UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

COUPLING DATA SCIENCE TECHNIQUES AND NUMERICAL WEATHER PREDICTION MODELS FOR HIGH-IMPACT WEATHER PREDICTION

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
DOCTOR OF PHILOSOPHY

By

DAVID JOHN GAGNE II
Norman, Oklahoma
2016

COUPLING DATA SCIENCE TECHNIQUES AND NUMERICAL WEATHER PREDICTION MODELS FOR HIGH-IMPACT WEATHER PREDICTION

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. Amy McGovern, Chair

Dr. Jeffrey Basara

Dr. Andrew Fagg

Dr. Michael Richman

Dr. John Williams

Dr. Ming Xue

Acknowledgements

The research in this dissertation would not be possible without the collaborations, resources, and discussions provided by a diverse and talented group of people from all over the country. First, I am grateful for my advisor, Dr. Amy McGovern. She first hired me as an REU student back in summer 2007 and has served as my primary advisor and mentor ever since then. She introduced me to the world of artificial intelligence and machine learning and encouraged me to undertake ambitious and challenging research projects. She has been a strong advocate for me and supported me through many stressful life milestones.

I would also like to thank all of my doctoral committee members: Dr. Jeffrey Basara, Dr. Andrew Fagg, Dr. John Williams, and Dr. Ming Xue. Their questions and comments during my general exam, seminar, dissertation defense, and other discussions really help strengthen the quality of the final dissertation. Special thanks also go to Dr. Sue Ellen Haupt for organizing and supervising my extended visit at NCAR. She and John also supported me through our weekly video conferences where they provided lots of feedback and ideas.

My hail forecasting research has been heavily supported by the Center for Analysis and Prediction of Storms through the Severe Hail Analysis, Representation, and Prediction (SHARP) project. I would like to thank fellow SHARP team members Dr. Nate Snook, Dr. Youngsun Jung, and Jon Labriola. Many other CAPS scientists and staff assisted with data access, computer support, and general advice and discussions: Tim Supinie, Dr. Kevin Thomas, Dr. Fanyou

Kong, Dr. Derek Stratman, Dr. Bryan Putnam, Dr. Keith Brewster, Chris Cook, Eileen Hasselwander, and Debra Farmer.

Through the National Center for Atmospheric Research (NCAR) Advanced Study Program (ASP) Graduate Visitor Program, I had the opportunity to spend a year working at NCAR. During that time I developed many collaborations that led to research included in this dissertation. For assistance with solar energy forecasting research I would like to thank Seth Linden, Dr. Jared Lee, Dr. Gerry Wiener, Dr. Pedro Jimenez, Dr. Julia Pearson, Dr. Tara Jensen, Dr. Tyler McCandless, Jim Cowie, Bill Petzke, Thomas Brummet, and Dr. Randy Bullock. For their work on the NCAR ensemble and help with incorporating my hail forecasting method in the operational system, I would like to thank Dr. Ryan Sobash, Dr. Glen Romine, and Craig Schwartz. Dr. Greg Thompson and Dr. Hugh Morrison helped me gain more insight into the workings of their microphysics schemes. I also greatly appreciate the many great conversations with my friends over lunch and tea that really enriched my time at NCAR.

The scientists at the NOAA National Severe Storms Laboratory (NSSL), the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS), and Storm Prediction Center have helped provide me with lots of data, research ideas, helpful discussion and feedback, and the opportunity to test my hail forecasting model in real-time through the Hazardous Weather Testbed. For assistance with participating in the HWT and the data and ideas that flowed from it, I wish to thank Dr. Jimmy Correia, Dr. Michael Coniglio, Dr. Adam Clark, Kent Knopfmeier, Chris Melick, Burkely Gallo, Dr. Chris Karstens, Dr. Israel Jirak, Dr. Patrick Marsh, and Dave Imy. Kiel Ortega, Travis Smith, and Carrie Langston helped me acquire the gridded radar products from NMQ, MRMS, and MYRRORS that were my key source of hail observations. For their extensive discussions, lessons, ideas, and feedback, I also thank Dr. Lou Wicker, Dr. Kim Elmore, Dr. Ted Mansell, Dr. Kristin Calhoun, Dr. Conrad Ziegler, and Dr. Harold Brooks. I also greatly appreciated the friendship and conversations with Jessica Erlingis Lamers, Greg Blumberg, Dr. Zac Flamig, Dr. Race Clark, Tiffany Meyer, Jill Hardy, Brandon Smith, and Alex Zwink.

I also greatly appreciate the support I received from the fellow students and alumni of the IDEA Lab: Ryan Lagerquist, David Harrison, Thibault Lucidarme, Andrew MacKenzie, Andrea Balfour, Taner Davis, and Carmen Chilson.

Very special thanks go to my parents, David and Diana Gagne, for all of the love and support they have given me throughout my life.

The research in this dissertation was funded under the NSF Graduate Research Fellowship Program Grant 2011099434, the NSF Severe Hail Analysis, Representation, and Prediction (SHARP) Grant AGS-1261776, and the National Center for Atmospheric Research Research Applications Lab and Advanced Study Program Graduate Visitor Program. The Oklahoma Mesonet is funded by the taxpayers of Oklahoma through the Oklahoma State Regents for Higher Education and the Oklahoma Department of Public Safety. Simulations and analysis used in this dissertation were generated on the NSF XSEDE TACC Stampede and the NCAR Yellowstone supercomputers.

Finally, I want to give a huge amount of thanks and gratitude to my lovely wife Cathy Bolene Gagne for being my partner on this long and wild journey. She has supported me through the many trials of graduate school, moving to Colorado and back, and many late nights writing and coding. She has been an amazing mother to our son, Robert Blaise, and has made it possible for me to finish this dissertation in time to begin the next stage of our lives.

Table of Contents

List Of Tables

List Of Figures

Abstract

Meteorologists have access to more model guidance and observations than ever before, but this additional information does not necessarily lead to better forecasts. New tools are needed to reduce the cognitive load on forecasters and to provide them with accurate, reliable consensus guidance. Techniques from the data science community, such as machine learning and image processing, have the potential to summarize and calibrate numerical weather prediction model output and to generate deterministic and probabilistic forecasts of high-impact weather. In this dissertation, I developed data-science-based approaches to improve the predictions of two high-impact weather domains: hail and solar irradiance. Both hail and solar irradiance produce large economic impacts, have non-Gaussian distributions of occurrence, are poorly observed, and are partially driven by processes too small to be resolved by numerical weather prediction models.

