshold, the PP output (Fig. 5.3c) shows probabilities ranging from 15% to 23%, again primarily over eastern Colorado.

The uncalibrated NMEP hail prediction (Fig. 5.4a) valid 1200 UTC contains probabilities of severe hail exceeding 5% over a similar region as the SPC hail outlook (Fig. 5.3a) and PP output (Fig. 5.3b). However, uncalibrated probabilities exceed 15% over Arizona, a region without predicted probabilities in the SPC hail outlook. No severe hail was reported in Arizona, however there was one report in nearby southern Nevada (Fig. 5.4a). Neither the uncalibrated prediction nor the SPC outlook indicates a severe hail threat in North Dakota or Minnesota, where two isolated instances of severe hail were reported. In general, the uncalibrated probabilities of severe hail are substantially greater than those in the SPC hail outlook, with probabilities exceeding 60% over eastern Colorado. The uncalibrated sig-severe hail prediction (Fig. 5.4b) also displays high probabilities, though not as substantial, in eastern Colorado where values exceed 15%. In general, the region of highest probabilities at both hail thresholds is closely co-located with the bulk of the observed hail reports. However, some regions of severe probabilities over Arizona and southeastern Oklahoma are not in the vicinity of any hail reports (Fig. 5.3a). Despite the high uncalibrated probabilities of severe and sig-severe hail, the high-end day's ML predictions are comparable with the PP output in terms of spatial extent and location, more so than in the marginal hail case.

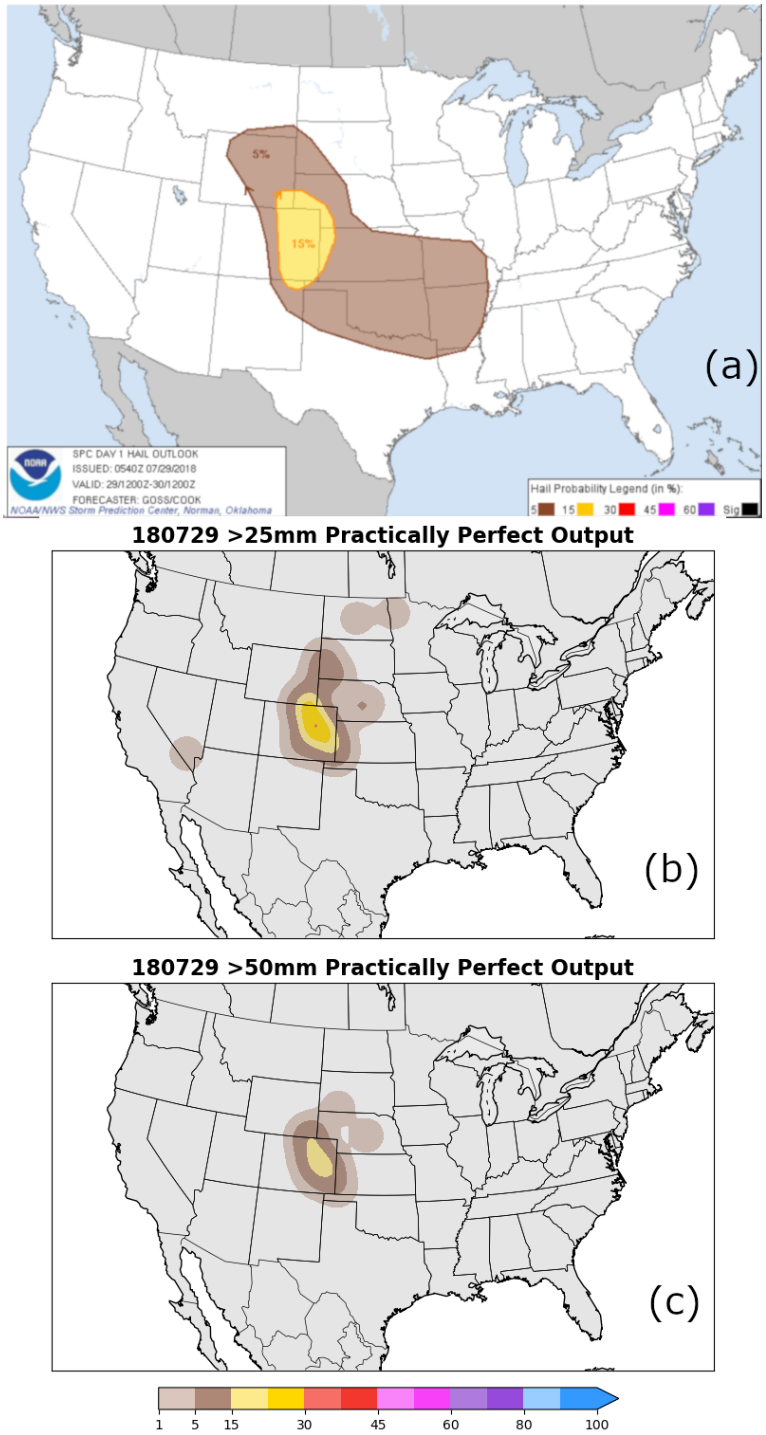

5%
15%
(a)
SPC DAY 1 HAIL OUTLOOK
ISSUED: 0540Z 07/29/2018
VALID: 29/1200Z-30/1200Z
FORECASTER: GOSS/COOK
NOAA/NWS Storm Prediction Center, Norman, Oklahoma
Hail Probability Legend (in %):
5 15 30 45 60 Sig
180729 >25mm Practically Perfect Output
(b)
180729 >50mm Practically Perfect Output
(c)
1 5 15 30 45 60 80 100

Figure 5.3: Similar to Fig. 5.1, SPC forecasts including the (a) day 1 hail outlook valid 1200 UTC, (b) practically perfect output at the severe hail threshold, and (c) significant severe threshold, all valid 29 July 2018.

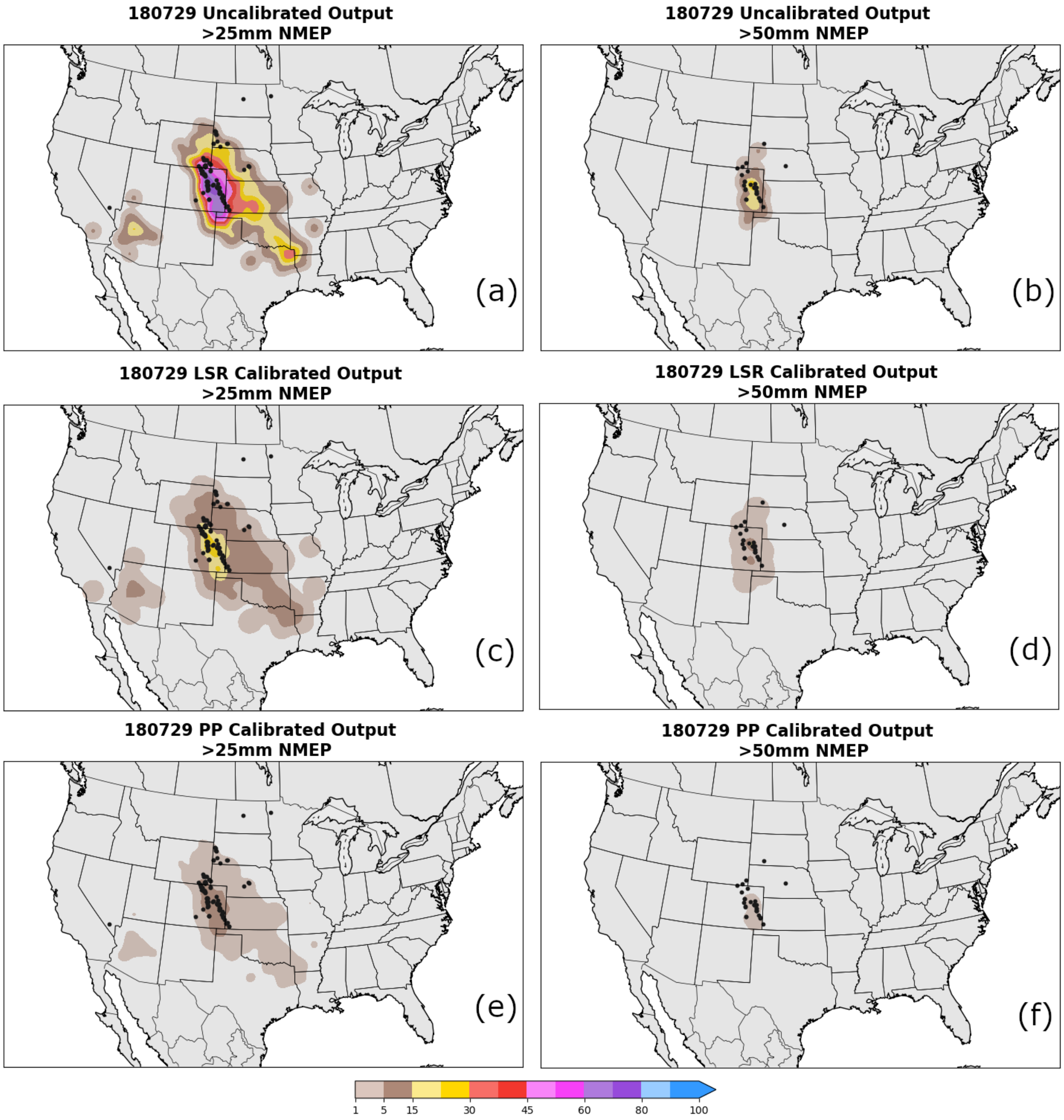

180729 Uncalibrated Output
>25mm NMEP
(a)
180729 Uncalibrated Output
>50mm NMEP
(b)
180729 LSR Calibrated Output
>25mm NMEP
(c)
180729 LSR Calibrated Output
>50mm NMEP
(d)
180729 PP Calibrated Output
>25mm NMEP
(e)
180729 PP Calibrated Output
>50mm NMEP
(f)
1
5
15
30
45
60
80
100

Figure 5.4: Similar to the Fig. 5.2, ML neighborhood maximum ensemble probabilities of hail valid 29 July 2018. Predictions are produced at the (a,c,e) severe and (b,d,f) significant severe hail thresholds. Calibration is accomplished using local storm reports (LSRs) and practically perfect (PP) probabilities. The black dots are severe and significant severe hail reports.

Compared to the uncalibrated hail output, the LSR calibrated severe hail NMEP (Fig. 5.4c) displays probabilities closer in magnitude to the PP output. The severe LSR calibrated prediction exhibits a large region of probabilities between 5% and 15%, with a maximum around 22% in eastern Colorado, a value slightly lower than the PP output (Fig. 5.3b). As in the uncalibrated prediction, the LSR calibrated model does not predict storms in North Dakota and eastern Minnesota. For sig-severe hail, the LSR calibrated probabilities (Fig. 5.4d) reach at least 14% over some areas in Colorado but are lower than the 15% maximum indicated in the PP output (Fig. 5.3c). Additionally, there is a slightly larger spatial extent of lower probabilities in the sig-severe LSR calibrated prediction, similar to the marginal hail case. The larger spatial extent and under-prediction of probability magnitudes associated with the LSR calibrated sig-severe prediction results in part from the small training dataset, and the overall rarity of sig-severe hail. In general, for probability magnitude and distribution, the severe LSR calibrated prediction resembles the PP output but under-predicts the sig-severe magnitudes.

As in the marginal hail case study, the severe PP calibrated prediction (Fig. 5.4e) under-forecasts the severe hail probability magnitudes, compared to the SPC hail outlook (Fig. 5.3a) and PP output (Fig. 5.3b). The severe PP calibrated prediction exhibits very low probabilities of severe hail, less than 15%, with a small region of probability exceeding 5% over eastern Colorado (in comparison, the severe PP probability reaches 30% in this vicinity). The sig-severe PP calibrated predictions also under-forecast compared to the PP output (Fig. 5.4f). Like the marginal hail case, while the PP calibrated model could be useful as a conservative estimate of hail, the LSR calibrated severe hail predictions better resemble both the SPC outlook and PP output in terms of probability magnitudes. Also, more sig-severe hail reports could decrease the slightly high spatial extent bias in the LSR calibrated model.

5.3 Quantitative Verification

All of the CONUS-trained ML models, both calibrated and uncalibrated, are quantitatively verified through reliability, ETS, and bias. The verification metrics are assessed over the isotonic regression test set which consists of 37 days chosen at random out of 121, from 1 May to 31 August, 2018. All of the metrics evaluate predictions over the day-ahead (12-36 hr) forecast period, with additional periods included in the ETS and bias plots analyzing the LSR calibrated output.

5.3.1 Reliability

Evidenced by the reliability diagrams, where the LSRs are observational truth, the PP output exhibits a consistent low bias and the uncalibrated ML model a consistent high bias at both the severe and sig-severe thresholds (Fig. 5.5). Additionally, the uncalibrated ML output and UH proxy display comparable forecasting bias at the severe threshold (Fig. 5.5a), with the uncalibrated more reliable at probabilities below 25% and exceeding 80%. At the sig-severe threshold, the UH proxy persists with a high bias, while the uncalibrated ML output displays near-perfect reliability up to about 5%, after which it begins over-forecasting and is less reliable than UH after about 15%. Next, we expect the PP and LSR calibrated predictions to have comparable reliability characteristics as the target datasets they are regressed towards (PP output or perfect reliability for the LSRs). As expected, the LSR calibrated output exhibits near perfect reliability between 0 to 25% at the severe hail threshold, with a slight high bias between 25 and 35% (Fig. 5.5a). At the sig-severe threshold, the LSR calibrated predictions consistently over-forecast across the output forecast probabilities (Fig. 5.5b). The LSR calibrated curves do not extend past 40% and 20% at the severe and sig-severe thresholds respectively. This behavior is more pronounced for the PP calibrated output, which rarely predicts probabilities exceeding 20% and 1% at the severe and sig-severe thresholds. Nevertheless, the ML PP calibrated and PP output exhibit comparable reliability behavior at the severe hail threshold up to 20% (Fig. 5.5a). The PP calibrated probabilities at the sig-severe threshold are sufficiently low that they do not appear (Fig. 5.5b).

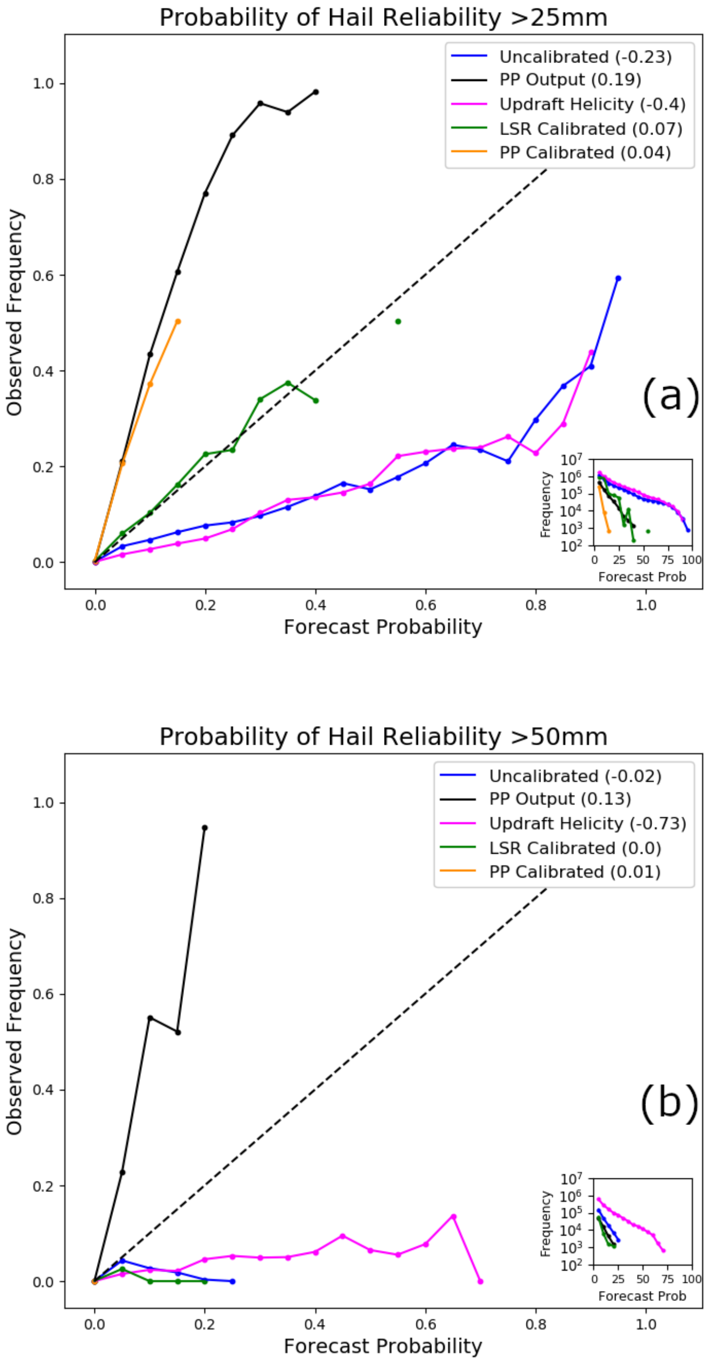


Figure 5.5: Reliability diagrams of the 24-hour ML predictions, practically perfect output, and updraft helicity proxy over the test dataset. Predictions are verified at the (a) severe and (b) significant severe hail thresholds. The legend displays the Brier Skill Score across all forecast probabilities.

5.3.2 Equitable Threat Score and Bias

With respect to skill, ideal hail forecasts maximize ETS while retaining a bias of 1. As the PP output is the ideal probabilistic prediction based on the LSRs, the differing forecasts are compared to the PP data. In Fig. 5.6, ETS and bias are calculated to examine the different hail predictions with the LSR dataset as observational truth. The LSR calibrated and PP output exhibit similar patterns in ETS at the severe hail threshold (Fig. 5.6a), with both datasets maximized at 15% forecast probability. However, the PP output achieves an ETS value of 0.35 while the LSR calibrated only reaches about 0.12. The PP calibrated probabilities also exhibit a clear maximum in skill, similar in magnitude to the LSR calibrated but at a lower forecast probability (around 5%). Both the uncalibrated output and UH probabilities achieve higher skill than the PP output past 25%, 0.12 at 35% for the uncalibrated output and 0.14 at 37% for the UH proxy. However, both outputs have the highest bias (Fig. 5.6b) among the datasets. It should be noted that the uncalibrated model displays lower bias before 60%, compared to the UH proxy. Of the ML predictions, the LSR calibrated probabilities demonstrate higher skill past 10% (Fig. 5.6a), while also retaining bias values closer to the optimal PP output (Fig. 5.6b).