Hail forecasts were produced with convection-allowing model output from the Center for Analysis and Prediction of Storms and National Center for Atmospheric Research ensembles. The machine learning hail forecasts were compared against storm surrogate variables and physics-based diagnostic models of hail size. Initial machine learning hail forecasts reduced size errors but struggled with predicting extreme events. By coupling the machine learning model to predicting hail size distributions and estimating the distribution parameters jointly, the machine learning methods were able to show skill and reliability in predicting both severe and significant hail.

Machine learning model and data configurations for gridded solar irradiance forecasting were evaluated on two numerical modeling systems. The evaluation determined how machine learning model choice, closeness of fit to training data, training data aggregation, and interpolation method affected forecasts of clearness index at Oklahoma Mesonet sites not included in the training data. The choice of machine learning model, interpolation scheme, and loss function had the biggest impacts on performance. Errors tended to be lower at testing sites with sunnier weather and those that were closer to training sites. All of the machine learning methods produced reliable predictions but underestimated the frequency of cloudiness compared to observations.

Chapter 1

Introduction

The weather community sails on a turbulent sea of noisy data. The accuracy of weather forecasts is driven by two main factors: (1) the quality of the observational data and model guidance available to forecasters, and (2) the ability of the forecasters to properly interpret that information (Stewart 2001). The effect of improving guidance quality is evident in tornado warning verification trends (Fig. 1.1) with the largest, recent increases in probability of detection occurring with the deployment of the NEXRAD network in the early 1990s (Brooks 2004). The finer resolution of numerical weather prediction modeling systems and the increasing availability of ensemble guidance require forecasters to analyze and consider the likelihood of a wider range of possible weather scenarios. Radar and satellite observations are also increasing in temporal and spatial resolution, revealing new weather features but also requiring more time to analyze. Smart phones, crowdsourcing, and personal weather stations are conveying all kinds of potentially relevant data that has shown the ability to improve forecasts (Mass and Madaus 2014), but they also bring a lot of noise due to poor calibration, siting issues, and human biases. Even with all this change, weather forecaster displays and methods have not been evolving fast enough to keep pace with the explosion of data. Poor integration of new data sources with forecasters could result in static or decreasing performance along with greater incentive for more forecaster duties to be replaced with automation (Snellman 1977).

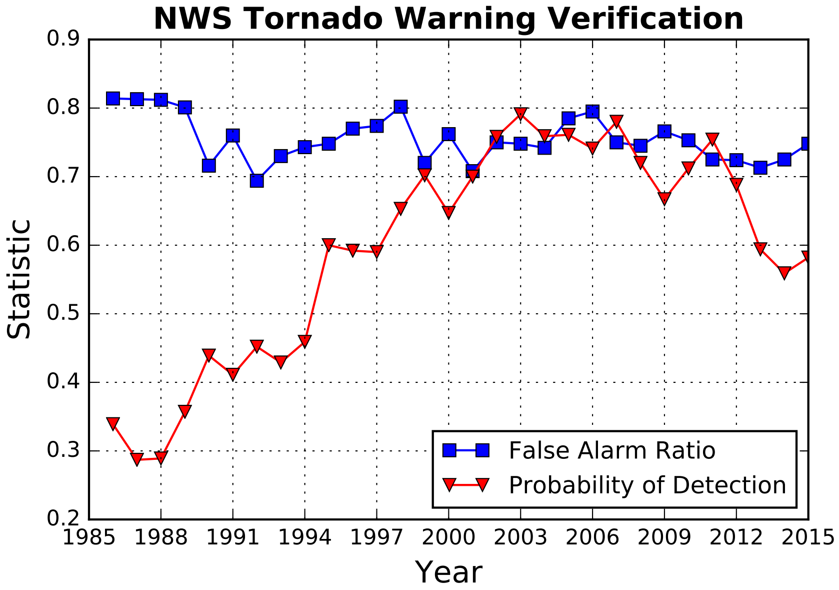


Figure 1.1: Yearly National Weather Service tornado warning verification statistics for the US. Source: NOAA.

The heuristics and biases present in human cognition have a limiting effect on the ability of using more information to improve forecasts (Stewart et al. 1992; Doswell III 2004). While providing forecasters with more information does have a slightly positive impact on accuracy, it can lead to negative impacts on reliability as forecasters become overconfident in their decision (Stewart et al. 1992, 1997). Through education, training, and experience, forecasters develop a set of rules and heuristics for evaluating weather guidance and generating forecasts (Doswell III 2004). Forecasters may not benefit from additional guidance if it does not match with their conceptual model of the situation (representativeness bias), if it is a tool or situation outside of the previous experience of the forecaster (availability bias), or if the subsequent information differs significantly from initial guidance (anchoring) (Tversky and Kahneman 1974; Doswell III 2004). The anchoring effect can lead to confirmation bias when forecasters accept additional information that supports their initial instinct, while rejecting anything contrary. In order to minimize the effects of these biases, the guidance used should be as relevant, reliable, and unbiased as possible (Stewart 2001).

Statistically-corrected consensus ensemble prediction systems would then be the logical choice for initial guidance, but forecasters have expressed distrust of ensembles and statistical correction methods (Novak et al. 2008). Instead, they prefer to use deterministic numerical weather prediction (NWP) models or a subset of individual ensemble members to predict high impact events. Consensus ensemble guidance in the form of ensemble mean or probability tends to smooth gradients and decrease the amplitude of extremes. Deterministic guidance is more physically consistent and can show extreme events, but it may overstate the likelihood of an extreme event occurring, especially if forecasters are biased toward looking at the extreme solutions. Existing linear bias-correction techniques like Model Output Statistics (Glahn and Lowry 1972) can correct for some issues with ensembles but have a limited ability to improve the representation of extreme events. Forecasters have expressed that they are more likely to adopt a new technique if it demonstrates a significant improvement over existing approaches and gives more direct guidance for the phenomenon being forecast (Morss and Ralph 2007).