At the sig-severe threshold, the LSR calibrated and uncalibrated predictions exhibit similar skill, maximized around 0.03 (Fig. 5.6c). After 2%, the uncalibrated predictions have greater skill compared to the LSR calibrated. The PP calibrated output does not exceed skill greater than 0.01 past about 2% at the sig-severe threshold. As with severe hail, the UH field achieves higher skill compared to the PP output past 20%, but again has the highest bias (Fig. 5.6d). Unlike the severe threshold, the sig-severe uncalibrated predictions exhibit bias much closer to the PP output. Also, while the LSR calibrated predictions have lower skill, the model's bias is comparable with the

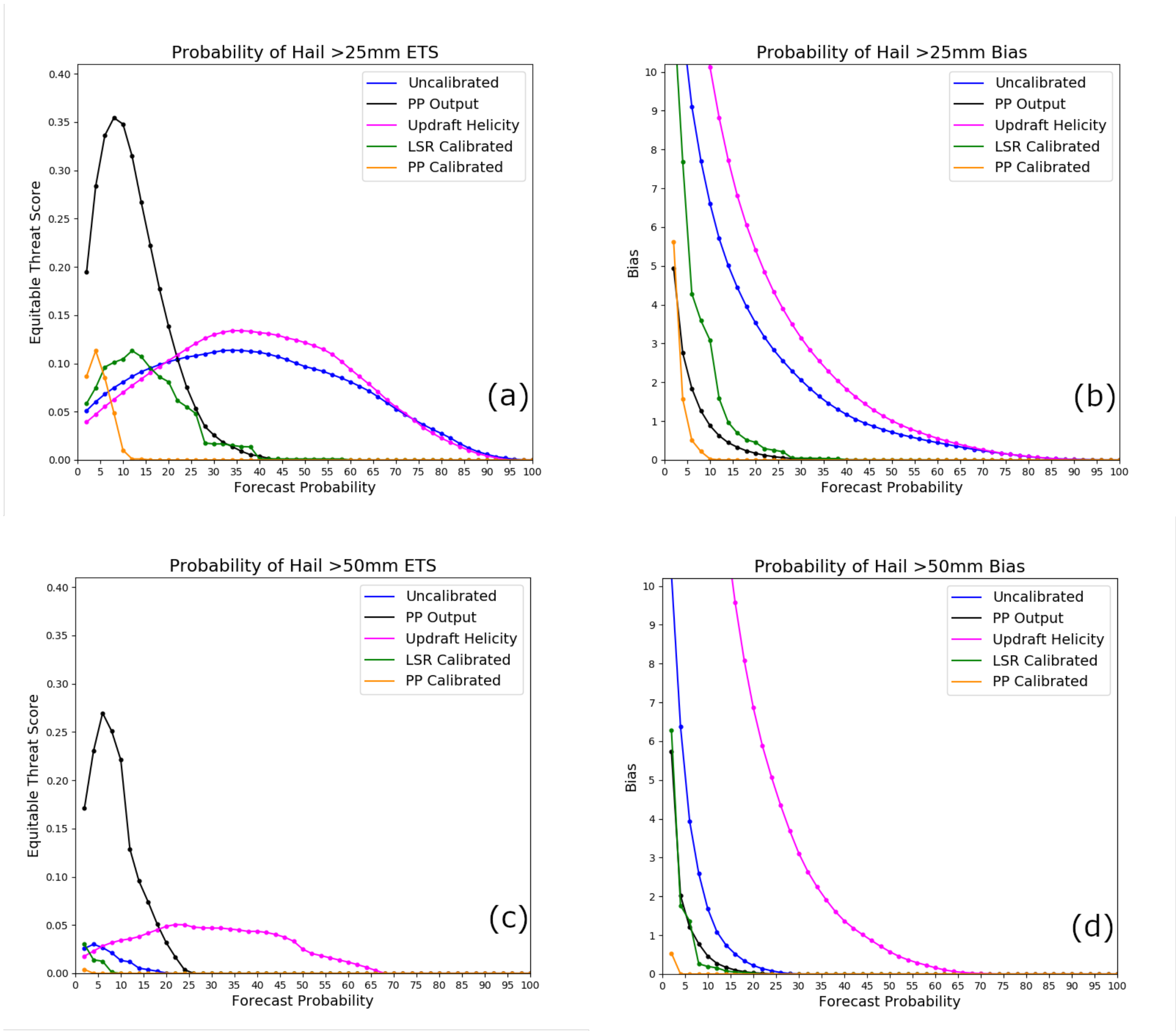


Figure 5.6: Equitable threat score (ETS) and bias plots evaluating the 24-hour ML output and practically perfect probabilities, with local storm reports as observations. ETS and bias are calculated at the (a,b) severe and (c,d) significant severe hail thresholds.

PP output at the sig-severe threshold. For forecasting, bias could potentially be more important to forecasters than higher skill given the differing aspects of a forecast that each skill score prioritizes.

Next, ETS and bias are calculated using the PP dataset as observational truth (Fig. 5.7) to analyze each prediction's performance against the optimal SPC dataset. The PP calibrated output displays the highest skill (0.25) at the severe hail threshold, however rapidly decreases in skill after 5% (Fig. 5.7a). Additionally, the PP calibrated predictions under-forecast severe hail (Fig. 5.7b) after 2% forecast probability with a bias of 0 by 10%. The LSR calibrated severe hail output produces the highest skill after 5% (Fig. 5.7a) while also exhibiting lower bias than both the UH proxy and the uncalibrated output (Fig. 5.7b). There is a spike in bias of the LSR calibrated output at 35%, where the model over-forecasts compared to the PP observations. This could be a function of the conservative nature of the PP observations as probability magnitudes past 35% are rarely issued (Fig. 5.6a,b). Both the uncalibrated ML predictions and UH proxy show near identical skill, but with lower uncalibrated forecast bias compared to the UH field (Fig. 5.7a,b).

Predictions verified against the PP dataset at the sig-severe hail threshold (Fig. 5.7c,d) indicate the UH field has the highest skill. However, the UH proxy also has the highest bias of all the predictions (Fig. 5.7d). Similar to the models verified against the sig-severe LSRs, the LSR calibrated model has lower skill than the uncalibrated but with the bias closest to the PP output bias, or a value of 1 in this case. The uncalibrated model has similar skill as the UH field for the lower probabilities, or about 0.09 at 5% and quickly decreasing to 0 by 10% (Fig. 5.7c). As with the severe threshold, the uncalibrated model output has lower bias than the UH proxy across all of the probability thresholds (Fig. 5.7d). The sig-severe PP calibrated output displays low

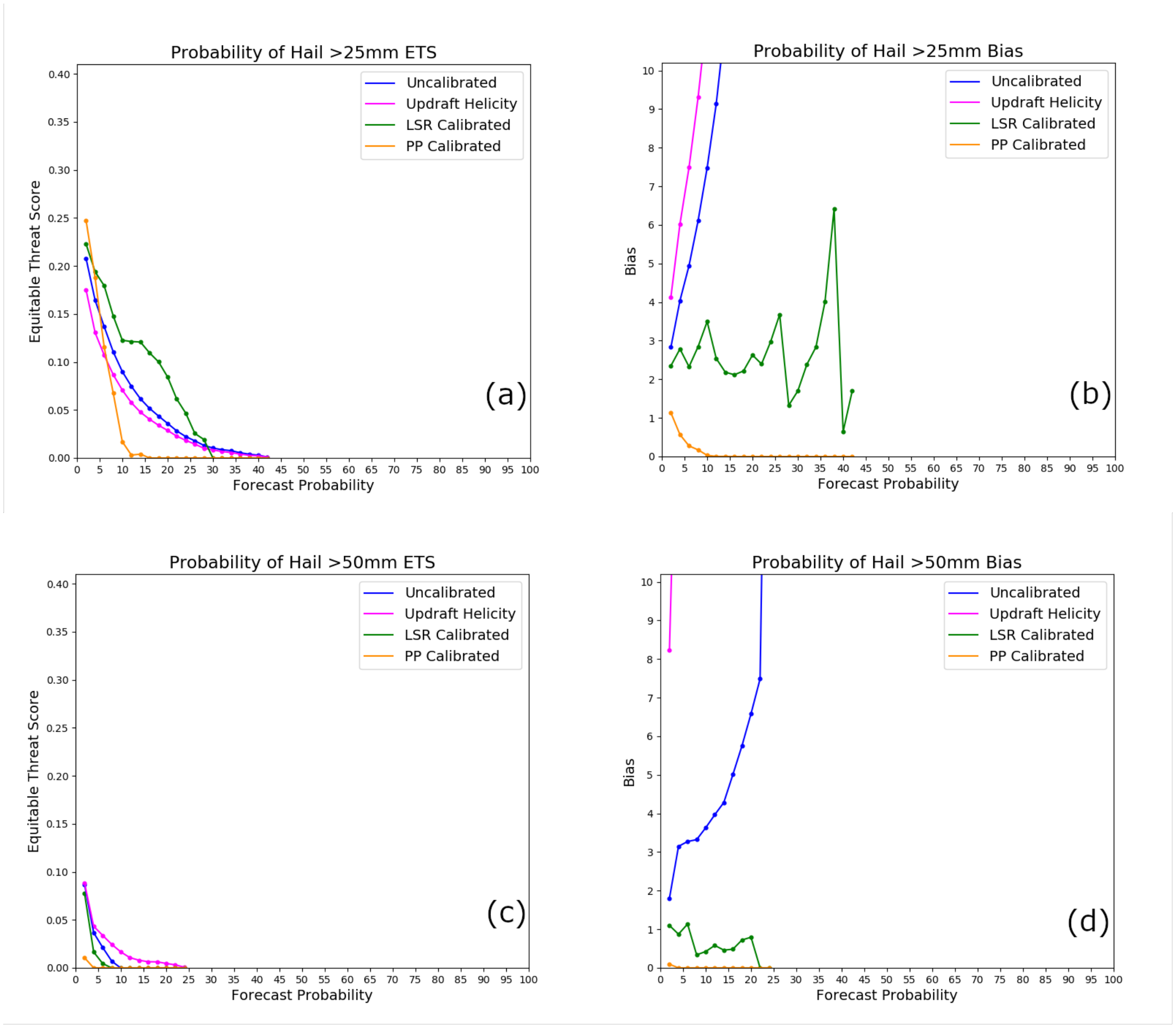


Figure 5.7: Similar to Fig. 5.6, equitable threat score (ETS) and bias plots evaluate the 24-hour ML hail predictions, with practically perfect probabilities as observations. ETS and bias are calculated at the (a,b) severe and (c,d) significant severe hail thresholds

skill, maximizing at 0.01 around 2%, with near-zero bias across all probabilities (Fig. 5.7c,d). In general, the LSR calibrated output verified against the PP probabilities has low ETS skill, but displays the most similar bias to the PP output (compared to the uncalibrated and proxy output). These observations are similar to the sig-severe ETS and bias diagrams analyzed against the LSRs (Fig. 5.6c,d).

Finally, the LSR calibrated outputs are analyzed over different forecast periods (1700-2100, 1900-2300, 2100-0100, and 1200-1200 UTC the next day) against both the PP and LSR as observations. The LSR calibrated model is analyzed due to the subjective success of the model output from the previous verification metrics. Only the severe hail threshold is evaluated due to the rarity of sig-severe hail reports, especially over four-hour periods. A similar analysis at the sig-severe threshold could be enlightening, however a much larger dataset would be needed. The predictions evaluated against both the LSRs (Fig. 5.8a) and PP output (Fig. 5.8c) display generally improved skill later in time. Analyzing the ETS score associated with the LSRs as observations (Fig. 5.8a) shows that the latest four-hour time period, 2100-0100 UTC, exhibits the greatest skill after 15% forecast probability. At the lower probabilities, the 2100-0100 UTC time period has comparable skill as the day-ahead forecasts (1200-1200 UTC). Verifying against the PP observations (Fig. 5.8c) shows that the 2100-0100 UTC and day-ahead periods also produce the highest skill, with the day-ahead predictions achieving slightly greater skill at higher probabilities.

With regards to bias, the predictions verified against the LSRs (Fig. 5.8b) show similar bias across the different forecast periods. Comparatively, the predictions verified against the PP dataset (Fig. 5.8d) exhibit relatively dissimilar bias values. The 1700-2100 UTC period has consistently low bias compared to all the other periods. Lack of severe hail samples within the training set for the 1700-2100 UTC period may be why

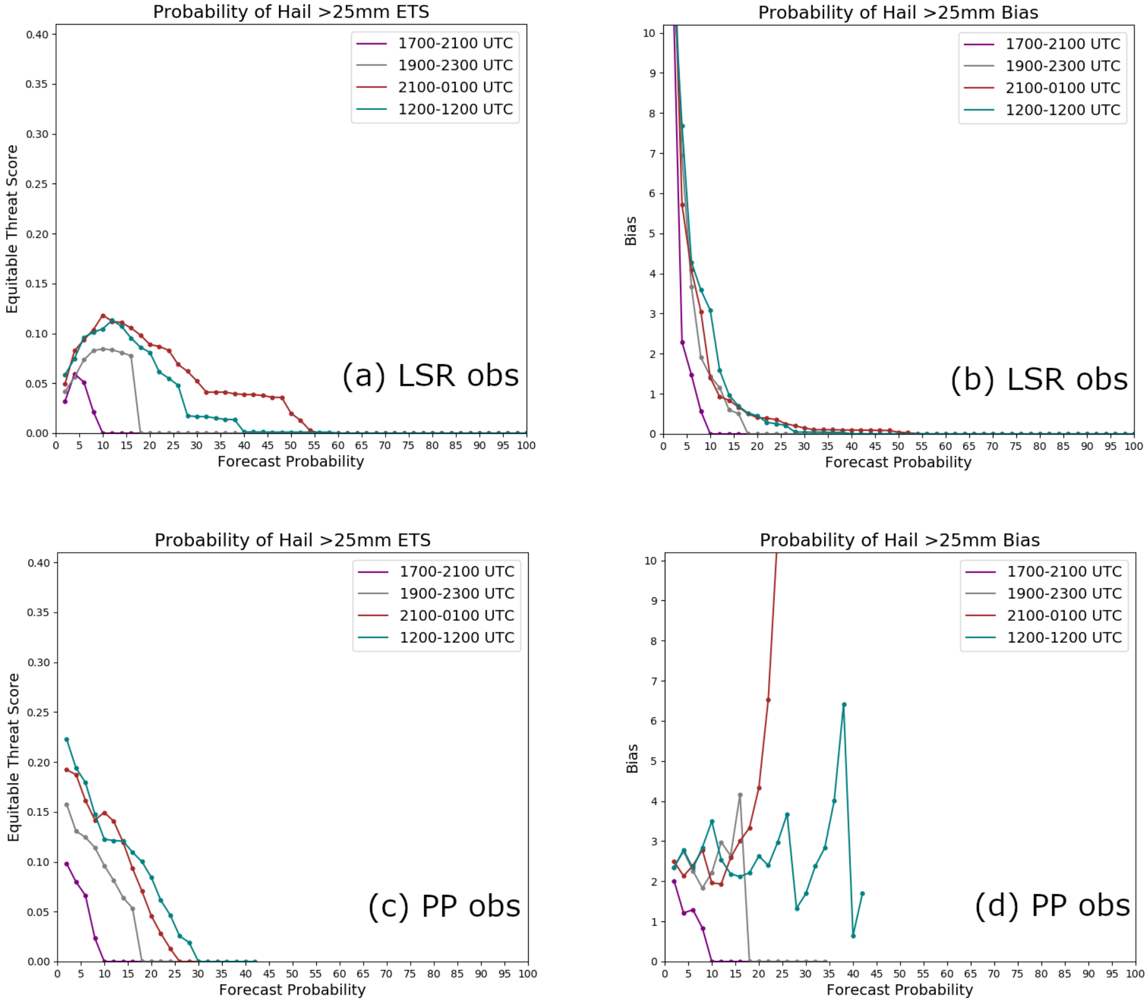


Figure 5.8: Plots of equitable threat score (ETS) and bias evaluating the LSR calibrated output at varying forecast periods. Predictions are verified at the severe hail threshold using the (a,b) local storm reports and separately (c,d) practically probabilities as observations.

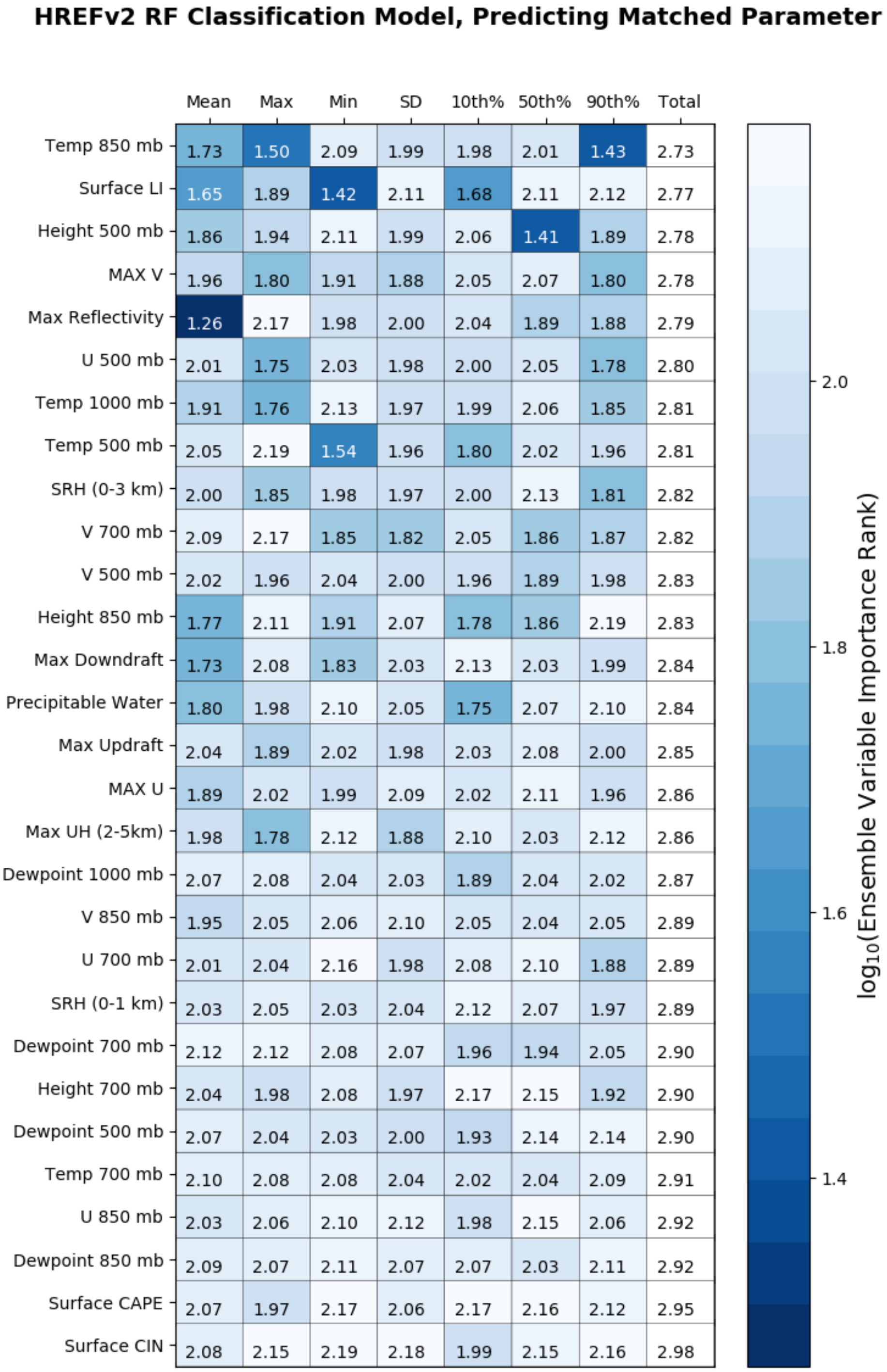
the early period exhibits overall decreases in skill, as the ML model rarely predicts severe hail for this time. After 15%, the 1900-2300 UTC period peaks in bias around 17% before decreasing in value. For the day-ahead forecasts, bias fluctuates between 2 and 3.5 for low probabilities, with a spike to 6.5 at around 35%. Bias for the 2100-0100 UTC period rises quickly for probabilities above 15%, exceeding a bias of 10 for probabilities of 30% or higher. In general, the later time periods exhibit the greatest performance, regardless of the observational verification dataset.

5.4 Model Interpretation

The permutation variable importance algorithm evaluating the CONUS-trained ML models is trained and tested over the 2017 and 2018 seasons, respectively. In the testing dataset, the same random 37 days are examined by each HREFv2 member as were investigated quantitatively above. A description of the variable importance algorithm can be found in chapter 4, where correlated variables are addressed within the multipass method. Starting with the RF classification model, the 850 mb temperature, surface lifted index, and 500 mb height fields display the lowest total rankings, or greatest importance (Fig. 5.9) for classifying HREFv2 storm objects as hailstorms. The most important variable of all the 2D fields for classifying CONUS HREFv2 tracks is the mean hourly maximum reflectivity, followed by the median 500 mb height field. Other surface (CAPE and CIN) and environmental fields (U wind and dewpoint 850 mb) are least important when classifying the HREFv2 storm objects.