In this dissertation, I develop data-science-based approaches to improve the predictions of two high-impact weather domains: hail and solar irradiance. While the two areas may seem to share little in common at first glance, both phenomena produce large economic impacts on a frequent basis. Hail causes billions of dollars in property and crop damage in the United States each year (Changnon 2009) and is a major yearly liability for insurance companies (Brown et al. 2015) with $850 million in average annual claims. Urban sprawl and population growth in large cities such as Dallas/Fort Worth, St. Louis, Chicago, and Denver have made large amounts of property damage from hail events more likely (Rosencrants and Ashley 2015). Solar energy is a rapidly growing source of electricity whose variability needs to be predicted accurately, so electricity loads can be properly balanced and operational costs can be minimized. While solar irradiance itself does not cause disaster, poor forecasts and underestimation of the uncertainty could lead to major monetary losses for electric utilities and energy trading firms, and brownouts and blackouts due to an inadequate electricity supply could occur.

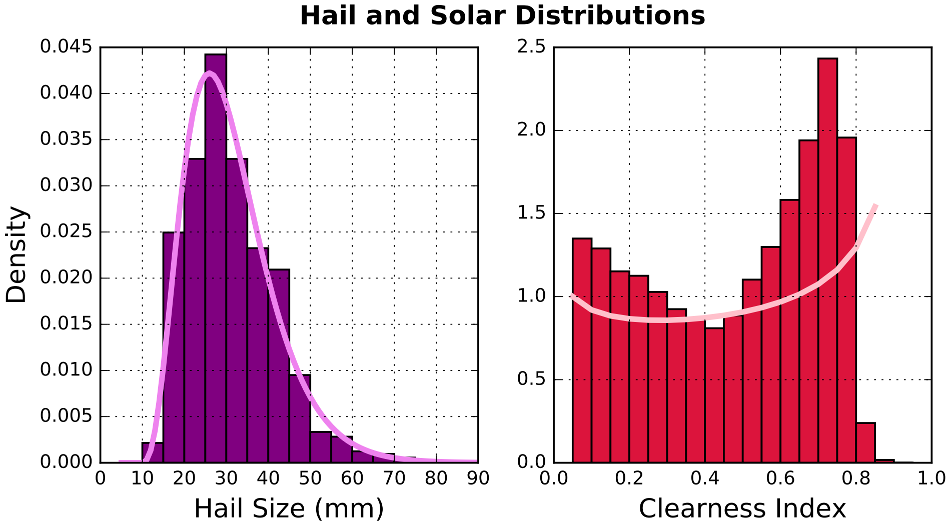


Figure 1.2: Observed distributions of hail sizes and clearness index during May and June 2015. A gamma distribution is fitted to the hail size values, and a beta distribution is fitted to the clearness index values.

Both hail and solar irradiance exhibit non-Gaussian distributions in their range of values (Fig. 1.2). Hail size follows a gamma distribution, and clearness index, a scaled measure of observed versus idealized irradiance, exhibits a beta distribution (Falls 1974) with a larger peak near the maximum (Jurado et al. 1995). Large hail occurs rarely at a given location but happens almost daily somewhere in the contiguous US during the months from April through

July. While partly cloudy skies are rarer than mostly clear or cloudy days, they still occur fairly frequently at most locations. Both are rare events but common enough that collecting forecasts and observations over the period of a few months and a wide area can capture much of the variability associated with each phenomenon. This level of relative rareness makes these phenomena more amenable to prediction using statistical and machine learning methods.

Techniques from the data science community have the potential to address the needs of high-impact weather event forecasters by integrating the large amounts of information available into reliable hazard forecasts. Advances in computing and storage have allowed us to amass vast archives of data and process it in real-time. Image processing techniques can filter gridded data to identify salient features, which can make the analysis of data over long periods of time easier and more consistent (Lakshmanan and Smith 2009). Machine learning methods can generate highly accurate predictive models from complex, multidimensional datasets by discovering underlying structures in the data with less reliance on theoretical assumptions about its origins (Breiman 2001b). The best machine learning methods balance predictive accuracy with robustness to noisy data and provide some level of interpretability. Ensemble decision tree methods, including random forests (Breiman 2001a) and gradient boosted regression (Friedman 2001), use interpretable base models while providing consistently high performance across many datasets (Caruana and Niculescu-Mizil 2006; McGovern et al. 2015). To produce these significant performance gains, however, machine learning models require one crucial item: a large set of forecast data paired with observations. High-impact weather events tend to be both extreme and rare, so a large amount of data is generally required to capture them well. Physical understanding is also helpful in constraining the data sources needed for prediction and ensuring that the predictions are physically consistent.

How much of a role should automated guidance play in the forecast process? For forecasts of standard weather variables, such as temperature and precipitation, the National Weather Service (NWS) currently operates with a human-in-the-loop paradigm in which forecasters subjectively blend and adjust multiple sources of guidance to create a final forecast. While the NWS approach is very time and labor intensive, local offices may be able to add predictive value in situations where local effects have a larger impact on the forecast. At the NWS Weather Prediction Center, which issues temperature and precipitation forecasts over the entire US, the human forecasts now perform significantly worse than down-scaled, bias-corrected ensemble forecasts for temperature and precipitation (Novak et al. 2014). Official NWS track forecasts of hurricanes, a major form of high impact weather, also perform worse than weighted ensemble consensus forecasts (Cangialosi and Franklin 2015). There are also issues with spatial discontinuities in forecasts and warnings between the domains of different forecast offices (Gilbert et al. 2015). Many private weather firms, including the Weather Company, operate in a human-over-the-loop paradigm in which an optimal blend of bias-corrected model output is generated as needed by users, and human forecasters can adjust blending weights to account for observed short-term biases or data quality issues (Williams et al. 2016). This approach scales easily and only requires a small team of meteorologists to oversee a mostly automated system. The downside of a heavily automated approach is that forecasters may become disengaged from the forecast process (Pliske et al. 2004) and struggle to take appropriate corrective action when automation fails (Skitka et al. 1999; Pagano et al. 2016). By studying the error characteristics of different machine learning methods in high-impact weather situations, researchers and forecasters can identify when the automated guidance should be trusted and when it is more likely to struggle.