In predicting the shape parameter of the gamma MESH distribution, the CONUS-trained RF regression model distinguishes the U wind field at differing levels as most important (Fig. 5.10). The mean 700 mb U wind achieves the lowest ranking of all the variables, followed by the minimum 1000 mb temperature. Generally, the RF regression model predicting the shape parameter favors low-level environmental fields over storm variables, differing from the RF classification model. The low-level variables are also important in predicting the scale parameter of the MESH gamma distribution (Fig. 5.11). The overall most important variables to the RF regression model also include the 850 and 700 mb U wind, with the addition of precipitable water as a relevant field. The most important 2D variable is the maximum 850 mb dewpoint, with the mean 850 mb U wind ranking second. Compared to the other two models, the storm variables are the least important when predicting the scale parameter, including the variable chosen for inital HREFv2 object identification.

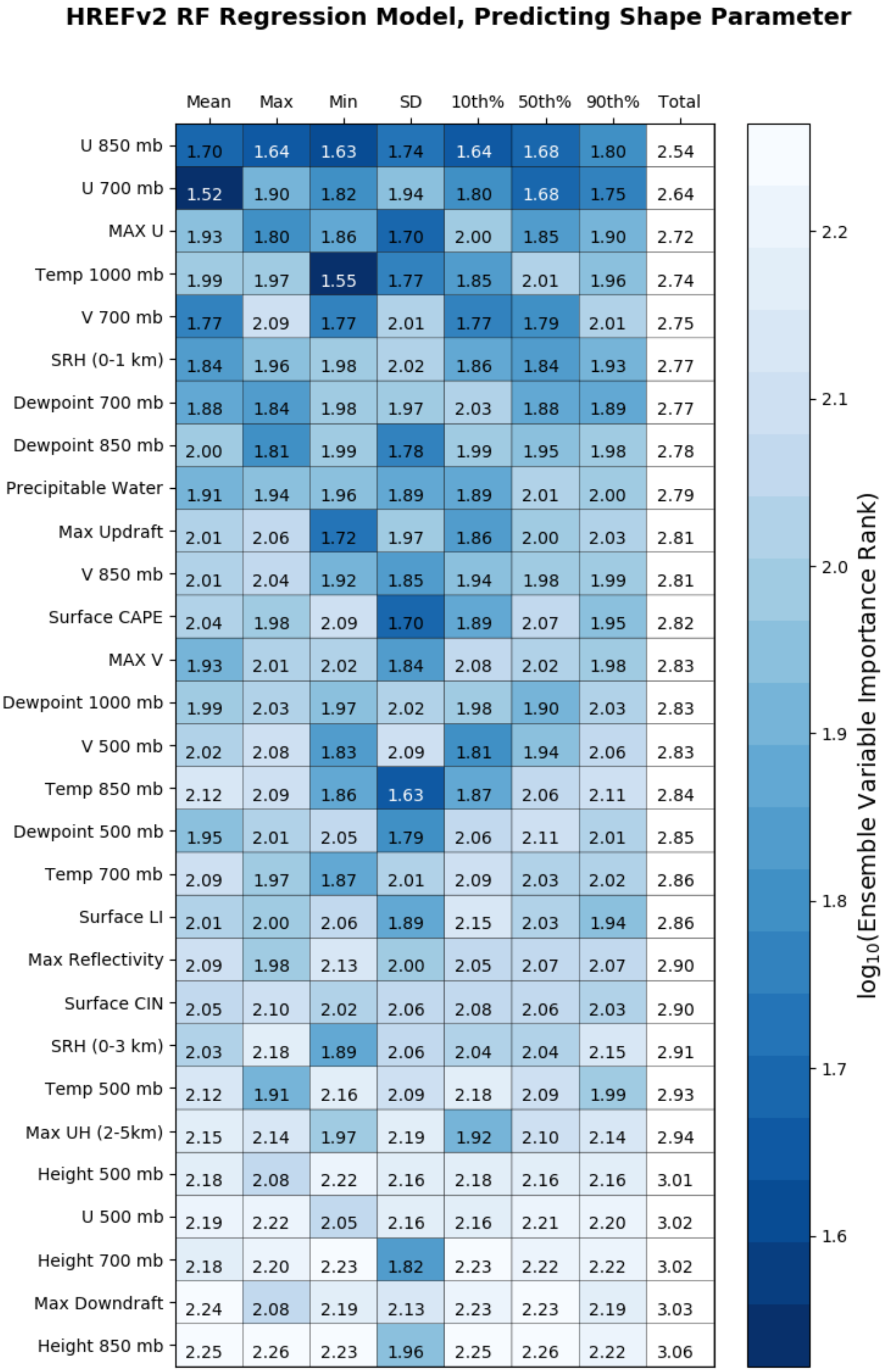
In general, there exists a relatively wide range of important variables for each step of the ML-based hail prediction process. Aspects of the wind field are important in all the ML models, with the greatest importance when predicting hail size. Instability and storm variables are most important when classifying the HREFv2 storm tracks.



HREFv2 RF Classification Model, Predicting Matched Parameter

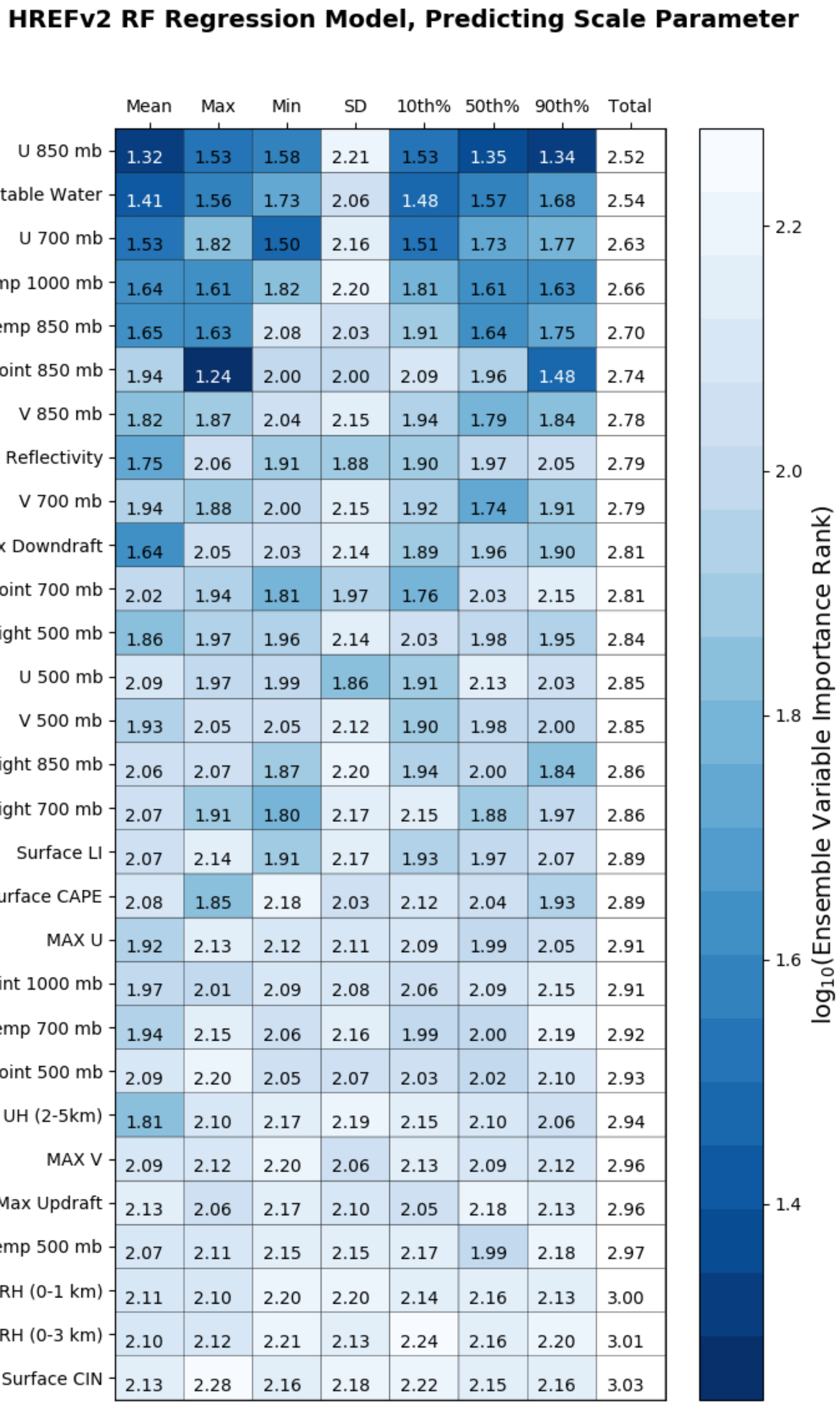
	Mean	Max	Min	SD	10th%	50th%	90th%	Total
Temp 850 mb	1.73	1.50	2.09	1.99	1.98	2.01	1.43	2.73
Surface LI	1.65	1.89	1.42	2.11	1.68	2.11	2.12	2.77
Height 500 mb	1.86	1.94	2.11	1.99	2.06	1.41	1.89	2.78
MAX V	1.96	1.80	1.91	1.88	2.05	2.07	1.80	2.78
Max Reflectivity	1.26	2.17	1.98	2.00	2.04	1.89	1.88	2.79
U 500 mb	2.01	1.75	2.03	1.98	2.00	2.05	1.78	2.80
Temp 1000 mb	1.91	1.76	2.13	1.97	1.99	2.06	1.85	2.81
Temp 500 mb	2.05	2.19	1.54	1.96	1.80	2.02	1.96	2.81
SRH (0-3 km)	2.00	1.85	1.98	1.97	2.00	2.13	1.81	2.82
V 700 mb	2.09	2.17	1.85	1.82	2.05	1.86	1.87	2.82
V 500 mb	2.02	1.96	2.04	2.00	1.96	1.89	1.98	2.83
Height 850 mb	1.77	2.11	1.91	2.07	1.78	1.86	2.19	2.83
Max Downdraft	1.73	2.08	1.83	2.03	2.13	2.03	1.99	2.84
Precipitable Water	1.80	1.98	2.10	2.05	1.75	2.07	2.10	2.84
Max Updraft	2.04	1.89	2.02	1.98	2.03	2.08	2.00	2.85
MAX U	1.89	2.02	1.99	2.09	2.02	2.11	1.96	2.86
Max UH (2-5km)	1.98	1.78	2.12	1.88	2.10	2.03	2.12	2.86
Dewpoint 1000 mb	2.07	2.08	2.04	2.03	1.89	2.04	2.02	2.87
V 850 mb	1.95	2.05	2.06	2.10	2.05	2.04	2.05	2.89
U 700 mb	2.01	2.04	2.16	1.98	2.08	2.10	1.88	2.89
SRH (0-1 km)	2.03	2.05	2.03	2.04	2.12	2.07	1.97	2.89
Dewpoint 700 mb	2.12	2.12	2.08	2.07	1.96	1.94	2.05	2.90
Height 700 mb	2.04	1.98	2.08	1.97	2.17	2.15	1.92	2.90
Dewpoint 500 mb	2.07	2.04	2.03	2.00	1.93	2.14	2.14	2.90
Temp 700 mb	2.10	2.08	2.08	2.04	2.02	2.04	2.09	2.91
U 850 mb	2.03	2.06	2.10	2.12	1.98	2.15	2.06	2.92
Dewpoint 850 mb	2.09	2.07	2.11	2.07	2.07	2.03	2.11	2.92
Surface CAPE	2.07	1.97	2.17	2.06	2.17	2.16	2.12	2.95
Surface CIN	2.08	2.15	2.19	2.18	1.99	2.15	2.16	2.98

Figure 5.9: Permutation variable importance ranks for a classification Random Forest predicting full CONUS HREFv2 storm objects as possible hail objects. Input variables include the 29 HREFv2 variables and the statistics applied to each storm object. Darker colors indicate greater feature importance.



	Mean	Max	Min	SD	10th%	50th%	90th%	Total
U 850 mb	1.70	1.64	1.63	1.74	1.64	1.68	1.80	2.54
U 700 mb	1.52	1.90	1.82	1.94	1.80	1.68	1.75	2.64
MAX U	1.93	1.80	1.86	1.70	2.00	1.85	1.90	2.72
Temp 1000 mb	1.99	1.97	1.55	1.77	1.85	2.01	1.96	2.74
V 700 mb	1.77	2.09	1.77	2.01	1.77	1.79	2.01	2.75
SRH (0-1 km)	1.84	1.96	1.98	2.02	1.86	1.84	1.93	2.77
Dewpoint 700 mb	1.88	1.84	1.98	1.97	2.03	1.88	1.89	2.77
Dewpoint 850 mb	2.00	1.81	1.99	1.78	1.99	1.95	1.98	2.78
Precipitable Water	1.91	1.94	1.96	1.89	1.89	2.01	2.00	2.79
Max Updraft	2.01	2.06	1.72	1.97	1.86	2.00	2.03	2.81
V 850 mb	2.01	2.04	1.92	1.85	1.94	1.98	1.99	2.81
Surface CAPE	2.04	1.98	2.09	1.70	1.89	2.07	1.95	2.82
MAX V	1.93	2.01	2.02	1.84	2.08	2.02	1.98	2.83
Dewpoint 1000 mb	1.99	2.03	1.97	2.02	1.98	1.90	2.03	2.83
V 500 mb	2.02	2.08	1.83	2.09	1.81	1.94	2.06	2.83
Temp 850 mb	2.12	2.09	1.86	1.63	1.87	2.06	2.11	2.84
Dewpoint 500 mb	1.95	2.01	2.05	1.79	2.06	2.11	2.01	2.85
Temp 700 mb	2.09	1.97	1.87	2.01	2.09	2.03	2.02	2.86
Surface LI	2.01	2.00	2.06	1.89	2.15	2.03	1.94	2.86
Max Reflectivity	2.09	1.98	2.13	2.00	2.05	2.07	2.07	2.90
Surface CIN	2.05	2.10	2.02	2.06	2.08	2.06	2.03	2.90
SRH (0-3 km)	2.03	2.18	1.89	2.06	2.04	2.04	2.15	2.91
Temp 500 mb	2.12	1.91	2.16	2.09	2.18	2.09	1.99	2.93
Max UH (2-5km)	2.15	2.14	1.97	2.19	1.92	2.10	2.14	2.94
Height 500 mb	2.18	2.08	2.22	2.16	2.18	2.16	2.16	3.01
U 500 mb	2.19	2.22	2.05	2.16	2.16	2.21	2.20	3.02
Height 700 mb	2.18	2.20	2.23	1.82	2.23	2.22	2.22	3.02
Max Downdraft	2.24	2.08	2.19	2.13	2.23	2.23	2.19	3.03
Height 850 mb	2.25	2.26	2.23	1.96	2.25	2.26	2.22	3.06

Figure 5.10: Similar to Fig. 5.9, now using a random forest regression model. The regression model predicts the shape parameter of a MESH gamma distribution for each storm object associated with hail from the random forest classification model



HREFv2 RF Regression Model, Predicting Scale Parameter

	Mean	Max	Min	SD	10th%	50th%	90th%	Total
U 850 mb	1.32	1.53	1.58	2.21	1.53	1.35	1.34	2.52
Precipitable Water	1.41	1.56	1.73	2.06	1.48	1.57	1.68	2.54
U 700 mb	1.53	1.82	1.50	2.16	1.51	1.73	1.77	2.63
Temp 1000 mb	1.64	1.61	1.82	2.20	1.81	1.61	1.63	2.66
Temp 850 mb	1.65	1.63	2.08	2.03	1.91	1.64	1.75	2.70
Dewpoint 850 mb	1.94	1.24	2.00	2.00	2.09	1.96	1.48	2.74
V 850 mb	1.82	1.87	2.04	2.15	1.94	1.79	1.84	2.78
Max Reflectivity	1.75	2.06	1.91	1.88	1.90	1.97	2.05	2.79
V 700 mb	1.94	1.88	2.00	2.15	1.92	1.74	1.91	2.79
Max Downdraft	1.64	2.05	2.03	2.14	1.89	1.96	1.90	2.81
Dewpoint 700 mb	2.02	1.94	1.81	1.97	1.76	2.03	2.15	2.81
Height 500 mb	1.86	1.97	1.96	2.14	2.03	1.98	1.95	2.84
U 500 mb	2.09	1.97	1.99	1.86	1.91	2.13	2.03	2.85
V 500 mb	1.93	2.05	2.05	2.12	1.90	1.98	2.00	2.85
Height 850 mb	2.06	2.07	1.87	2.20	1.94	2.00	1.84	2.86
Height 700 mb	2.07	1.91	1.80	2.17	2.15	1.88	1.97	2.86
Surface LI	2.07	2.14	1.91	2.17	1.93	1.97	2.07	2.89
Surface CAPE	2.08	1.85	2.18	2.03	2.12	2.04	1.93	2.89
MAX U	1.92	2.13	2.12	2.11	2.09	1.99	2.05	2.91
Dewpoint 1000 mb	1.97	2.01	2.09	2.08	2.06	2.09	2.15	2.91
Temp 700 mb	1.94	2.15	2.06	2.16	1.99	2.00	2.19	2.92
Dewpoint 500 mb	2.09	2.20	2.05	2.07	2.03	2.02	2.10	2.93
Max UH (2-5km)	1.81	2.10	2.17	2.19	2.15	2.10	2.06	2.94
MAX V	2.09	2.12	2.20	2.06	2.13	2.09	2.12	2.96
Max Updraft	2.13	2.06	2.17	2.10	2.05	2.18	2.13	2.96
Temp 500 mb	2.07	2.11	2.15	2.15	2.17	1.99	2.18	2.97
SRH (0-1 km)	2.11	2.10	2.20	2.20	2.14	2.16	2.13	3.00
SRH (0-3 km)	2.10	2.12	2.21	2.13	2.24	2.16	2.20	3.01
Surface CIN	2.13	2.28	2.16	2.18	2.22	2.15	2.16	3.03

Figure 5.11: Similar to Fig. 5.10, permutation variable importance is calculated using a random forest regression model to predict the scale parameter of a MESH gamma distribution.

5.5 Discussion

Performance of three ML-based hail prediction models was investigated over the CONUS, including subjective (case study) and objective (reliability, ETS, and bias) examinations as well as a model interpretation technique. Of the different ML models evaluated, the LSR calibrated model displayed the closest resemblance to the SPC day 1 hail outlook and practically perfect data, as well as the highest reliability towards the LSRs. However, the predictions displayed a high spatial bias especially at the significant severe (sig-severe) threshold. The over-forecasting of sig-severe hail is due in part to rarity of the event, and also the algorithm for creating hail size predictions. Predictions are produced by each HREFv2 member based off a distribution of storm object values (MAXUVV for this thesis) that is compared to a distribution of predicted MESH hail sizes. These predictions are then aggregated to create the NMEPs. However, predicting hail size from each member localizes the output such that sig-severe hail is rarely predicted. When an ensemble approach was investigated, in which hail was predicted from a distribution of ensemble storm objects rather than individual members, sig-severe hail forecasts exhibited improved reliability performance. However, the ensemble approach also increases predicted severe hail, leading to over-forecasting at that threshold. In short, an approach designed to produce optimal predictions of severe hail did not produce optimal sig-severe hail predictions, and vice versa. It is possible that developing two different models, one for predicting severe hail and another for sig-severe hail, could result in the best possible forecast performance.

Next, the ML hail predictions over the CONUS were evaluated through ETS and bias calculations. The LSR calibrated predictions had comparable skill and bias with the PP output at the severe hail threshold, regardless of the observational dataset used. The PP calibrated predictions demonstrated the lowest skill and bias across both observational datasets and hail thresholds. Additional evaluation included an updraft helicity (UH) proxy variable to compare the ML models against a non-ML baseline. The uncalibrated ML output had similar ETS values compared to the UH proxy at the severe threshold, while UH exhibited the highest skill of all the outputs at the sig-severe threshold. The greater UH forecast skill is most likely because ETS prioritizes hits, causing the UH field which spatially over-forecasts to exhibit the highest skill. Combining the ETS and bias information indicates that the ML models reduce false alarms rather than increase hits.

The last objective evaluation analyzed the LSR calibrated output across different forecast hours of interest to the SPC. The 1200-1200 UTC and 2100-0100 UTC time periods displayed the highest skill across both verification datasets. Greater performance with these two periods likely results from the higher prevalence of severe weather, including severe hail, during the afternoon and evening hours. Additionally, the increased performance associated with the later time periods may indicate that the LSR calibrated ML model exhibits an over-forecasting bias during early time periods.

Finally, a permutation variable importance algorithm determined relevant variables for CONUS ML-based hail prediction. When classifying HREFv2 storm objects, instability, 850 mb temperature, and 500 mb height fields were most important, most likely for their dynamical contributions to storm development. In fact, Allen et al. (2015) describes how the thermodynamic environment, especially steep mid-level lapse rates, are important for hail occurrence. In predicting the shape parameter, which affects the skewness of the MESH gamma distribution, the U wind field at differing levels was most important. Similarly, Dennis and Kumjian (2017) found that increased east-west deep layer shear leads to large hail growth. For the third ML model used to predict hail, precipitable water, U wind (850 and 700 mb), and various temperature fields (1000 and 850 mb) were found important in predicting the scale parameter. The relevant variables suggest that available moisture, instability, and shear all contribute to the range of hail sizes a storm will produce, again related to discussion of hail occurrence in Allen et al. (2015). Decreased storm variable importance within the ML models may result from the time period studied (forecast hours 12-36), where environmental features have a greater impact over small scale fields. Also, the lack of importance of MAXUVV indicates that while strong updrafts are an important basis for identifying storms, other factors are crucial for hail growth. In fact, hailstone trajectories critical for growth are affected more by the broadness and tilt of an updraft, rather than only the intensity (e.g., Weisman and Klemp, 1982; Nelson, 1983; Foote, 1984; Dennis and Kumjian, 2017).

Overall, the calibrated CONUS ML hail guidance provided operationally useful predictions, while highlighting physically relevant variables for hail prediction. All of the ML models over-forecast hail in terms of spatial extent, but this could benefit forecasters preparing forecasts in terms of highlighting potential hail risk areas and addressing storm placement uncertainty. The severe hail predictions were particularly skillful after calibration, which is encouraging for an initial application of the technique to the HREFv2 dataset. The calibrated predictions produce output that, compared to the uncalibrated predictions, better resemble a SPC hail outlook and practically perfect output. The SPC hail outlook resemblance indicates that the calibrated ML models focus on similar hail areas as forecasters, and may increase operational forecaster trust in the automated predictions.

Chapter 6

Regional ML Hail Forecasts

The algorithm for ML-based hail forecasting has shown success, especially the LSR calibrated model, on the CONUS grid. However, hail prediction can vary substantially within different regions of the CONUS. Hail producing storms in the southern plains form under different conditions than the southeast or northern plains, and vice versa (e.g., Frisby, 1963; Nelson and Young, 1979; Schaefer et al., 2004). Therefore, investigating ML models trained over smaller subdomains may produce hail forecasts with superior performance over the full CONUS-trained model. In this chapter, comparisons of localized models are performed in areas frequently impacted by hail, particularly severe hail. Several methods for regionalization were tested, beginning with a direct application of the 9 regions from the SPC HREF viewer.[1] As these regions are already used in SPC operations, they provide a natural basis for developing regional products. However, initial tests indicated this approach is unworkable, as limiting training to these regions with an already insufficient training dataset resulted in poor ML performance. Limiting the dataset to this extent caused the ML model to never predict hail.

Another tested method divided the CONUS into three separate areas: the West (West Coast to approximately the Rocky Mountains), Central (Rocky Mountains to the Mississippi River), and East (Mississippi River to the East Coast). While the larger sectors improved the ML model performance compared to the mesoscale regions, again

[1] https://www.spc.noaa.gov/exper/href/

hail storms in South Dakota form under different environmental conditions than those in Texas. Therefore, it might not be optimal to combine these areas in a single region.

To address this issue, regions were defined from areas that receive high impact-hail, as defined by the SPC climatology of all hail and sig-severe hail between 1982 to 2011.[2] The regions include the northern plains, central plains, southern plains, and the southeast (Fig. 6.1a). The southeast region was chosen to evaluate the ML model because of the lack of climatologically large hail, compared to the plains, but still constitutes an area impacted by hail events. The northern plains sector includes Montana, Wyoming, North Dakota, South Dakota, and Minnesota. The central plains region encompasses Utah, Wyoming, Colorado, Kansas, Nebraska, Iowa, and Missouri. Rather than splitting Wyoming in half between the northern and central plains regions, it was included twice for this initial regional analysis. The southern plains sector covers New Mexico, Texas, Oklahoma, Arkansas, and Louisiana. Finally, the southeast sector includes Mississippi, Alabama, Tennessee, North Carolina, South Carolina, Georgia, and Florida. Each sector's states were chosen because they experience climatologically larger hail, and also surround a city that experiences relatively high-impact hail (Fig. 6.1b). The high impact cities include Rapid City, SD; Denver, CO; Dallas, TX; and Atlanta, GA for the northern plains, central plains, southern plains, and southeast regions, respectively.

The process for training and testing the regional ML models is described in section 6.1. Similar to verification of the CONUS-trained ML models, the ML models tested over each region are verified both subjectively and objectively. A case study (section 6.2) is examined within each region, while objective evaluations are performed using reliability (section 6.3.1), equitable threat score (ETS)/bias (section 6.3.2), and statistical significance testing (section 6.4). The regional model(s) with the greatest

[2] https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=allHail

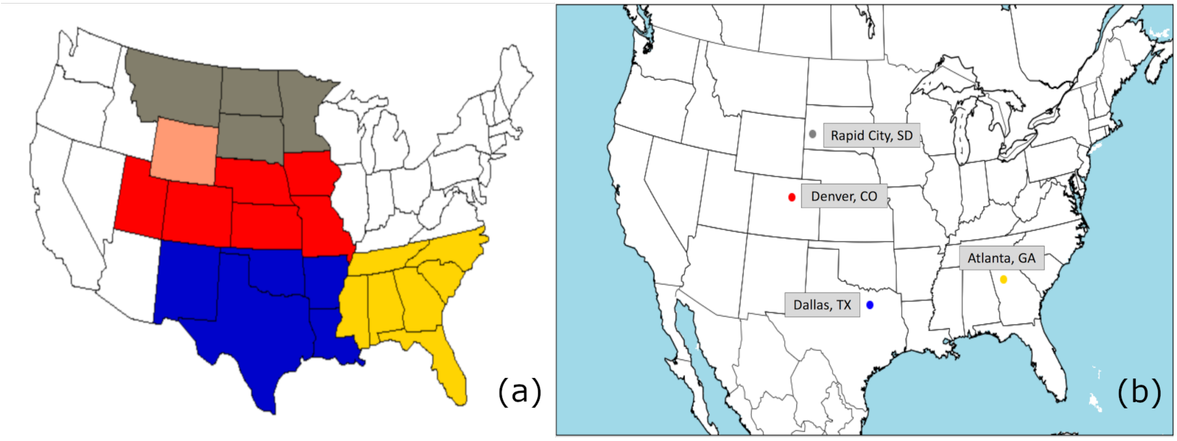


Figure 6.1: Maps displaying (a) the regions for testing the regional ML models, and (b) the center points for training the regional ML models. Regions include the northern plains (gray), central plains (red), southern plains (blue), and southeast (yellow). Wyoming (pink) is included in both the northern and central plains regions.

performance are investigated using permutation variable importance, to examine any additional information the model(s) provide (section 6.5)). Finally, a discussion of the hail prediction results from localized ML modeling is included in section 6.6.

6.1 Machine Learning Process

The high-impact cities mentioned above act as center points to train the regional ML models. Only the storms within each region, defined by the state boundaries for simplicity as a first analysis, are tested by the ML model. Limiting training data to a small domain has proven unsuccessful from preliminary tests mentioned above. Therefore each storm within the CONUS is included in the training dataset, however the storms are weighted based on their proximity to the centerpoint of a region. An exponential decay function weights the storm distances such that,

$$weights = \exp(\frac{-Distances}{2}) * 10$$

The storms at larger distances from the regional centerpoint are less important for training the ML models, causing the ML models to place greater importance on features commonly responsible for severe hail events in the areas of interest. Figure 6.2 is a visual representation of the weights applied to each storm for a given region.

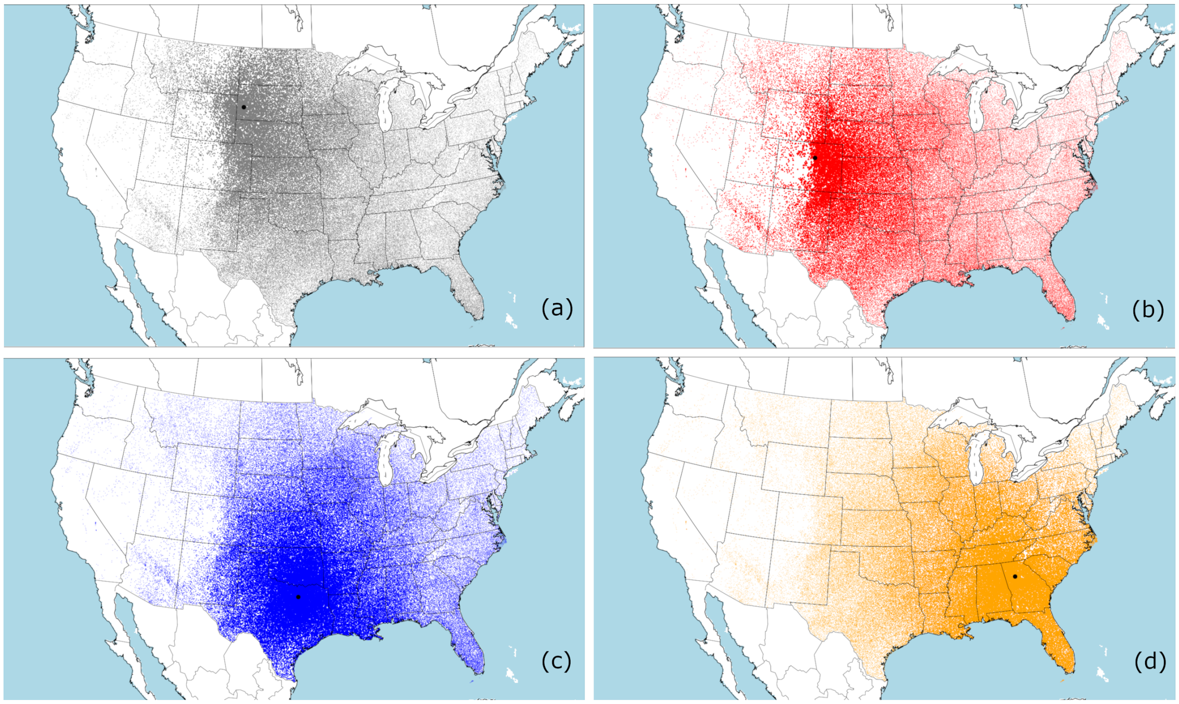


Figure 6.2: Identified HREFv2 storm objects, where the storm size in each region represents the weight received during training. Regional maps include the (a) northern plains, (b) central plains, (c) southern plains, and (d) southeast.

Differing from the CONUS-trained model, the regional storm weights are added when training the RF models and only the storms within each region's states are tested. Both the regional and CONUS models are evaluated, to investigate differences in performance and determine whether localized modeling result in superior regional hail forecasts. Each region is trained over the same 2017 season as the CONUS-trained model and tested over the 2018 season. However, the days that evaluate the performance of the regional and CONUS-trained models vary per region, given the available hail days within the testing set for each sector. For comparison between training type, the hail forecasts are not calibrated, to focus on differences in resulting regionalizations. Also, only the severe hail threshold is examined due to the limitation in dataset size for both training and testing at the significant severe threshold. A larger dataset, encompassing multiple years, could be sufficient for evaluating the significant severe threshold.

6.2 Regional Case Studies

In the northern plains sector, an upper level ridge developed over the high plains and Canadian prairies on 8 June 2018, with a surface trough advecting dewpoints in the mid 60s. A combination of cooler air aloft and warm surface air created a highly unstable profile, leading to severe hail production despite a lack of environmental wind shear. The PP forecast for 8 June (Fig. 6.3a) displayed probabilities of severe hail up to 22% over western South Dakota, surrounded by a larger area of 2 to 14%. Smaller areas of up to 14% were located on the North Dakota/Minnesota border and over northern Iowa.

Both models, CONUS and sector trained, predict non-zero probabilities of severe hail from northeastern Wyoming to southwestern South Dakota. However, the sector model subjectively decreases the false alarm area of values greater than 15% that were associated with a cluster of hail reports in this area (Fig. 6.3b,c). Comparatively, the CONUS model (Fig. 6.3b) has a larger area of probabilities exceeding 15%, with slight eastern displacement of the south central South Dakota hail threat. Additionally, the

CONUS model misses reports in North Dakota/Minnesota while the sector model has a closed probability contour. Both models pick up on the hail reports in northern Iowa. Overall, the regional model subjectively outperforms the CONUS model in terms of severe hail forecast skill for this case.

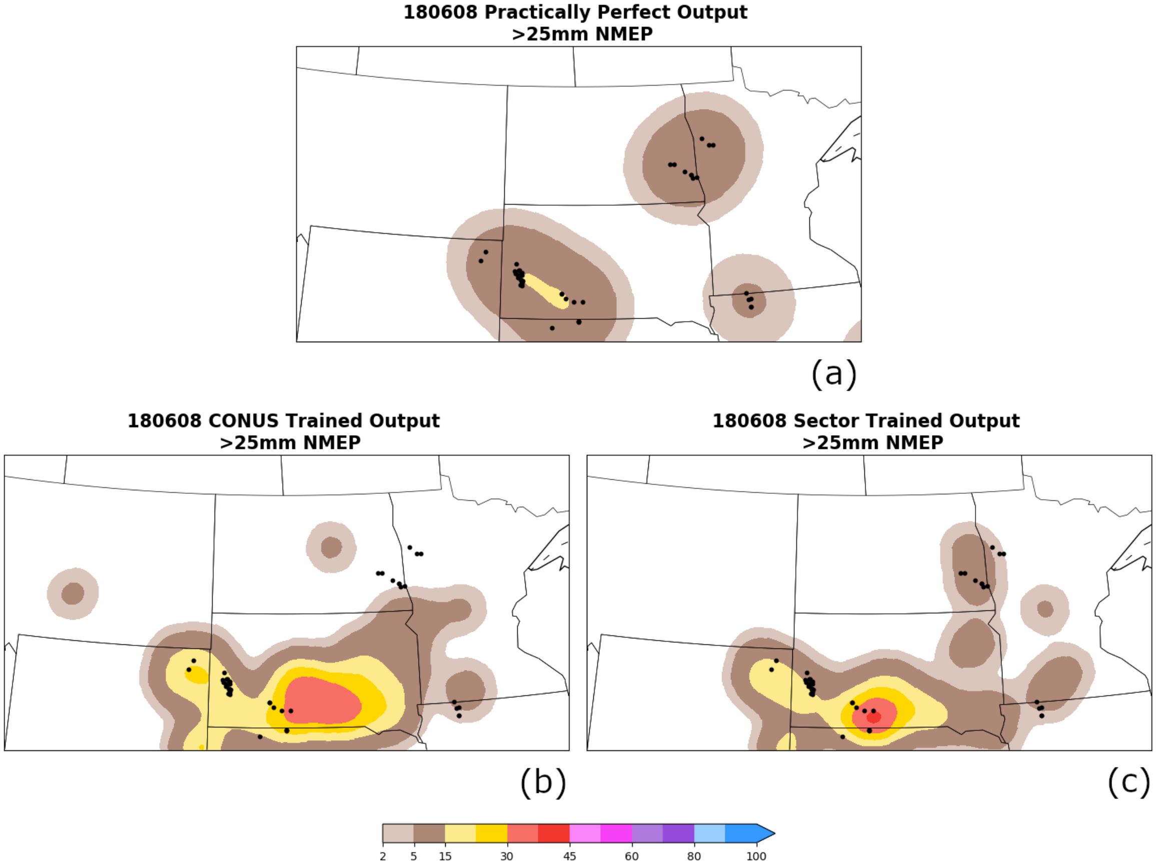

180608 Practically Perfect Output
>25mm NMEP
(a)
180608 CONUS Trained Output
>25mm NMEP
(b)
180608 Sector Trained Output
>25mm NMEP
(c)
2
5
15
30
45
60
80
100

Figure 6.3: Northern plains case study for ML-based regional severe hail prediction, valid 8 June 2018. Maps include the (a) practically perfect output, (b) ML model trained on unweighted CONUS storms, and (c) regional model trained on weighted CONUS storms.

In the central plains case, the synoptic and mesoscale setup was previously identified as the high-end hail case for the CONUS-trained hail forecasts. The reader is referred to section 5.2 for the environmental setup on 27 July 2018 in eastern Colorado. The case study was chosen again to further analyze the forecast and compare the CONUS and regional model performance. The PP output displays the highest probabilities over east-central Colorado, exceeding 30%.

The CONUS and regional model outputs in the central plains exhibit similar forecast probability magnitudes and areal coverage. Small differences include a smaller area of false alarms in the CONUS-trained model (Fig. 6.4b) over Missouri. In addition, the sector model (Fig. 6.4c) outputs lower probabilities over central Nebraska, up to 30% compared to the CONUS model's 37%, where there are no PP values. Lastly, the sector model outputs tighter contours in eastern Colorado, decreasing the false alarm area for probabilities between 15% to 60%. After 60% both models predict the same hail threat area in eastern Colorado. Overall, there is little subjective difference between the two model outputs for this case in the central plains.

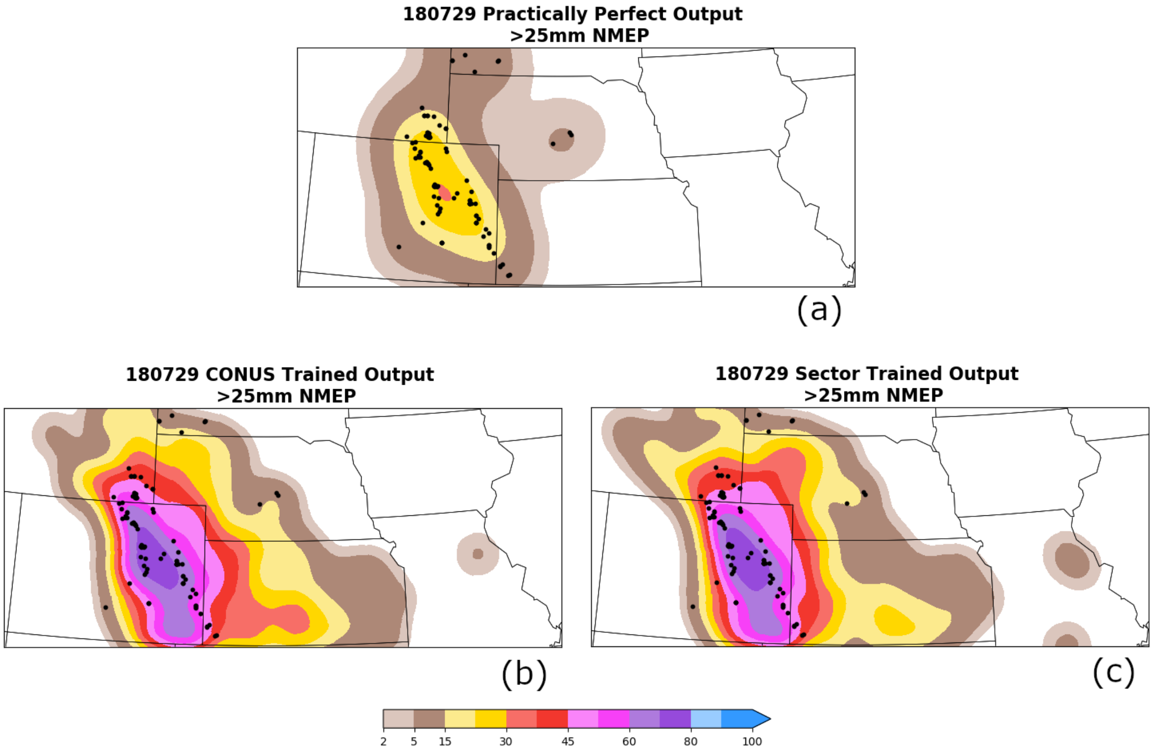

180729 Practically Perfect Output
>25mm NMEP
(a)
180729 CONUS Trained Output
>25mm NMEP
(b)
180729 Sector Trained Output
>25mm NMEP
(c)
2
5
15
30
45
60
80
100

Figure 6.4: Central plains regional case study, valid 29 July 2018. Maps include the (a) practically perfect output, (b) ML model trained on unweighted CONUS storms, and (c) regional model trained on weighted CONUS storms.