The primary hypothesis of this dissertation is that properly configured decision tree ensemble machine learning models will produce day-ahead predictions of hail size and solar irradiance that show significantly more skill than raw NWP model output, physics-based diagnostic models, and linear regression. The secondary hypothesis is that properly configured decision tree ensemble machine learning models will produce distributions of forecasts that are physically consistent with distributions of observations. In Chapter 3, I evaluate machine learning regression models that directly predict maximum hail size from convection-allowing model output against a physics-based diagnostic method and determine whether any of the approaches can produce both low size errors and identify extreme hail events accurately. Utilizing the experiences from Chapter 3, in Chapter 4 I develop machine learning models with constraints in their pre-processing and training procedures aimed at producing more accurate and physically-realistic hail size forecasts. The machine learning hail forecasts are evaluated against other physics-based methods to determine how well each method detects hail events and how likely false alarms are for a given probability of detection. In Chapter 5, I create solar irradiance forecasting systems with varied machine learning model configurations. I evaluate the prediction errors to determine which models and configurations have a significant impact on performance. I also stratify errors by forecast value and site to identify the major physical sources of error in the predictions. In Chapter 6, I discuss the insights gained from this investigation and future directions for this research.

Chapter 2

Background

Meteorologists create high-quality weather forecasts by synthesizing information from an array of observations and guidance provided by numerical weather prediction (NWP) models. In order for NWP model guidance to be interpretable and useful, it must first be post-processed. Post-processing in general transforms raw model output into a form that is more easily interpreted by the forecaster. Basic post-processing involves interpolating NWP model output from the original grid to other coordinate systems, such as constant pressure levels or height above ground, and calculating derived quantities and diagnostic variables from the fundamental prognostic variables. Feature identification catalogs different areas of interest for particular forecasting tasks. Statistical post-processing combines NWP model output with outside observations and other data sources to produce a calibrated forecast product. This chapter describes these different post-processing methods in more detail and discusses the merits and limitations of different approaches. The ingredients and forecasting approaches for hail and solar irradiance are then examined.

2.1 Convection-Allowing Model Ensembles

As computational speed and storage capacities have increased in the past 20 years, research groups and operational weather forecast centers have run real-time NWP models capable of explicitly representing deep, moist convection (Kain et al. 2008). NWP models with grid spacing larger than 4 km must parameterize deep convection in order to capture the thermodynamic and precipitation effects from convective overturning properly. At coarser grid spacings, convection occurs on a slower time scale than observed, resulting in convective structures, heat and moisture fluxes, and precipitation amounts being represented incorrectly (Weisman et al. 1997). While not all convective processes are adequately resolved between 1 and 4 km, individual updrafts and their associated heat and moisture fluxes can be represented with a reasonable degree of accuracy. Because some convective processes are not fully resolved, these models are called Convection Allowing Models (CAMs). Even with their imperfect representation of convection, CAMs have shown skill over mesoscale NWP models in precipitation forecasting and in forecasting convective evolution and morphology (Weisman et al. 2008). However, errors in the initial conditions and model physics lead to spatial errors in storm placement and timing errors in convective initiation and storm evolution. In order to account for the uncertainty associated with these errors, modelers have developed CAM ensembles that perturb the initial and boundary conditions, physics parameterizations, and the dynamical cores to create a set of realizations that capture the range of possible convective solutions for a given day. CAM ensembles of many different configurations have run in real time as part of the NOAA Hazardous Weather Testbed Spring Experiment since 2007 (Clark et al. 2012a). The National Weather Service is already running deterministic CAMs operationally and is planning to deploy CAM ensembles in the near future (Benjamin 2014).

While CAMs cannot directly resolve the severe hazards associated with thunderstorms, which include tornadoes, hail, high winds, and flash floods, diagnostic information extracted from individual storms has shown skill in inferring the potential for these hazards. Because storms can greatly evolve in structure and intensity on the order of minutes, hourly instantaneous snapshots of CAM output will likely miss the time when the storm is at its greatest intensity. More frequent model output may capture these extremes, but requires more storage space. A useful compromise is to track the maximum value of a quantity at a grid point over a given time period and output that field every hour. These hourly maximum fields can provide a proxy for both storm track and intensity (Kain et al. 2010). For severe weather forecasting, updraft helicity, the integrated product of vertical velocity and vertical vorticity, between 2 and 5 km is a good proxy for strong, rotating updrafts (Kain et al. 2008) and shows some correlation with tornado path length (Clark et al. 2013). Other hourly maximum fields include updraft and downdraft speeds, radar reflectivity at the -10C level and column-integrated graupel.

Products derived from the hourly maximum fields can help diagnose severe weather likelihood and intensity. The length and direction of the tracks provides an estimate of storm speed and motion. Reliable probabilistic guidance for severe weather can be derived from hourly maximum fields by selecting a threshold, marking grid points where the threshold was exceeded within a given radius, and then applying a Gaussian smoother to the surrogate severe report grid (Sobash et al. 2011). These probabilities can be generated for both deterministic and ensemble configurations (Fig. 2.1) and are very computationally efficient to generate. Varying the standard deviation of the Gaussian smoother