On 2 May 2018 in the southern plains region, a shortwave trough accompanied by a dryline extending from the central to southern plains, set up ascent in the Texas panhandle, western Oklahoma, and Big-Bend area of Texas. Despite areas of convective inhibition, strong moisture return and a discrete storm mode, which favors supercell development and in turn large hail, contributed to severe hail production. The PP output valid 2 May 2018 (Fig. 6.5a) displays two areas of up to 22% probability of severe hail over southwestern Oklahoma and the Oklahoma panhandle. Probabilities up to 14% are displayed over the Big-Bend and also extending from north-central Texas northward through Oklahoma.

The CONUS (Fig. 6.5b) and sector (Fig. 6.5c) trained model forecasts in the southern plains region exhibit the greatest subjective differences of all the regions, both in terms of probability magnitudes and area. The regional model predicts lower probabilities over eastern Texas, resulting in reduced false alarms. Additionally, the sector model overall outputs smaller probability areas exceeding 30%, again decreasing false alarms compared to the PP output. Both models highlight southwestern Texas to Western Oklahoma as areas of hail threat, while also displacing higher probabilities for the southern hail threat over the Big Bend. In general, the regional model reduces false alarms compared to the CONUS-trained model, while still accurately predicting the severe hail threat in areas where it was observed.

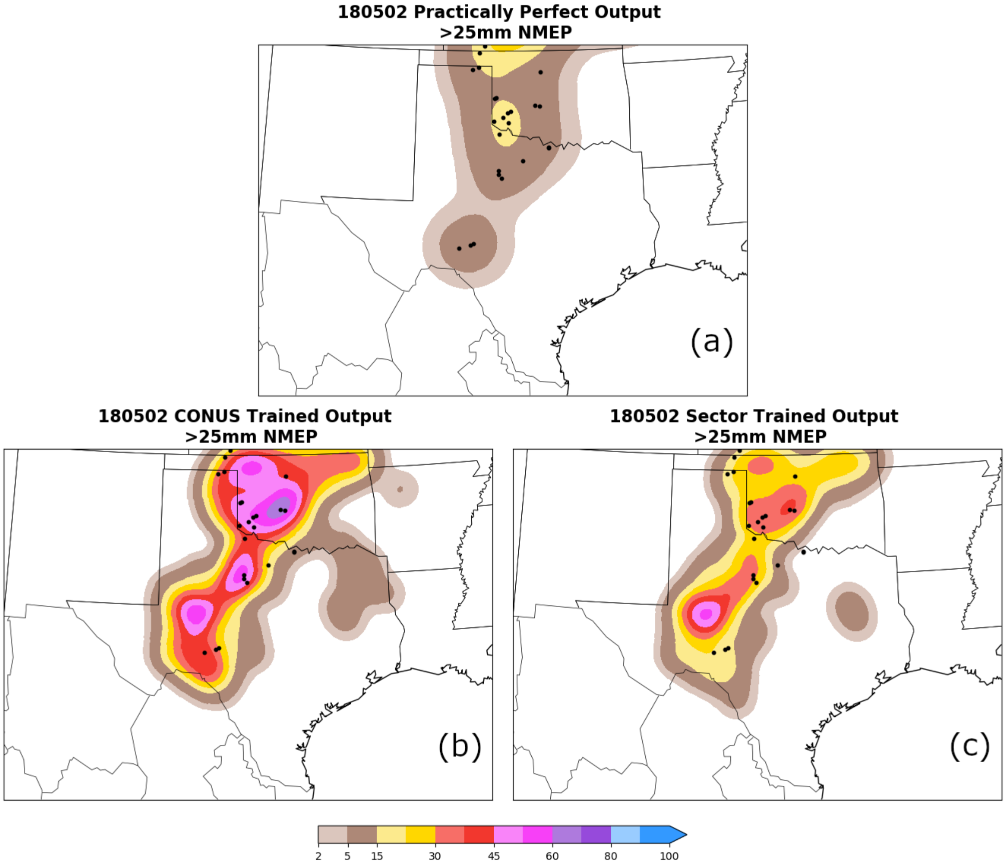

180502 Practically Perfect Output
>25mm NMEP
(a)
180502 CONUS Trained Output
>25mm NMEP
(b)
180502 Sector Trained Output
>25mm NMEP
(c)
2
5
15
30
45
60
80
100

Figure 6.5: Southern plains regional case study, valid 2 May 2018. Maps include the (a) practically perfect output, (b) ML model trained on unweighted CONUS storms, and (c) regional model trained on weighted CONUS storms.

Lastly, an upper-level ridge over the southern plains brought northwesterly flow into the mid-Mississippi valley, creating an unstable profile for the southeast region on 21 July 2018. A shortwave trough responsible for storms early in the morning over eastern Tennessee/Kentucky, as well as dew points in the lower 70's °F, led to scattered thunderstorm development. The PP output (Fig. 6.6a) for 21 July displays probabilities up to 29% on the Alabama/Georgia border and up to 22% extending from the same border northeastward to North Carolina.

Similar to the central plains region, there are few subjective differences between the CONUS (Fig. 6.6b) and sector (Fig. 6.6c) trained models for this region's case study. Both models predict the greatest severe hail threat over southwestern Alabama, while hail reports occurred in eastern Alabama up into North Carolina. The sector model slightly decreases the false alarm area of probabilities over Florida and Georgia, as well as decreases probabilities in northern Georgia and South Carolina. However, differences between the two models are minimal, and both over-forecast the coverage and magnitude of severe hail.

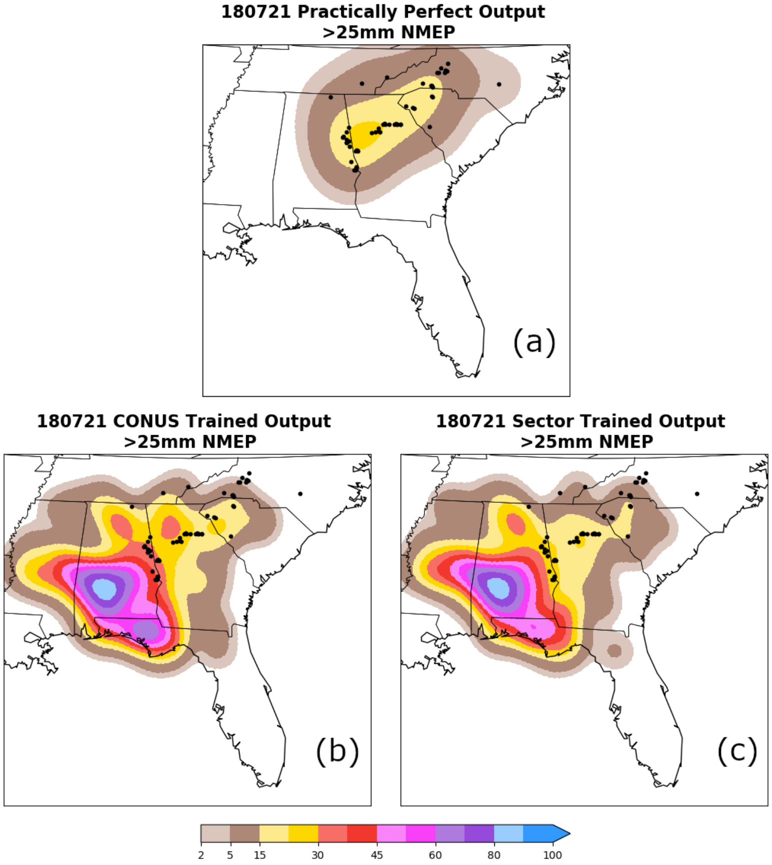

180721 Practically Perfect Output
>25mm NMEP
(a)
180721 CONUS Trained Output
>25mm NMEP
(b)
180721 Sector Trained Output
>25mm NMEP
(c)
2
5
15
30
45
60
80
100

Figure 6.6: Southeast regional case study, valid 21 July 2018. Maps include the (a) practically perfect output, (b) ML model trained on unweighted CONUS storms, and (c) regional model trained on weighted CONUS storms.

6.3 Quantitative Verification

6.3.1 Reliability

In addition to the subjective evaluations presented above, the CONUS and regional ML models are analyzed quantitatively, first using reliability diagrams with LSRs as the observational dataset (Fig. 6.7). In the northern plains sector (Fig. 6.7a), both models are reliable up to about 5%. For higher forecast probabilities, the models over-predict hail. There is little difference in reliability between the two models until 60%, where the sector model slightly over-forecasts compared to the CONUS model then becomes more reliable above 80%. In addition to the few subjective difference in reliability, the northern plains sector model achieves a Brier Skill Score (BSS) of -0.15, barely surpassing the CONUS model at -0.16.

The central plains models (Fig. 6.7b) also exhibit near-perfect reliability up to 5%, after which both models over-forecast with respect to the LSRs. There is very little difference in reliability between the two models, which is also shown in the BSS (both -0.13). Although the central and northern plains models display little difference in training type, the central plains forecasts are generally more reliable than the northern plains.

Next, the southern plains models exhibit the greatest difference in reliability (Fig. 6.7c). The majority of forecast probabilities indicate a more reliable sector model, with the exception around 70% where the CONUS model becomes slightly more reliable. Differences also exists with the BSS, which is higher for the sector model (-0.09) compared to the CONUS-trained model (-0.36).

Finally, the sector-trained model in the southeast region is near-perfectly reliable up to 20% above-which it begins over-forecasting (Fig. 6.7d). Both models over-forecast probabilities greater than 20%, however the sector model is more reliable between 40%

to 60%. Regarding model performance, the sector model displays a slightly higher BSS value (-0.12) than the CONUS model (-0.15). It should be noted that the southeast forecasts had the greatest over-forecasting bias of all the regions.

Among the four regional models, the southern plains trained model exhibits the greatest subjective difference from the full CONUS-trained model. The southeast region demonstrates the second greatest difference in model type while the northern and central plains models are subjectively similar.

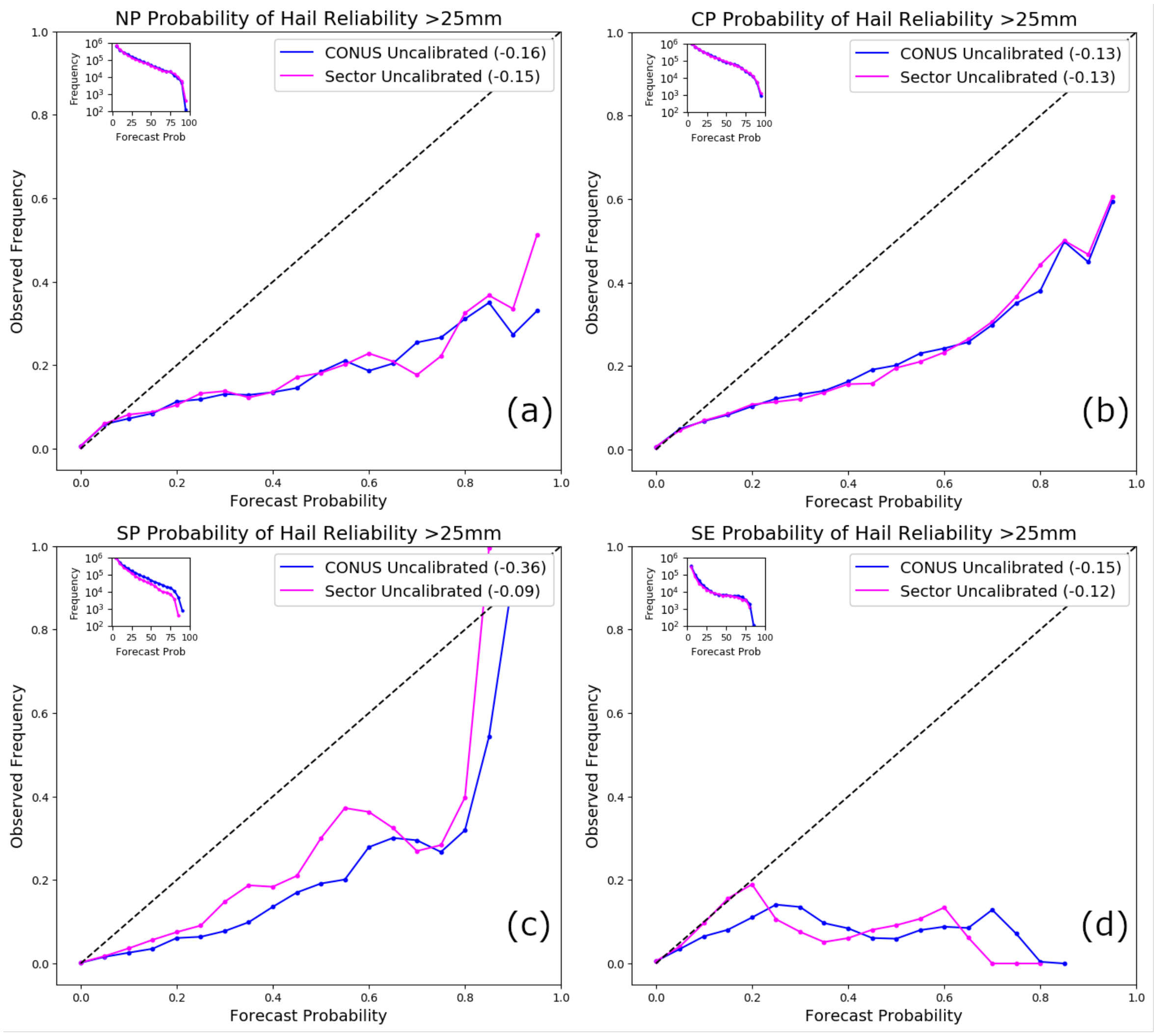


Figure 6.7: Reliability diagrams evaluating the (a) northern plains, (b) central plains, (c) southern plains, and (d) southeast. Both the unweighted CONUS and weighted sector trained ML models are evaluated against the local storm reports as observations. The legend displays the Brier Skill Score across all forecast probabilities.

6.3.2 Equitable Threat Score and Bias

In addition to reliability, ETS and bias are calculated against the LSRs and PP output to evaluate the regional hail forecasts. In the northern plains sector, the two models are nearly identical in skill when compared to both the LSRs (Fig. 6.8a) and PP output (Fig. 6.8c). Values of ETS, calculated using the LSRs, are maximized (0.12) for both the regional and CONUS-trained models at 25% forecast probability. Comparatively, the two models achieve an ETS value of 0.27 when the PP output is used as observations. The model biases are also nearly identical when compared to the LSRs (Fig. 6.8b) and PP output (Fig. 6.8d). The sector model shows slightly lower bias than the CONUS model between 10 and 30%, with the LSRs as truth. Any differences are diminished when the forecasts are compared to the PP output. Overall, there is very little difference in skill between the two model types in the northern plains sector, except for a slight decrease in bias of the sector model when compared against the LSRs.

For the central plains sector (Fig. 6.9), a similar pattern emerges regarding model similarity. The CONUS model achieves slightly greater skill (0.14 at 45%) when calculated against the LSRs (Fig. 6.9a). Larger forecast probabilities show the sector model as minutely more skillful. Using the PP output as observations, the CONUS-trained model is slightly more skillful with a maximum score of 0.26 at 2% (Fig. 6.9c). Any model difference is decreased when evaluating bias against the LSRs (Fig. 6.9b) and PP output (Fig. 6.9d). Compared to the northern plains sector, there is even less difference in ETS and bias between model type for the central plains region.

Next, the southern plains region once again presents the greatest subjective difference in model type of all the regions (Fig. 6.10). The sector model achieves the highest skill (0.16 at 35%) when calculated against the LSRs (Fig. 6.10c), before becoming less skillful at higher probabilities. Even though the sector model is less skillful after about 40%, it displays lower bias across all output probabilities, compared to the CONUS-trained model (Fig. 6.10b). When evaluated against the PP output, the sector model has the highest skill (0.19 at 2%) across all output probabilities (Fig. 6.10c). Similarly, the sector model exhibits lower bias across all probabilities when calculated with the PP output (Fig. 6.10d). In general, the sector model outperforms the CONUS-trained model in the southern plains region.

Lastly, the southeast region's sector model achieves the highest ETS skill (0.07) at 10% probability when using the LSRs as observations. Above 10%, the regional model decreases in skill compared to the CONUS -trained model (Fig. 6.11a). The models have similar bias, producing values closest to PP among all four regions (Fig. 6.11b), with the regional model exhibiting slightly lower bias up to 20%. Using the PP output as observations, the sector-trained model achieves the highest ETS value (0.14 at 2%), and decreases in skill compared to the CONUS model above 15% (Fig. 6.11c). The sector model exhibits slightly lower bias than the CONUS model when analyzed against the PP output (Fig. 6.11d). Of all the regions, the southeast models contain bias values closest to 1 when compared to the PP. Overall, few differences exist between the trained models in the southeast region, however the differences are not as great as the southern plains.

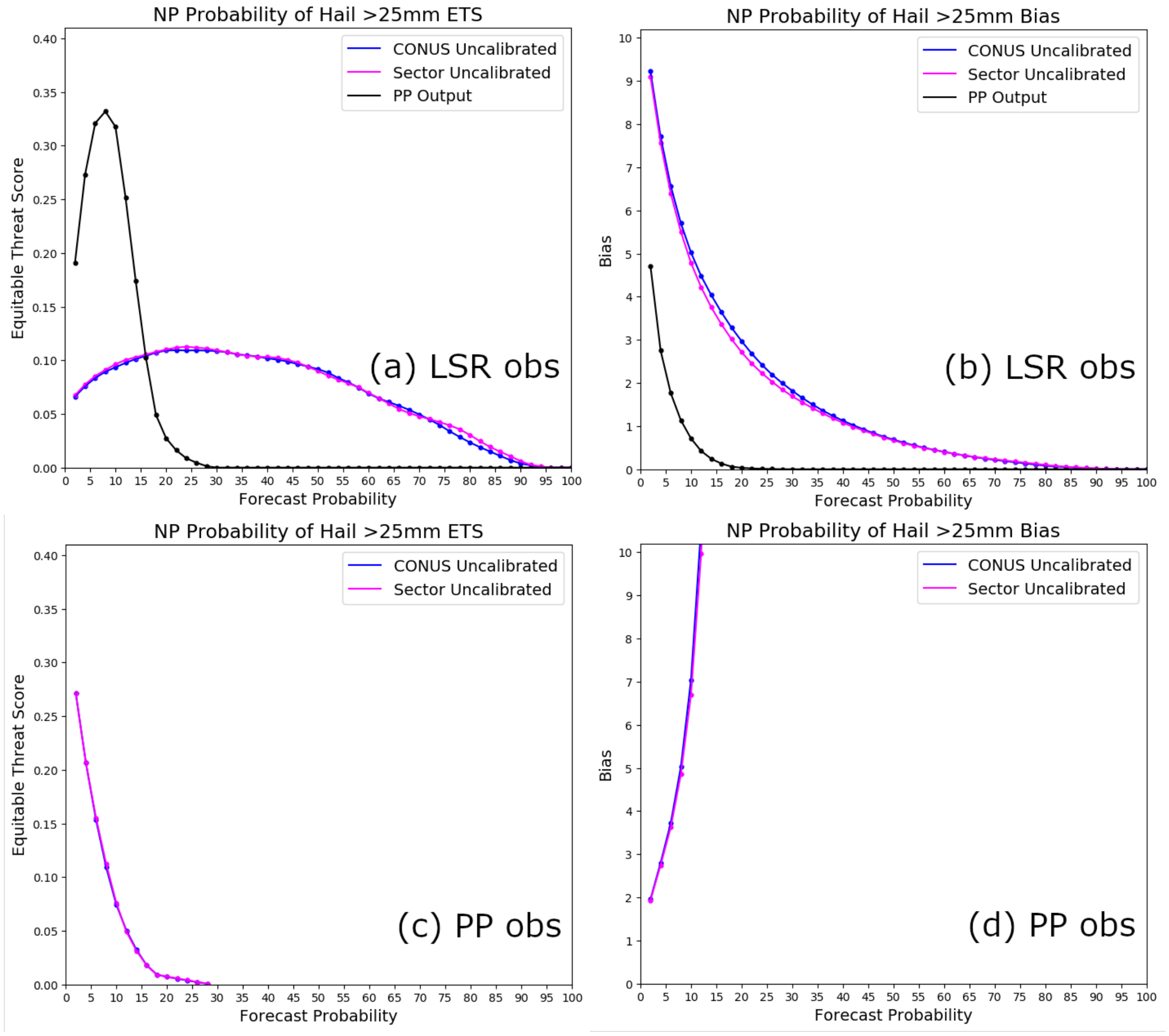


Figure 6.8: Plots of equitable threat score (ETS) and bias examining the central plains regional ML hail forecasts. The observational dataset comprises of the (a,b) local storm reports and (c,d) practically perfect output.

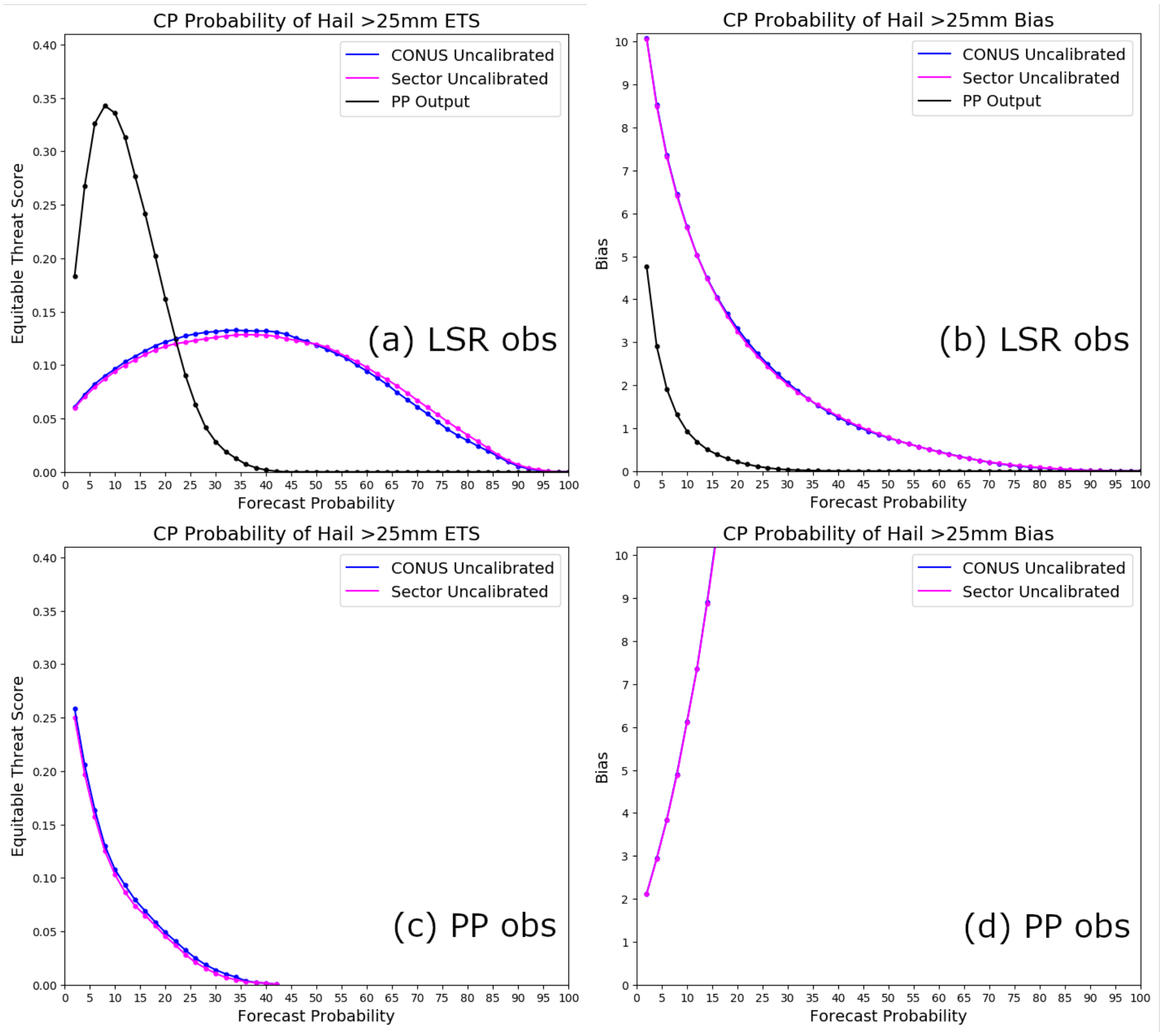


Figure 6.9: Plots of equitable threat score (ETS) and bias examining the central plains regional ML hail forecasts. Observational datasets are similar to Fig. 6.8.

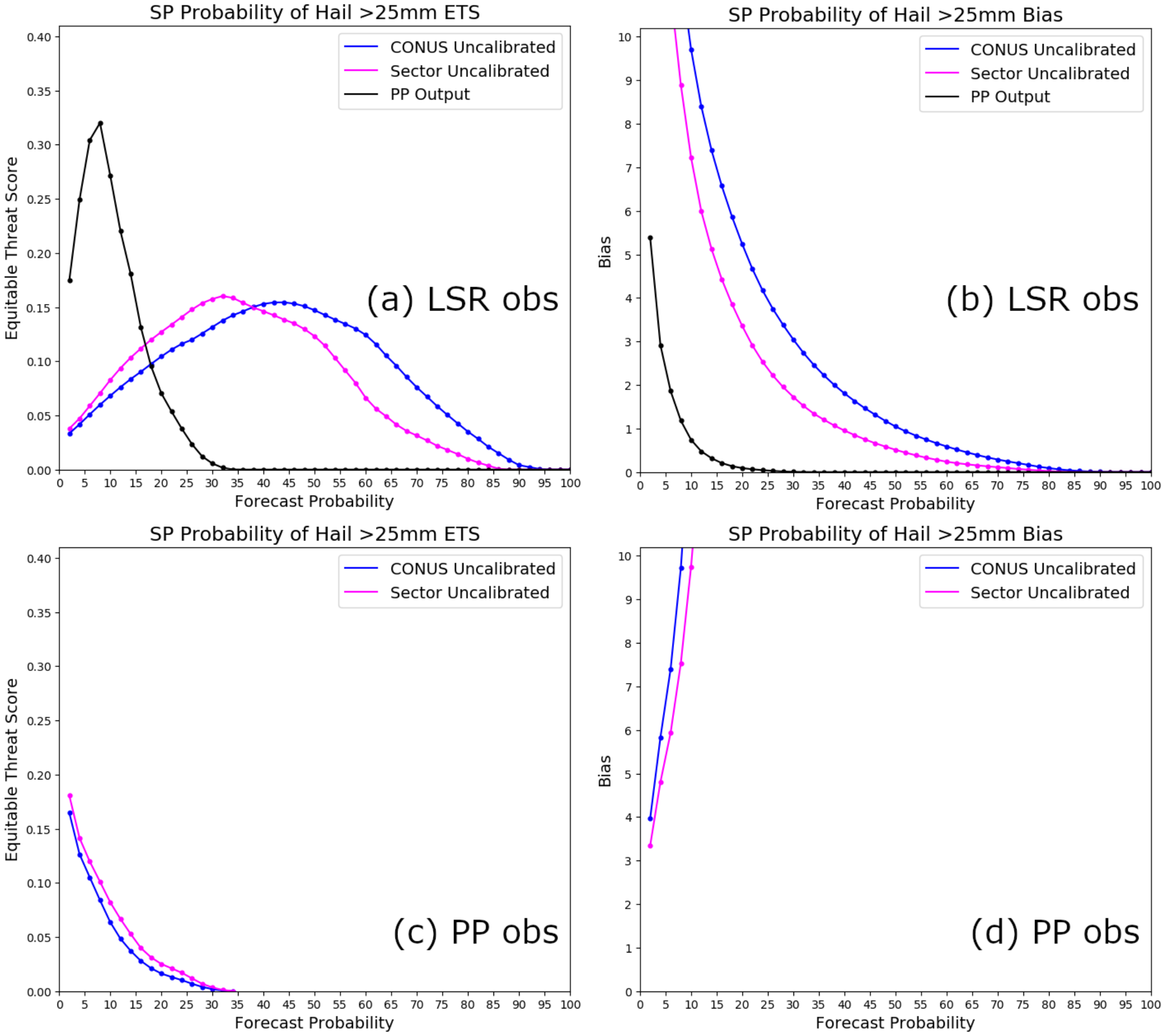


Figure 6.10: Plots of equitable threat score (ETS) and bias examining the southern plains regional ML hail forecasts. Observational datasets are also similar to Fig. 6.8.

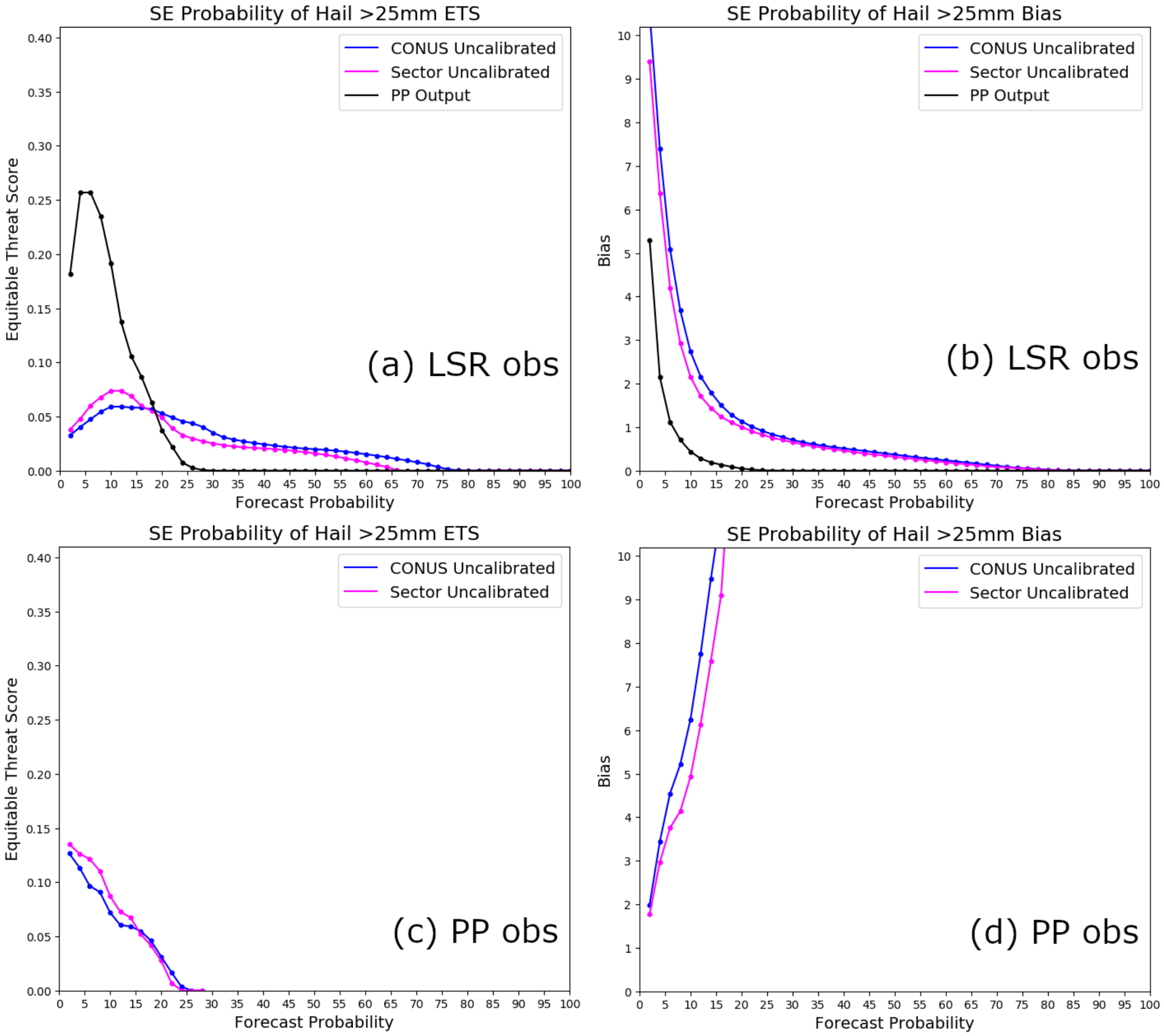


Figure 6.11: Plots of equitable threat score (ETS) and bias examining the southeast regional ML hail forecasts. Observational datasets are similar to Fig. 6.8.

6.4 Statistical Significance Testing

To evaluate significant differences in performance between the regional and CONUS-trained models, differences in ETS and bias are tested. A two-tail paired resampling test with an alpha of 0.05 investigates the null hypothesis, or the performance distributions of the CONUS and regional models in each region are not significantly different.

The algorithm for calculating performance significance is based on Hamill (1999), with applications from Schwartz and Sobash (2017). Only the LSR dataset is tested against as observational truth because any significant differences in predicting the LSRs also impact the PP dataset, based on its definition. For the significance algorithm, the CONUS and sector models are investigated from 1 May to 31 August 2018, with lag correlation tests identifying any serial correlations within the regional ML datasets. The northern plains region displays independence for daily sampling, every-other day sampling is required for the southern and central plains regions, and the southeast sector needs four days in between samples. Each forecast probability threshold undergoes significance testing with 1000 resampling iterations of the adjusted datasets.

To determine the p-value, differences in performance are first calculated between the CONUS and regional models. Then, performance is evaluated where each day within the test set is randomly chosen to assess either the CONUS or regional models. Performance is also calculated over a second dataset, consisting of the opposite model chosen each day. To determine statistical significance, differences in skill from the two randomized datasets are compared to the difference in skill from the original CONUS and regional models. The p-value is determined where the absolute differences of the random dataset metric scores are greater than the difference of the original score. A p-value ≤ 0.05 delineates statistically significant differences in performance.

The results of the significance testing are presented in Figure 6.12. It should be noted, all of the regions exhibit high p-values at higher forecast probabilities because the two models typically do not output probabilities above 90%. In the northern plains region (Fig. 6.12a), only one forecast probability (80%) indicates ETS is significantly different between the two models. For bias, values between between 10 and 30% are significantly different, occurring with smaller sector model output. In the central plains (Fig. 6.12b), bias is not significantly different between the two model types, however

ETS values are significantly different between 70 and 80%. Said probabilities are associated with greater sector model skill, compared to the CONUS-trained model output.

The majority of southern plains sector probabilities (Fig. 6.12c) display significant differences in both performance metrics. The sector model exhibits higher ETS values where significant differences exist between 5 and 30%, while the CONUS-trained model exhibits significantly greater ETS values for probabilities 55 to 75%. Significant differences in bias up to 80% occur where the regional model outputs lower bias values. Finally, in the southeastern region (Fig. 6.12d), significant differences in ETS exist at 5% forecast probability, coinciding with greater sector-trained model skill. However, all other probabilities do not exhibit significant differences in either skill score between the ML models.

Overall, the southern plains region outputs the highest number of forecast probabilities with statistically significant performance differences. All of the models indicate some significant difference in at least one performance metric across the forecast probabilities, with the majority of differences occurring where the sector trained model outperforms the CONUS trained.

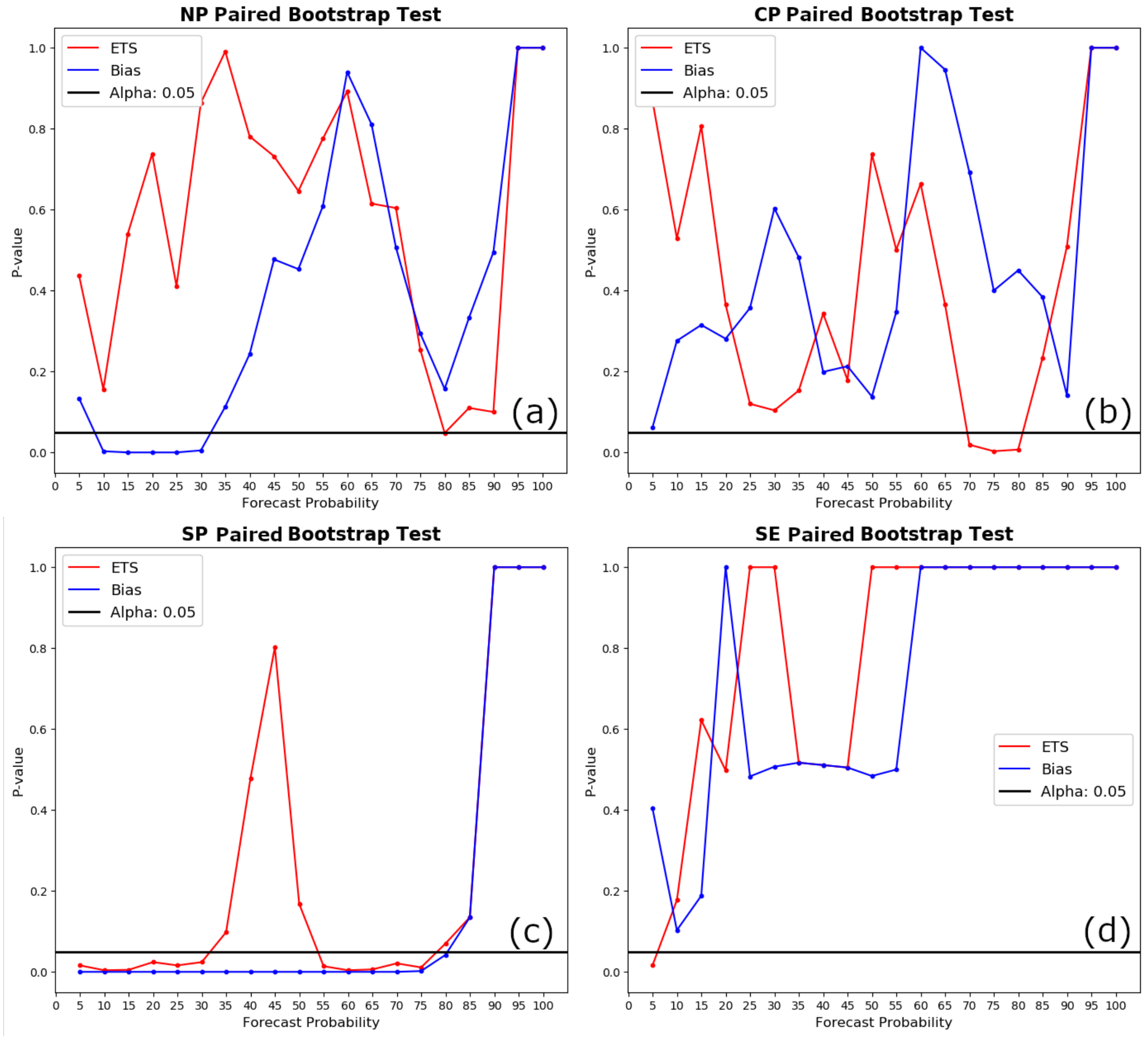


Figure 6.12: Plots of two-tail resampled significance tests, examining equitable threat score (ETS) and bias p-values at varying forecast probabilities. Regions analyzed include the (a) northern plains, (b) central plains, (c) southern plains, and (d) southeast.