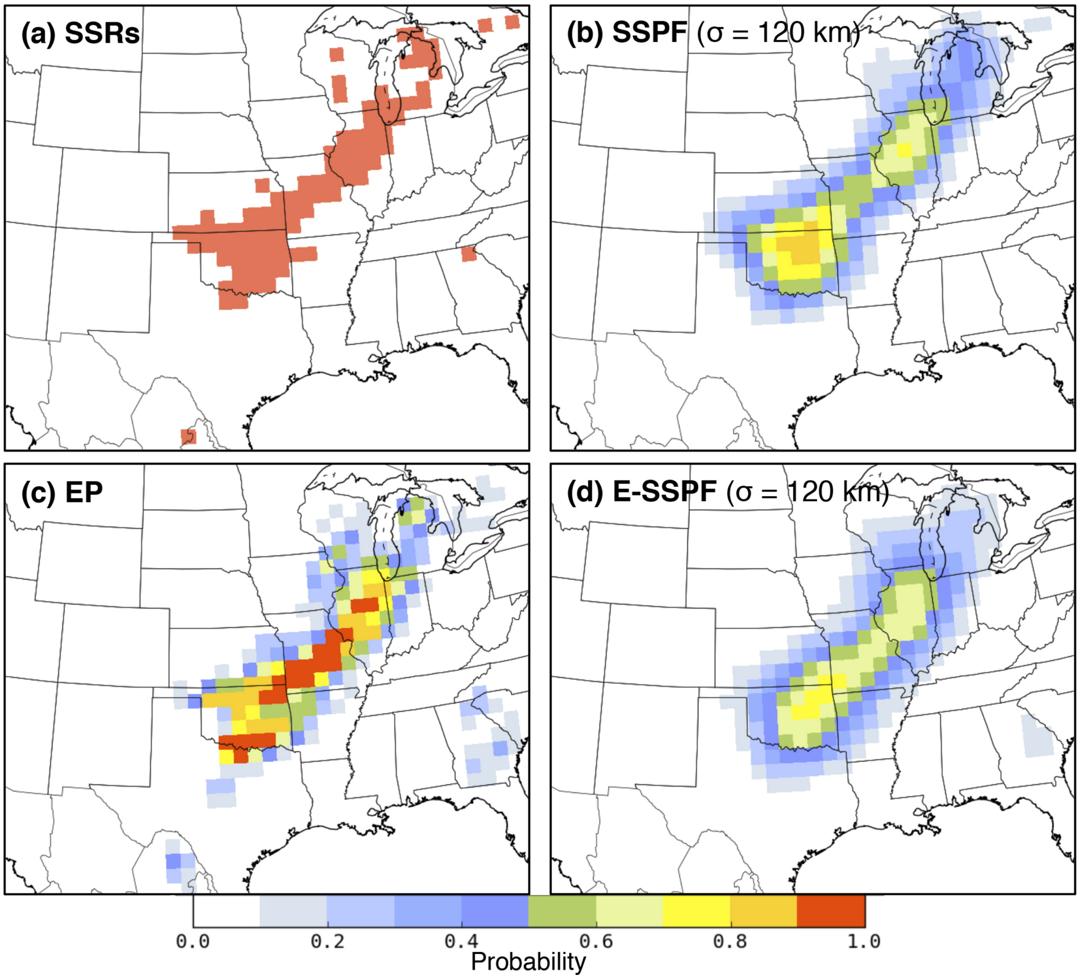


Figure 2.1: (a) Storm Surrogate Reports on an 80 km grid. (b) Storm Surrogate Probability Forecast from a single ensemble member. (c) Ensemble probability (mean of (a)) on coarse grid. (d) Ensemble Storm Surrogate Probability Forecast. Figure from Sobash et al. (2016).

can allow the user to calibrate the storm surrogate probabilities based on the severe weather report of choice (Sobash et al. 2016).

Storm surrogate products have some inherent limitations. The current storm surrogate fields do not directly predict the occurrence of severe weather hazards. The distributions of intensities of the storm surrogate products are model core, grid spacing, and parameterization scheme dependent(Kain et al. 2008), which can make them more challenging for forecasters to process when examining model output from multiple ensemble prediction systems. Most of the storm surrogate products are only indirectly verifiable in that the surrogate variables are not observed but can be correlated with severe reports. A forecast product that directly forecasts the chance or occurrence of a particular severe threat would be more useful if the predictions were well-calibrated and if they were physically consistent with the model forecast.

2.2 Feature Identification and Tracking Methods

Feature-based post-processing techniques are often used for forecasting atmospheric phenomena that occupy a discrete area and move and evolve with time. Commonly tracked features include cyclones (Blend and Schubert 2000), precipitation areas (Davis et al. 2009), thunderstorms (Dixon and Wiener 1993), fronts (Hewson 1998), and jet streams (Limbach et al. 2012). With feature identification and tracking, scientists can catalog the locations, intensities, and durations of features and compare them across space, time, and different datasets. Feature-based datasets generally require much less data to be stored per event, which allows for larger archives given fixed storage amounts. High-impact weather events tend to be associated with discrete features, so using feature-based analysis for forecasting can reduce the computational and cognitive load for both forecasters and their guidance models. In this dissertation, I use feature identification and tracking to identify potential hailstorms. I can then extract information about each feature and feed it into machine learning models to produce predictions.

This analysis can be done either subjectively by individuals who hand-label each feature or objectively by automated algorithms that apply the same criteria to every event. Subjective feature identification takes advantage of people's natural pattern recognition skills and is better for capturing complex features and edge cases. However, the process is very labor intensive and time-consuming (Lakshmanan and Smith 2009), and it often requires someone with training and expertise on the given phenomenon. Subjective identification can also be inconsistent with different experts, or even the same expert at different times, analyzing the same feature or similar features differently depending on their experience or fatigue levels (Hewson 1998). Objective, or automated, feature identification tends to be faster and more consistent than humans and can be run in real-time or archival situations. It can be scaled very cheaply across multiple processors or machines, and it can be performed with different settings on the same dataset to ensure greater robustness. On the other hand, automated approaches generally require a lot of up-front labor to develop, and often require data to first be quality-controlled and smoothed in order to produce good results. While many automated techniques are available for feature identification, all of them have to be fine-tuned for the needs and challenges of a particular domain before being used operationally.

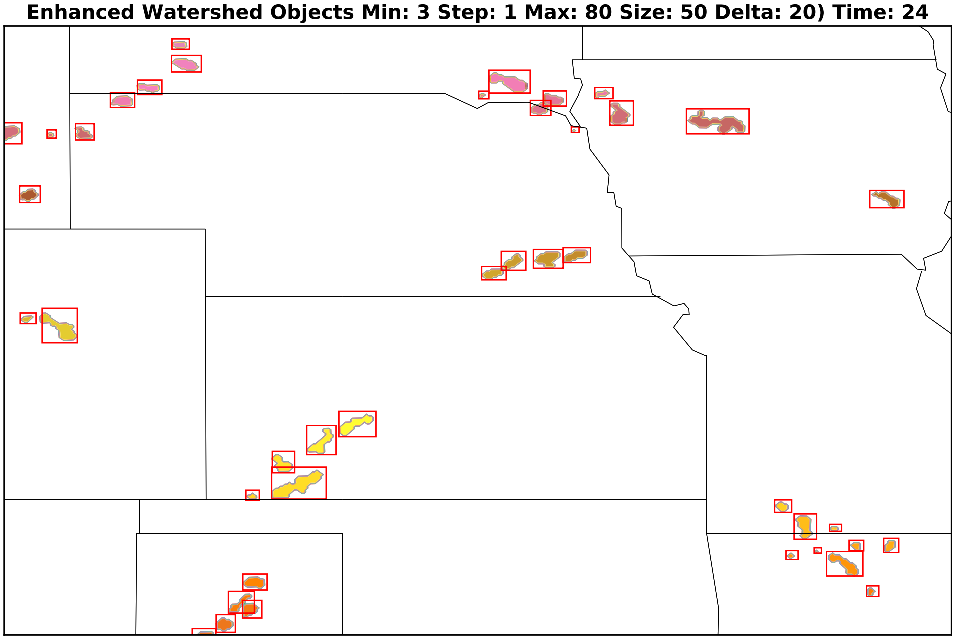

Enhanced Watershed Objects Min: 3 Step: 1 Max: 80 Size: 50 Delta: 20) Time: 24

Figure 2.2: An example of features identified in a total column graupel field using the enhanced watershed technique on convection-allowing model output. The features are colored by ID number. The red rectangles show the bounding box around each feature.