6.5 Model Interpretation

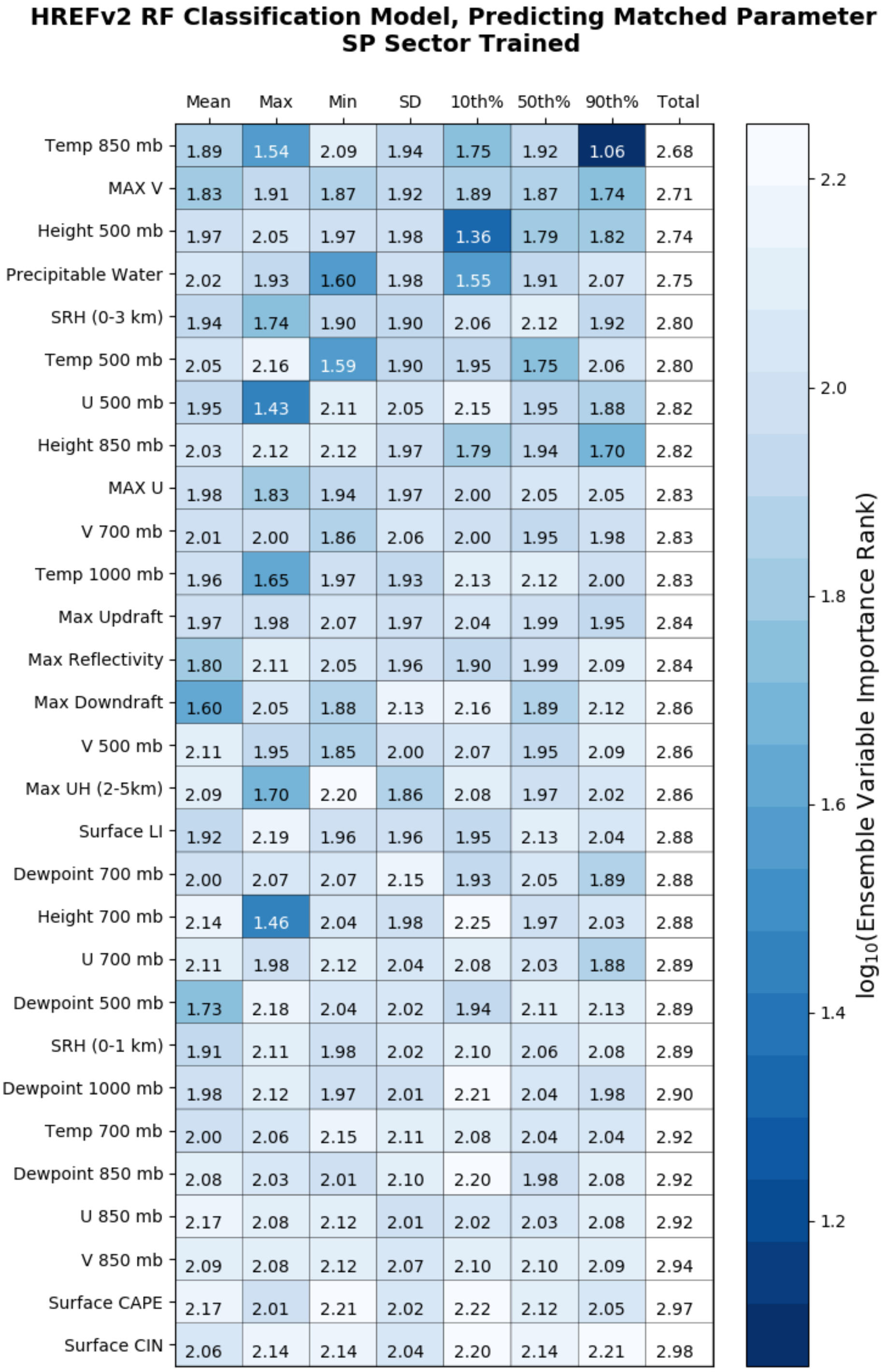
Only the southern plains region is evaluated using permutation variable importance because of the performance differences discovered between the regional and CONUS-trained models. The other regions lack subjectively substantial differences in performance, possibly due to the limited datasets for training and testing. Chapter 4 includes a description of the importance algorithm, now pertaining to regional data. The process for regional hail prediction can be found in section 6.1, where the importance algorithm is evaluated over the same training (2017) and testing data (2018).

The first step of the ML-based hail prediction algorithm classifies HREFv2 storm objects for hail association. The most important variables for classification in the southern plains sector are the 850 mb temperature, maximum hourly V wind, and 500 mb height fields (Fig. 6.13), similar to the CONUS-trained model (Fig. 5.9). However, precipitable water and 0-3km SRF increase in importance in the southern plains region, whereas the maximum hourly reflectivity and surface lifted index decrease. The most important field in the regionally trained model is the 90th percentile of the 850 mb temperature field, followed by the 10th percentile of the 500 mb height field. Overall, both the regional and CONUS-trained classification models feature environmental variables as important, however the regional model highlights different dynamical aspects.

The RF regression model predicting the shape variable in the southern plains region (Fig. 6.14) distinguishes the U wind fields as important, again similar to the CONUS-trained models (Fig. 5.10). However, moisture fields including precipitable water and dewpoint (1000 and 850 mb) are more relevant in the southern plains. Also differing from the CONUS-trained model, the standard deviation of the 700 mb temperature is most important 2D variable, compared to the 700 mb U wind in the CONUS model. The second most important regional variable is the median 850 mb U wind. Generally, to predict the shape parameter, the southern plains model emphasizes moisture and temperature fields.

When predicting the scale parameter, the regional (Fig. 6.15) and CONUS-trained (Fig. 5.11) models are most similar in their total variable rankings. Both models focus on precipitable water, U wind (850 and 700 mb), and the 1000 mb temperature. Precipitable water is slightly more important for the sector model, as well as surface CAPE and 0-3km SRH. The most important variable within both models is the maximum 850 mb dewpoint, however the southern plains model highlights the 90th percentile 1000 mb temperature as second important over the CONUS-trained's mean 850 mb U wind. Beyond the most important variables, slight differences exist in the rankings of environmental and storm variables within the regional model predicting the scale parameter.

In general, moisture and temperature variables are more important for hail prediction in the southern plains, compared to the full CONUS. The greatest difference in relevant variables for prediction occurs with classifying the HREFv2 storm objects, while the most similarities exist when predicting the shape parameter of the gamma MESH distribution.



	Mean	Max	Min	SD	10th%	50th%	90th%	Total
Temp 850 mb	1.89	1.54	2.09	1.94	1.75	1.92	1.06	2.68
MAX V	1.83	1.91	1.87	1.92	1.89	1.87	1.74	2.71
Height 500 mb	1.97	2.05	1.97	1.98	1.36	1.79	1.82	2.74
Precipitable Water	2.02	1.93	1.60	1.98	1.55	1.91	2.07	2.75
SRH (0-3 km)	1.94	1.74	1.90	1.90	2.06	2.12	1.92	2.80
Temp 500 mb	2.05	2.16	1.59	1.90	1.95	1.75	2.06	2.80
U 500 mb	1.95	1.43	2.11	2.05	2.15	1.95	1.88	2.82
Height 850 mb	2.03	2.12	2.12	1.97	1.79	1.94	1.70	2.82
MAX U	1.98	1.83	1.94	1.97	2.00	2.05	2.05	2.83
V 700 mb	2.01	2.00	1.86	2.06	2.00	1.95	1.98	2.83
Temp 1000 mb	1.96	1.65	1.97	1.93	2.13	2.12	2.00	2.83
Max Updraft	1.97	1.98	2.07	1.97	2.04	1.99	1.95	2.84
Max Reflectivity	1.80	2.11	2.05	1.96	1.90	1.99	2.09	2.84
Max Downdraft	1.60	2.05	1.88	2.13	2.16	1.89	2.12	2.86
V 500 mb	2.11	1.95	1.85	2.00	2.07	1.95	2.09	2.86
Max UH (2-5km)	2.09	1.70	2.20	1.86	2.08	1.97	2.02	2.86
Surface LI	1.92	2.19	1.96	1.96	1.95	2.13	2.04	2.88
Dewpoint 700 mb	2.00	2.07	2.07	2.15	1.93	2.05	1.89	2.88
Height 700 mb	2.14	1.46	2.04	1.98	2.25	1.97	2.03	2.88
U 700 mb	2.11	1.98	2.12	2.04	2.08	2.03	1.88	2.89
Dewpoint 500 mb	1.73	2.18	2.04	2.02	1.94	2.11	2.13	2.89
SRH (0-1 km)	1.91	2.11	1.98	2.02	2.10	2.06	2.08	2.89
Dewpoint 1000 mb	1.98	2.12	1.97	2.01	2.21	2.04	1.98	2.90
Temp 700 mb	2.00	2.06	2.15	2.11	2.08	2.04	2.04	2.92
Dewpoint 850 mb	2.08	2.03	2.01	2.10	2.20	1.98	2.08	2.92
U 850 mb	2.17	2.08	2.12	2.01	2.02	2.03	2.08	2.92
V 850 mb	2.09	2.08	2.12	2.07	2.10	2.10	2.09	2.94
Surface CAPE	2.17	2.01	2.21	2.02	2.22	2.12	2.05	2.97
Surface CIN	2.06	2.14	2.14	2.04	2.20	2.14	2.21	2.98

Figure 6.13: Permutation variable importance scores for a random forest classification model predicting HREFv2 storm objects for hail association in the southern plains region. The model is trained using Dallas, TX as a centerpoint. Input variables are the same as in Fig. 5.9.

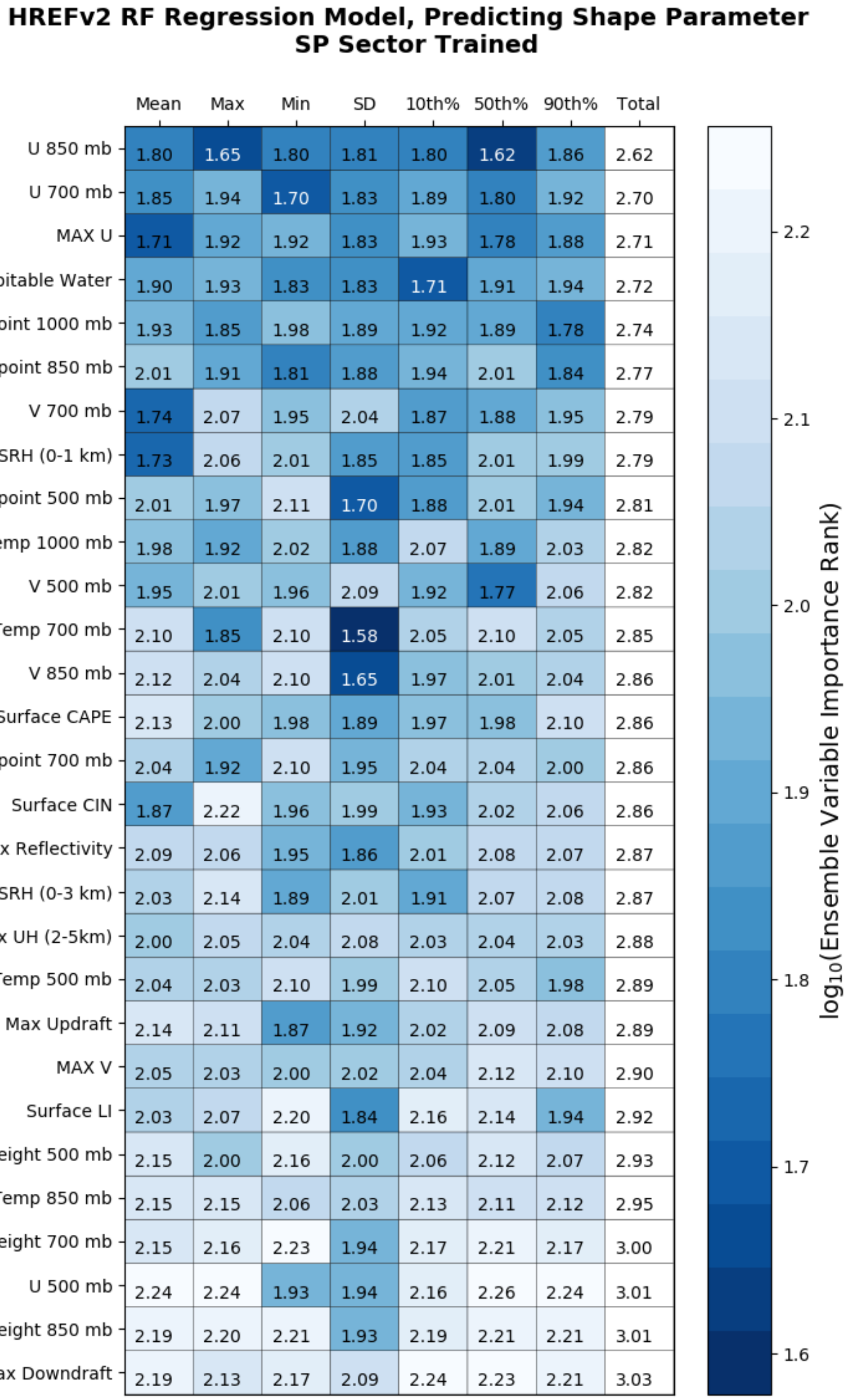
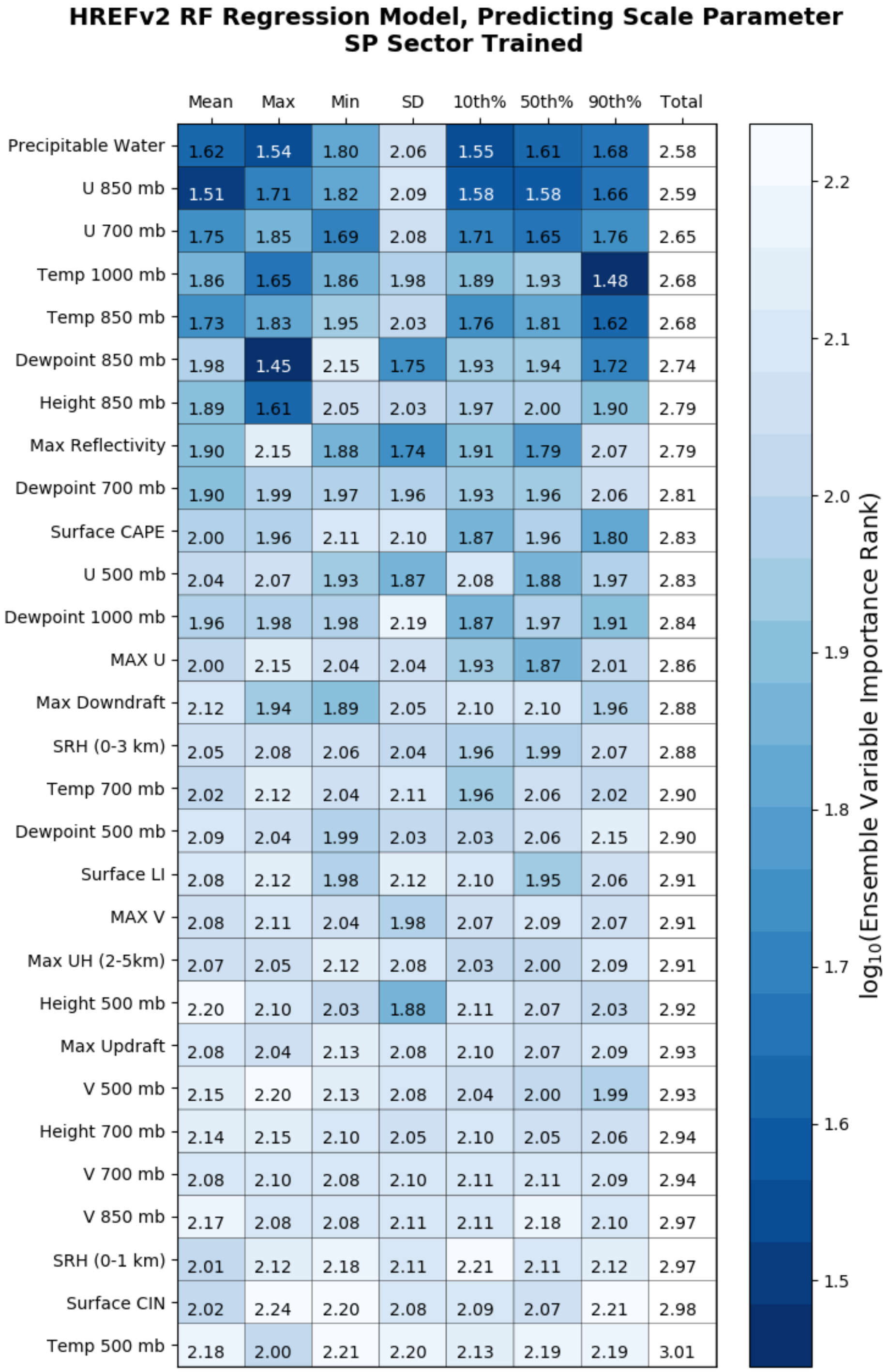


Figure 6.14: As in Fig. 6.13, now evaluating a random forest regression model predicting the shape parameter of a MESH gamma distribution for hail storm objects within the southern plains.



HREFv2 RF Regression Model, Predicting Scale Parameter
SP Sector Trained

	Mean	Max	Min	SD	10th%	50th%	90th%	Total
Precipitable Water	1.62	1.54	1.80	2.06	1.55	1.61	1.68	2.58
U 850 mb	1.51	1.71	1.82	2.09	1.58	1.58	1.66	2.59
U 700 mb	1.75	1.85	1.69	2.08	1.71	1.65	1.76	2.65
Temp 1000 mb	1.86	1.65	1.86	1.98	1.89	1.93	1.48	2.68
Temp 850 mb	1.73	1.83	1.95	2.03	1.76	1.81	1.62	2.68
Dewpoint 850 mb	1.98	1.45	2.15	1.75	1.93	1.94	1.72	2.74
Height 850 mb	1.89	1.61	2.05	2.03	1.97	2.00	1.90	2.79
Max Reflectivity	1.90	2.15	1.88	1.74	1.91	1.79	2.07	2.79
Dewpoint 700 mb	1.90	1.99	1.97	1.96	1.93	1.96	2.06	2.81
Surface CAPE	2.00	1.96	2.11	2.10	1.87	1.96	1.80	2.83
U 500 mb	2.04	2.07	1.93	1.87	2.08	1.88	1.97	2.83
Dewpoint 1000 mb	1.96	1.98	1.98	2.19	1.87	1.97	1.91	2.84
MAX U	2.00	2.15	2.04	2.04	1.93	1.87	2.01	2.86
Max Downdraft	2.12	1.94	1.89	2.05	2.10	2.10	1.96	2.88
SRH (0-3 km)	2.05	2.08	2.06	2.04	1.96	1.99	2.07	2.88
Temp 700 mb	2.02	2.12	2.04	2.11	1.96	2.06	2.02	2.90
Dewpoint 500 mb	2.09	2.04	1.99	2.03	2.03	2.06	2.15	2.90
Surface LI	2.08	2.12	1.98	2.12	2.10	1.95	2.06	2.91
MAX V	2.08	2.11	2.04	1.98	2.07	2.09	2.07	2.91
Max UH (2-5km)	2.07	2.05	2.12	2.08	2.03	2.00	2.09	2.91
Height 500 mb	2.20	2.10	2.03	1.88	2.11	2.07	2.03	2.92
Max Updraft	2.08	2.04	2.13	2.08	2.10	2.07	2.09	2.93
V 500 mb	2.15	2.20	2.13	2.08	2.04	2.00	1.99	2.93
Height 700 mb	2.14	2.15	2.10	2.05	2.10	2.05	2.06	2.94
V 700 mb	2.08	2.10	2.08	2.10	2.11	2.11	2.09	2.94
V 850 mb	2.17	2.08	2.08	2.11	2.11	2.18	2.10	2.97
SRH (0-1 km)	2.01	2.12	2.18	2.11	2.21	2.11	2.12	2.97
Surface CIN	2.02	2.24	2.20	2.08	2.09	2.07	2.21	2.98
Temp 500 mb	2.18	2.00	2.21	2.20	2.13	2.19	2.19	3.01

Figure 6.15: Similar to Fig. 6.14, permutation variable importance evaluates a random forest regression model predicting the scale parameter of a MESH gamma distribution.

6.6 Discussion

The four regions analyzed for superiority of a regional model over the CONUS-trained model were the northern plains, central plains, southern plains, and southeast. Of the four regions, the southern plains model exhibited the greatest difference between a regionally trained and CONUS-trained model. A majority of the hail forecast probabilities output in the southern plains exhibited significantly greater ETS and bias performance, while also improving Brier Skill Score (BSS) and reliability. Subjectively, a case study indicated that the sector-trained model outputs more realistic probability magnitudes and decreases false alarms with respect to both the LSRs and PP dataset.

Examination of the southern plains regional model using permutation variable importance identified 850 mb temperature, maximum hourly V wind, 500 mb heights, and precipitable water as the most important variables in identifying hailstorms. Similar to the CONUS-trained model, the ranked variables contribute to severe storm development. However, the temperature and moisture importance distinguishes the southern plains model, where greater low-level instability and available moisture within the hail growth zone is highlighted. Predicting the shape parameter also emphasizes moisture fields, along with differing levels of U wind. While the CONUS-trained model also highlights shear, similar to results from Dennis and Kumjian (2017), greater importance of deep-layer moisture characterizes hail size occurrence predictions in the southern plains. Finally, in determining the range of output hail sizes per storm object, the southern plains model stresses high 1000 mb temperatures in predicting the scale parameter. Increased importance of a thermodynamic field in the southern plains region highlights the influence of instability on hail production (Allen et al., 2015), most likely resulting from enhanced elevated mixed layers (EML) in the southern plains. Overall, the regional model places greater importance on moisture and temperature fields, rather than the wind field.

In the other regions, few subjective, objective, and statistical differences exist between the regional and CONUS-trained model. The southeastern model exhibited slight increases in performance at lower probability thresholds, especially with regards to reliability. However, the improvement did not yield large statistical differences between models. The success of the southern plains model, and partly the southeastern model, could be due to the greater number of hail cases within each region. Therefore, a larger training set of hail storms within each region could improve results in the remaining sectors. Additionally, verifying against a dataset without population biases could impact the differing forecast evaluations. Verifying against a larger dataset, in terms of areal coverage and dataset size, could improve the regional ML model's overall performance.

Chapter 7

Conclusions

A random forest machine learning (ML) method for day-ahead hail prediction, based on that of Gagne et al. (2017), was examined for predictability with data from HREFv2 and observations from the maximum expected size of hail (MESH) dataset. In general, favorable conditions for large hail growth include vertical wind shear to develop broad and sustained updrafts, available supercooled liquid, and horizontal trajectories of hailstones within the main updraft region (e.g., Heymsfield, 1982; Nelson, 1983; Foote, 1984; Miller et al., 1990; Dennis and Kumjian, 2017). The first part of this thesis explored hail prediction over the CONUS, where the ML models emphasized deep layer shear and other environmental variables as important for hail prediction, similar to the fields mentioned above. An additional step calibrated the ML hail predictions towards the LSRs and PP output, using isotonic regression. Calibration increased the reliability and skill of severe hail predictions, while also producing guidance similar to human-generated forecasts in terms of probability magnitudes, making the predictions more likely to be trusted by operational forecasters. The severe PP calibrated predictions were conservative in predicting severe hail, not exceeding 20% forecast probability, while the the LSR calibrated predictions did not exceed 40%. Subjective case study evaluations indicated that the LSR calibrated model most closely resembled the operational day 1 SPC hail outlook and PP output at the severe hail threshold.

In the second part of this thesis, regional domains were analyzed for superior performance of a localized ML model over the CONUS-trained model. Of the four regions, chosen for their prevalence of climatologically large hail, the southern plains regional forecasts exhibited the greatest objective, subjective, and statistical differences compared to the CONUS-trained ML output. Deep layer moisture and low-level temperatures were considered most important in the southern plains, highlighting instability and overall moisture fields compared to the CONUS-trained model. The other regional models displayed few differences when compared to the CONUS-trained ML model. Lack of highly weighted hailstorms could be the cause of the decreased regional performances, compared to the southern plains model. Further regional analyses that employ different storm weighting functions, and add additional years of data, could improve the performance of the other regional models.

Beyond greater data inclusion, verification and calibration using MESH observations could improve the overall performance of the ML hail prediction algorithm. Using a function of MESH derived in Murillo and Homeyer (2019) could improve hail forecasts, as they investigated a large hail dataset and calculated the 95 th percentile of hail sizes, for a better MESH approximation. Finally, Gagne et al. (2019) found that convolutional neural nets (CNNs) improved skill over statistical ML-based hail predictions. Differing from RFs, CNNs directly evaluate spatial information rather than statistical data approximations. Further exploration may reveal that CNNs can improve the spatial biases associated with the RF hail predictions, while bypassing the need for object-tracking.

Reference List

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within wrf. *Mon. Wea. Rev.*, **144 (12)**, 4919–4939, doi:10.1175/MWR-D-16-0027.1.

Aligo, E., B. Ferrier, J. Carley, E. Rogers, M. Pyle, S. Weiss, and I. Jirak, 2014: Modified microphysics for use in high-resolution NAM forecasts. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/27SLS/webprogram/Paper255732.html.].

Allen, J. T., M. K. Tippett, Y. Kaheil, A. H. Sobel, C. Lepore, S. Nong, and A. Muehlbauer, 2017: An extreme value model for u.s. hail size. *Monthly Weather Review*, **145**, 4501–4519, doi:10.1175/MWR-D-17-0119.1.

Allen, J. T., M. K. Tippett, and A. H. Sobel, 2015: An empirical model relating u.s. monthly hail occurrence to large-scale meteorological environment. *J. Adv. Model. Earth Syst.*, **7 (1)**, 226–243, doi:10.1002/2014MS000397.

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

Breiman, L., and P. Spector, 1992: Submodel selection and evaluation in regression. The X-random case. *International Statistical Review Revue Internationale de Statistique*, **60**, 291–319, doi:10.2307/1403680.

Brooks, H. E., C. A. Doswell, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the united states. *Wea. Forecasting*, **18**, 626–640, doi:10.1175/1520-0434(2003)018⟨0626:CEOLDT⟩2.0.CO;2.

Brooks, H. E., C. A. Doswell, and R. B. Wilhelmson, 1994: The role of midtropospheric winds in the evolution and maintenance of low-level mesocyclones. *Mon. Wea. Rev.*, **122**, 126–136, doi:10.1175/1520-0493(1994)122⟨0126:TROMWI⟩2.0.CO;2.

Brown, M. E., 2002: The spatial, temporal, and thermodynamic characteristics of southern-atlantic united states tornado events. *Physical Geography*, **23**, 401–417, doi:10.2747/0272-3646.23.5.401.

Cecil, D. J., and C. B. Blankenship, 2012: Toward a global climatology of severe hailstorms as estimated by satellite passive microwave imagers. *J. Climate*, **25**, 687–703, doi:10.1175/JCLI-D-11-00130.1.

Changnon, S. A., 1999: Data and approaches for determining hail risk in the contiguous united states. *J. Appl. Meteor.*, **38**, 1730–1739, doi:10.1175/1520-0450(1999)038⟨1730:DAAFDH⟩2.0.CO;2.

Changnon, S. A., and D. Changnon, 2000: Long-term fluctuations in hail incidences in the united states. *Journal of Climate*, **13**, 658–664, doi:10.1175/1520-0442(2000)013⟨0658:LTFIHI⟩2.0.CO;2.

Charba, J. P., and F. G. Samplatsky, 2011: Regionalization in fine-grid gfs mos 6-h quantitative precipitation forecasts. *Mon. Wea. Rev.*, **139**, 24–38, doi:10.1175/2010MWR2926.1.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An Objective High-Resolution Hail Climatology of the Contiguous United States. *Wea. Forecasting*, **27**, 1235–1248, doi:10.1175/WAF-D-11-00151.1.

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

Clark, A. J., and Coauthors, 2016: Preliminary findings and results-spring forecasting experiment 2016, [Available online at https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE_2016_preliminary_findings_final.pdf.].

Davis, C., and F. Carr, 2000: Summary of the 1998 Workshop on Mesoscale Model Verification. *Bull. Amer. Meteor.*, **81**, 809–819, doi:10.1175/1520-0477(2000)081⟨0809:SOTWOM⟩2.3.CO;2.

Dennis, E. J., and M. R. Kumjian, 2017: The impact of vertical wind shear on hail growth in simulated supercells. *J. Atmos. Sci.*, **74**, 641–663, doi:10.1175/JAS-D-16-0066.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, **129 (10)**, 2461–2480, doi:10.1175/1520-0493(2001)129⟨2461:AOAPMS⟩2.0.CO;2.

Foote, G. B., 1984: A study of hail growth utilizing observed storm conditions. *J. Climate Appl. Meteor.*, **23**, 84–101, doi:10.1175/1520-0450(1984)023⟨0084:ASOHGU⟩2.0.CO;2.

Frisby, E. M., 1963: Hailstorms of the upper great plains of the united states. *Journal of Applied Meteorology*, **2 (6)**, 759–766, doi:10.1175/1520-0450(1963)002⟨0759:HOTUGP⟩2.0.CO;2.

Gagne, D. J., 2016: Coupling Data Science Techniques and Numerical Weather Prediction Models for High-Impact Weather Prediction. Ph.D. thesis, University of Oklahoma, [Available online at https://shareok.org/handle/11244/44917.].

Gagne, D. J., S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **0 (0)**, doi:10.1175/MWR-D-18-0316.1.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1819–1840, doi:10.1175/WAF-D-17-0010.1.

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

Grant, L. D., and S. C. van den Heever, 2014: Microphysical and dynamical characteristics of low-precipitation and classic supercells. *Journal of the Atmospheric Sciences*, **71**, 2604–2624, doi:10.1175/JAS-D-13-0261.1.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, doi:10.1175/2007MWR2410.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014⟨0155:HTFENP⟩2.0.CO;2.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

Herman, G. R., and R. S. Schumacher, 2018a: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev*, **146**, 1571–1600, doi:10.1175/MWR-D-17-0250.1.

Herman, G. R., and R. S. Schumacher, 2018b: "Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation. *Mon. Wea. Rev*, **146**, 1785–1812, doi:10.1175/MWR-D-17-0307.1.

Heymsfield, A. J., 1982: A comparative study of the rates of development of potential graupel and hail embryos in high plains storms. *J. Atmos. Sci.*, **39**, 2867–2897, doi:10.1175/1520-0469(1982)039⟨2867:ACSOTR⟩2.0.CO;2.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, doi:10.1175/WAF-D-12-00113.1.

Hong, S.-Y., K.-S. S. Lim, Y.-H. Lee, J.-C. Ha, H.-W. Kim, S.-J. Ham, and J. Dudhia, 2010: Evaluation of the WRF Double-Moment 6-Class Microphysics Scheme for Precipitating Convection. *Adv. Meteorol.*, **2010**, 1–10, doi:10.1155/2010/707253.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes. *Mon. Wea. Rev*, **134**, 2318–2341, doi:10.1175/MWR3199.1.

Janjić, Z. I., 1990: The Step-Mountain Coordinate: Physical Package. *Mon. Wea. Rev*, **118**, 1429–1443, doi:10.1175/1520-0493(1990)118⟨1429:TSMCPP⟩2.0.CO;2.

Janjić, Z. I., 1994: The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes. *Mon. Wea. Rev*, **122**, 927–945, doi:10.1175/1520-0493(1994)122⟨0927:TSMECM⟩2.0.CO;2.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta Hail Growth Model Using Severe Hail Proximity Soundings from the United States. *Wea. Forecasting*, **24**, 1592–1609, doi:10.1175/2009WAF2222230.1.

Jirak, I. L., A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the Optimal Configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/29WAF25NWP/webprogram/Manuscript/Paper345640/waf_nwp_2018_jirak_href_config_ex_abs.pdf.].

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC Storm-Scale Ensemble of Opportunity: Overview and Results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. on Severe Local Storms*, [Available online at https://ams.confex.com/ams/26SLS/webprogram/Manuscript/Paper211729/2012_SLS_SSEO_exabs_Jirak_final.pdf.].

Johns, R. H., and C. A. Doswell, 1992: Severe Local Storms Forecasting. *Wea. Forecasting*, **7**, 588–612, doi:10.1175/1520-0434(1992)007⟨0588:SLSF⟩2.0.CO;2.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting Unique Information from High-Resolution Forecast Models: Monitoring Selected Fields and Phenomena Every Time Step. *Wea. Forecasting*, **25**, 1536–1542, doi:10.1175/2010WAF2222430.1.

Karstens, C. D., and Coauthors, 2015: Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, doi:10.1175/WAF-D-14-00163.1.

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

Labriola, J., N. Snook, Y. Jung, B. Putnam, and M. Xue, 2017: Ensemble Hail Prediction for the Storms of 10 May 2010 in South-Central Oklahoma Using Single- and Double-Moment Microphysical Schemes. *Mon. Wea. Rev*, **145**, 4911–4936, doi: 10.1175/MWR-D-17-0039.1.

Labriola, J., N. Snook, Y. Jung, and M. Xue, 2019: Explicit ensemble prediction of hail in 19 may 2013 oklahoma city thunderstorms and analysis of hail growth processes with several multimoment microphysics schemes. *Mon. Wea. Rev.*, **147 (4)**, 1193–1213, doi:10.1175/MWR-D-18-0266.1.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind. *Wea. Forecasting*, **32**, 2175–2193, doi:10.1175/WAF-D-17-0038.1.

Lai, Y., and D. A. Dzombak, 2019: Use of historical data to assess regional climate change. *J. Climate*, doi:10.1175/JCLI-D-18-0630.1.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An Efficient, General-Purpose Technique for Identifying Storm Cells in Geospatial Images. *J. Atmos. Oceanic Technol*, **26**, 523–537, doi:10.1175/2008JTECHA1153.1.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol*, **32**, 1209–1223, doi:10.1175/JTECH-D-13-00205.1.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble. *Wea. Forecasting*, **32**, 1403–1421, doi:10.1175/WAF-D-16-0200.1.

McCaul, E. W., and M. L. Weisman, 2001: The sensitivity of simulated supercell structure and intensity to variations in the shapes of environmental buoyancy and shear profiles. *Mon. Wea. Rev.*, **129**, 664–687, doi:10.1175/1520-0493(2001)129⟨0664:TSOSSS⟩2.0.CO;2.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bull. Amer. Meteor.*, **98**, 2073–2090, doi:10.1175/BAMS-D-16-0123.1.

Melick, C. J., I. L. Jirak, J. Correia, A. R. Dean, and S. J. Weiss, 2014: Exploration of the NSSL Maximum Expected Size of Hail (MESH) Product for Verifying Experimental Hail Forecasts in the 2014 Spring Forecasting Experiment. *27th Conf. on Severe Local Storms*, [Available online at https://ams.confex.com/ams/27SLS/webprogram/Manuscript/Paper254292/Melick_SLS2014_MESHVerif_SFE2014-Preprint_Final.pdf.].

Miller, L. J., J. D. Tuttle, and G. B. Foote, 1990: Precipitation production in a large montana hailstorm: Airflow and particle growth trajectories. *Journal of* *the Atmospheric Sciences*, **47**, 1619–1646, doi:10.1175/1520-0469(1990)047⟨1619:PPIALM⟩2.0.CO;2.

Molnar, C., 2019: *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Moore, J. T., and J. P. Pino, 1990: An Interactive Method for Estimating Maximum Hailstone Size from Forecast Soundings. *Wea. Forecasting*, **5**, 508–525, doi:10.1175/1520-0434(1990)005⟨0508:AIMFEM⟩2.0.CO;2.

Moore, T. W., 2018: Annual and seasonal tornado trends in the contiguous united states and its regions. *Int. J. Climatol.*, **38**, 1582–1594, doi:10.1002/joc.5285.

Murillo, E., and C. Homeyer, 2019: Severe hail fall and hail storm detection using remote sensing observations. *J. Appl. Meteor. Climatol.*, doi:0.1175/JAMC-D-18-0247.1.

Nelson, S. P., 1983: The influence of storm flow structure on hail growth. *J. Atmos. Sci.*, **40**, 1965–1983, doi:10.1175/1520-0469(1983)040⟨1965:TIOSFS⟩2.0.CO;2.

Nelson, S. P., and S. K. Young, 1979: Characteristics of oklahoma hailfalls and hailstorms. *J. Appl. Meteor.*, **18 (3)**, 339–347, doi:10.1175/1520-0450(1979)018⟨0339:COOHAH⟩2.0.CO;2.

Netzel, P., and T. Stepinski, 2016: On using a clustering approach for global climate classification. *J. Climate*, **29**, 3387–3401, doi:10.1175/JCLI-D-15-0640.1.

Ortega, K., 2018: Evaluating multi-radar, multi-sensor products for surface hailfall diagnosis. *Electronic J. Severe Storms Meteor*, **13**, 1–36, [Available online at http://ejssm.org/ojs/index.php/ejssm/article/view/163.].

Ortega, K. L., T. M. Smith, and G. J. Stumpf, 2006: Verification of multi-sensor, multi-radar hail diagnosis techniques. *Symp. on the Challenges of Severe Convective Storms*, Atlanta, GA, Amer. Meteor. Soc., [Available online at http://ams.confex.com/ams/pdfpapers/104885.pdf.].

Ortega, K. L., T. M. Smith, G. J. Stumpf, J. Hocker, and L. Lopez, 2005: A comparison of multi-sensor hail diagnosis techniques. *21st Conf. on Interactive Information Processing Systems*, San Diego, CA, Amer. Meteor. Soc., [Available online at http://ams.confex.com/ams/pdfpapers/87640.pdf.].

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Schaefer, J. T., J. J. Levit, S. J. Weiss, and D. W. McCarthy, 2004: The frequency of large hail over the contiguous United States. *14th Conf. on Applied Climatology*, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/pdfpapers/69834.pdf.].

Schwartz, C., and R. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, doi:10.1175/MWR-D-16-0400.1.

Snook, N., Y. Jung, J. Brotzge, B. Putnam, and M. Xue, 2016: Prediction and Ensemble Forecast Verification of Hail in the Supercell Storms of 20 May 2013. *Wea. Forecasting*, **31**, 811–825, doi:10.1175/WAF-D-15-0152.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Wea. Forecasting*, **26**, 714–728, doi:10.1175/WAF-D-10-05046.1.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307, doi:10.1186/1471-2105-9-307.

Svaldi, A., 2018: Damage from last year's massive front range hail storm cost $2.3 billion – $900 million more than first estimated. The Denver Post, [Available online at https://www.denverpost.com/2018/05/07/2017-front-range-hail-storm-damage-cost/.].

Thompson, R. L., B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous united states. part ii: Supercell and qlcs tornado environments. *Wea. Forecasting*, **27**, 1136–1154, doi:10.1175/WAF-D-11-00116.1.

Trapp, R. J., and H. E. Brooks, 2013: Regional characterization of tornado activity. *J. Appl. Meteor. Climatol.*, **52**, 654–659, doi:10.1175/JAMC-D-12-0173.1.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h Explicit Convective Forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, doi:10.1175/2007WAF2007005.1.

Weisman, M. L., and J. B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Mon. Wea. Rev.*, **110**, 504–520, doi:10.1175/1520-0493(1982)110⟨0504:TDONSC⟩2.0.CO;2.

Wilson, C. J., K. Ortega, and V. Lakshmanan, 2009: Evaluating Multi-Radar, Multi-Sensor Hail Diagnosis with High Resolution Hail Reports. *25th Conf. on Numerical Weather Prediction*, [Available online at https://ams.confex.com/ams/89annual/techprogram/paper_146206.htm.].

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An Enhanced Hail Detection Algorithm for the WSR-88d. *Wea. Forecasting*, **13**, 286–303, doi:10.1175/1520-0434(1998)013⟨0286:AEHDAF⟩2.0.CO;2.

Yang, S., and E. A. Smith, 2006: Mechanisms for diurnal variability of global tropical rainfall observed from trmm. *J. Climate*, **19**, 5190–5226, doi:10.1175/JCLI3883.1.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans. *Bull. Amer. Meteor.*, **92**, 1321–1338, doi:10.1175/2011BAMS-D-11-00047.1.

Zhang, Y., S. Moges, and P. Block, 2016: Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatial-temporal variability: Application to western ethiopia. *J. Climate*, **29**, 3697–3717, doi:10.1175/JCLI-D-15-0582.1.