Feature identification on a spatial grid typically involves a process of identifying candidate center points, growing regions of influence, merging features, and filtering those that do not meet certain criteria. Simple feature identification methods, such as those used in TITAN (Dixon and Wiener 1993), SCIT (Johnson et al. 1998), and MODE (Davis et al. 2009), look for contiguous areas that exceed a single intensity threshold, but they may capture too many spurious objects if the threshold is set too low or merge objects together that should be considered separate. Objects with maximum values near the threshold may also disappear and reappear due to small fluctuations and would be considered separately by the algorithm. The hysteresis method, which requires each feature to contain at least one point exceeding a higher threshold, in addition to all points exceeding a lower threshold, helps filter some spurious objects. The enhanced watershed method (Lakshmanan et al. 2009) grows objects until they reach a specified saliency, or area criterion. This change in criteria makes the method more scale-aware and reduces its sensitivity to the choice of intensity threshold. An example of the features identified by the enhanced watershed method is shown in Fig. 2.2.

Feature tracking methods use some combination of centroid matching and feature cross-correlation. In centroid matching methods, distances are computed from all centroids at one time step to all centroids at another time step. Then features meeting the minimum distance criteria are matched, and those without a matching pair are considered either terminated or new features. The TITAN storm-tracking algorithm (Dixon and Wiener 1993) uses a globally optimal matching algorithm (Munkres 1957) to find the best pairings of storms and to resolve track continuations in the cases of mergers and splits. Han et al. (2009) created an enhanced version of TITAN that first matched objects that overlapped spatially before matching with a global cost function. Lakshmanan and Smith (2010) evaluated 5 commonly used storm tracking algorithms and devised a new hybrid tracking algorithm that combined the best features from all of the other ones. One notable improvement in Lakshmanan and Smith (2010) was using a cross-correlation filter to estimate storm motion by finding the amount of translation that led to the highest spatial correlation with the grid at the previous time step. Limbach et al. (2012) used the overlap of jet stream features in time and space to perform 4-dimensional tracking. The MODE-Time Domain algorithm (Clark et al. 2014) also uses feature overlap to track objects in time.

2.3 Statistical Post-Processing

Statistical models for post-processing NWP model output have evolved within two general frameworks. Perfect-prog, the first framework, fits a statistical model between observed or analyzed variables and observations of a weather feature, such as temperature or precipitation (Klein et al. 1959). The statistical model is then applied to NWP forecasts that assume the model is perfect. Model Output Statistics (MOS), the second framework, fits a statistical model between NWP output at a given time horizon and observations at that time (Glahn and Lowry 1972). Because MOS fits to the NWP output directly, it can correct for systematic biases assuming that the configuration of the NWP model does not change. When NWP model configurations are updated, the MOS equations have to be regenerated after a sufficient number of new model forecasts are made. Perfect-prog models are generally less accurate than a well-tuned MOS model, but they are less sensitive to model configuration changes and tend to be more robust over time.

Both perfect-prog and MOS models have traditionally used multiple linear regression in which the input variables are selected through an iterative screening process. Hundreds of potential input variables are usually available, but the strength of the correlation with the predicted value may be low, or they may be highly correlated with other input variables (Glahn and Lowry 1972). Forward screening selection addresses this problem by fitting linear regression models to all input variables and selecting the model that produces the largest reduction in variance, then finding a second variable that provides the largest reduction in variance when combined with the first, and so on until some stopping criteria is met in terms of number of variables or minimal reduction in variance (Glahn and Lowry 1972). The method produces a set of strong predictors that also minimize cross-correlation, but also requires fitting thousands to millions of linear models before settling on a final one. MOS can produce deterministic, categorical, or probabilistic predictions that are constrained by transforming the input variables into binary values.

The training and forecast data for MOS models need to exhibit as much stationarity as possible in order to maximize predictive skill. MOS equations are generally trained for single sites at each lead time in order to capture local effects closely. For variables that have more skewed distributions, regional aggregation of similar sites helps capture rare events (Lowry and Glahn 1976). While regional models are expected to perform well within their specified region, application of these models on a high-resolution grid can lead to discontinuities at the borders of regions. When developing gridded MOS, Glahn et al. (2009) minimized spatial discontinuities by using a successive correction method (Cressman 1959), spatial smoothing, land-sea masking, and elevation information. The gridded MOS approach does produce effective forecasts, but it requires a large number of individual regression models for each location and lead time to produce effective forecasts.

2.4 Machine Learning

Modern datasets are constantly increasing in size and complexity, and greater demands are being placed on people to make predictions from these datasets. The assumptions of Gaussian-distributed data and constant variance that underly the simple multiple linear regression used by MOS mean that the approach does not scale well as more examples and predictors are added. Machine learning models, however, are designed to extract information from large, multidimensional, noisy datasets and produce predictions that generalize well. Unlike traditional statistical modeling approaches, which look for empirical distributions that best fit a smaller sample of data as closely as possible, machine learning modeling approaches derive as much information as possible from the data directly and focus on maximizing predictive performance on data independent of what was used to fit the model (Breiman 2001b). Machine learning models make gains in predictive performance by fitting to data in a way that is constrained by model structure and parameters to maximize signal and to reduce noise.

2.4.1 Regularized Linear Models

Regularization techniques allow linear models to fit noisy, high-dimensional data while minimizing the potential for overfitting. Simple linear regression determines its weights by minimizing the mean squared error between the model and observations, which can be done numerically using a gradient descent method. The loss function is defined in Eq. 2.1

$$\min_{w} \frac{1}{2n} \left\| Xw - y \right\|_2^2, \tag{2.1}$$

where w is the weight vector, X is the matrix of predictor values, n is the number of rows in the predictor matrix, and y is the predictand vector. If the input data to the model are noisy such that for a given set of input values there are multiple output values or contains many input variables with varying degrees of relevance, the least squares fit would either overfit the noise or not converge to an unbiased solution at all. A solution to this problem is to add a bias term to the minimization function that constrains the possible coefficient values.

Ridge regression (Hoerl and Kennard 1970) addresses this issue by adding a penalty term to the minimization function that is the L2 norm of the coefficient vector as shown in Eq. 2.2:

$$\min_{w} \frac{1}{2n} \left\| Xw - y \right\|_2^2 + \frac{\alpha}{2} \left\| w \right\|_2^2, \tag{2.2}$$

where α scales the effect of the Ridge term. The ridge term penalizes large-magnitude coefficients and shifts the optimal solution toward a set of small-magnitude coefficients. Lasso regression (Tibshirani 1996) uses the L1 norm of the weights in place of the L2 norm as shown in Eq. 2.3:

$$\min_{w} \frac{1}{2n} \left\| Xw - y \right\|_2^2 + \alpha \left\| w \right\|_1. \tag{2.3}$$

This minor change results in the potential for a sparse set of coefficients such that some of them are set to 0. This feature of Lasso allows it to do implicit variable selection as part of the optimization process. Unlike Ridge, Lasso has no analytical solution. The Elastic Net (Zou and Hastie 2005) combines the Ridge and Lasso terms in the loss function and balances their effects using a weighting term ρ (Eq. 2.4):

$$\min_{w} \frac{1}{2n} \left\| Xw - y \right\|_2^2 + \alpha\rho \left\| w \right\|_1 + \frac{\alpha(1-\rho)}{2} \left\| w \right\|_2^2 . \tag{2.4}$$

The magnitude of their effects is governed by α. By using cross-validation to determine the best ρ and α, a more robust set of coefficients can be found.

2.4.2 Decision Trees

Decision trees are a class of machine learning model that use a hierarchical set of decision thresholds to recursively partition a complex, multidimensional feature space into many, more uniform subsets (Breiman et al. 1984). A decision tree consists of binary decision nodes that ask yes-or-no questions about the input features, such as "Does the temperature exceed 20 °C?" The decision nodes eventually branch out to leaf nodes that output a prediction based on the training data that reached them. The prediction can be a class label, probability distribution, or continuous value depending on whether the decision tree is being used for classification or regression. An example of a decision tree can be found in Fig. 2.3.

Decision trees are grown by greedily determining the best input feature and threshold for splitting a given set of data into many more uniform subsets. In the traditional decision tree framework, all input features are evaluated at each

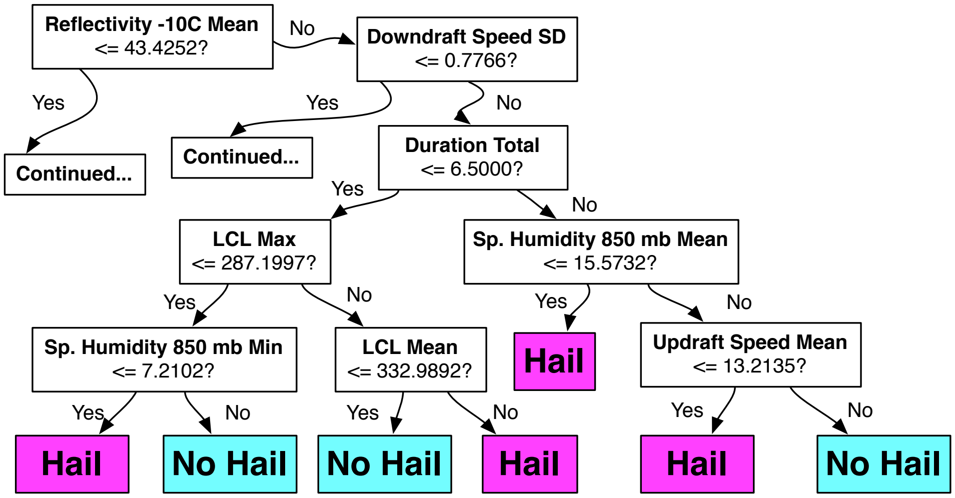


Figure 2.3: An example of a decision tree predicting whether or not hail will occur trained on data from the 2014 Center for Analysis and Prediction of Storms ensemble. See Chapter 4 for more details.

node, and candidate splitting thresholds are picked from transitions in labels for the training data. An impurity metric, such as cross-entropy or the Gini index for classification and mean squared error for regression (Hastie et al. 2009), is calculated for the current node and for both possible child nodes for every potential feature-threshold combination (Table 2.1). The combination with the largest reduction in impurity is selected for inclusion in the tree, and the training data are then split and sent downward where the growth process continues. Decision tree growth stops when a certain level of uniformity, a maximum depth, a minimum number of training cases, or a lack of significant decreases in error is reached. Through the tree-growing process, irrelevant features are not included, and a human-interpretable model is constructed. Because the predictions are only based on the training data labels, decision trees are resistant to noisy and missing input data, but they also do not extrapolate beyond values found in the training set.

Table 2.1: The definitions of commonly used impurity metrics for classification and regression decision trees (Hastie et al. 2009). $\hat{p}_{mk}$ is the probability of class k of classes K in decision node m. N_m is the number of training cases reaching node m. $x_i \in R_m$ describes the training cases that have a value x_i within region R_m. The training labels y_i are compared with the mean of the training cases within a node $\hat{c}_m$.

Impurity Metric	Equation
Gini Index	$\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
Cross-Entropy	$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$
Mean Squared Error	$\frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$

Decision trees also have many inherent limitations when used in isolation. When compared with other machine learning models, decision trees often produce less accurate predictions. The models are sensitive to slight changes in the training data, which can have a cascading effect on what input features are included in the tree. Deeper splits in the tree may not be informative if the number of training examples is small, resulting in overfitting (Hastie et al. 2009).

2.4.3 Decision Tree Ensembles

Some of the limitations of decision trees can be ameliorated through different forms of ensembles along with the addition of stochastic processes. Two of the most powerful decision tree ensemble techniques are random forests and gradient boosting trees. Random forests (Breiman 2001a) are unweighted ensembles of randomized decision trees. Each decision tree in the random forest is trained using a bootstrap-resampled dataset, and a random subset of the input variables are evaluated for inclusion in the tree at each node during the tree growth process. These randomization steps increase the independence of the trees in the ensemble and lead to greater exploration of the feature space (Ho 1998). Evaluating random subsets of the features instead of all possible ones also reduces the computational cost of building a large ensemble. Typically, the square root of the number of input variables is used for random subset sampling although a larger number may be necessary for datasets with many weak inputs (Hastie et al. 2009). By averaging the predictions from a set of low-training-set-error trees, the variance of the predictions is reduced, so the resulting predictions generally are smooth and have low error. Growing each tree independently also means that the tree growth can be easily parallelized for large datasets. Compared to other nonlinear machine learning methods, such as neural networks and support vector machines, random forests have fewer tunable parameters. In addition, a wide range of parameter values will generally produce good performance. Increasing the number of trees in the forest improves performance by reducing the variance of the ensemble average, but performance increases tend to diminish beyond about 100 trees (Breiman 2001a; Hastie et al. 2009). Adding more trees leads to the ensemble mean prediction converging toward a single value due to the central limit theorem and weak law of large numbers, so continuing to add trees will not result in overfitting (Breiman 2001a). The depth of the trees effects the variance of the predictions. A smaller tree depth results in more training samples at a given leaf node, resulting in the individual trees and thus the ensemble average having less variance and sharpness. Random forests are typically grown to the point where each leaf node has a very small number of examples to increase the sharpness, and the ensemble averaging process will then reduce the systematic bias and variance of the predictions (Breiman 2001a; Hastie et al. 2009). Due to their strong performance and relative ease of use, random forests have been increasingly adopted by the members of the weather community for a wide range of nowcasting and forecasting problems, including storm classification (Lakshmanan et al. 2010; Gagne II et al. 2009), convection initiation (Ahijevych et al. 2016), aircraft turbulence (Williams 2014), and hurricane power outages (Nateghi et al. 2014).

Gradient boosting trees (Friedman 2001) are additive, stagewise ensembles of decision trees. In a stagewise ensemble, a series of relatively weak models is fit sequentially to minimize the errors on the training set predictions made by the previous model. In the case of gradient boosting trees, the weak model is a decision tree. The initial tree is fit directly to the training set labels, while each additional model is fit to the negative gradient of the loss function. For the mean squared error loss function, the negative gradient is the residual of the training label and the prediction from the previous tree. Training examples that have larger residuals will receive higher weights in the fitting process. The predicted residual from each tree is then added to the sum of the previous tree predictions. Because adding more trees to boosting eventually leads to overfitting, each tree's residual prediction is multiplied by a learning rate parameter that reduces the magnitude of the residual by a constant value (Hastie et al. 2009). A smaller learning rate parameter requires more trees to achieve the same level of training error but tends to provide lower testing set error. Unlike a random forest, which tends to perform better with large trees, the gradient boosting procedure favors small trees for both faster computation and lower test set errors. The algorithm is not parallelizable but can be computationally efficient if the tree size is limited. Gradient boosting trees were not widely used in the meteorology community until the AMS Solar Energy Prediction Contest, in which the top 4 teams all used some variation of the model for their entry (McGovern et al. 2015).

2.5 Forecast Ingredients

2.5.1 Hail Forecasting

While hail growth and melting is governed by the complex microphysical processes that a hailstone experiences along its trajectory through a storm, hail forecasting methods have relied on a combination of coarse approximations of the expected storm environment and local hail climatology with varying degrees of success (Johns and Doswell III 1992). Most existing hail forecast methods (e.g., Fawbush and Miller 1953; Moore and Pino 1990; Brimelow et al. 2002) are based on information extracted from a sounding representative of the convective environment. All of these models estimate potential updraft strength from the buoyancy between the melting layer and the equilibrium level as the integrated buoyancy provides an estimate of the maximum potential updraft speed in a given environment. In order to determine the hail size at the ground, the hail size models also include a melting effects component based on the height of the wet-bulb temperature zero level, which is approximately where melting would begin in hail falling through downdraft air. Shallower, cooler, and drier melting layers lead to less melting; and larger hailstones have higher terminal fall speeds, leading to shorter transit times (Johns and Doswell III 1992).

The sounding-based hail forecasting methods have fundamental limitations that limit their skill. The sounding used to assess hail potential should be representative of the storm environment, which can be problematic when observed soundings are generally available only twice a day. This issue can be addressed by either correcting the sounding based on recent surface observations (Moore and Pino 1990) or by utilizing model forecast soundings near the time and location of expected hail. The updraft strength estimated from the sounding may not be representative of the updrafts that actually produce hail. Updraft speed computed from CAPE is estimated at the top of the updraft and not within the hail growth region. The largest hail may not come from the strongest part of the updraft because the hail embryos may be lofted out of the hail growth region if the speed is too high (Johns and Doswell III 1992). In storms with tilted updrafts, hailstones may have significant horizontal motions in their growth trajectories (Nelson 1983), leading to additional growing time.

Some of the issues with purely sounding-based hail diagnostics were addressed by feeding sounding information into a 1-dimensional hail growth model. This approach, called HAILCAST (Brimelow et al. 2002), creates an ensemble of updrafts based on perturbations of the temperature and moisture profile. A parameter called the Energy Shear Index governs the predicted lifetime of the updraft and the amount of entrainment that will reduce updraft speed. Hail embryos are then released into the updraft and grow until they can no longer be sustained by the updraft or the updraft collapses. A bulk melting scheme is then applied to approximate the hailstone size at the ground. Jewell and Brimelow (2009) found that HAILCAST generally provides a reliable forecast of hail size, especially in comparison to other sounding based methods. Adams-Selin et al. (2014) has incorporated a variation of HAILCAST directly into the WRF model that uses modeled vertical velocities in place of those estimated from a sounding profile.

2.5.2 Solar Irradiance Forecasting

With the rapid growth of electricity generation from solar energy, there is a greatly increased interest in forecasting solar irradiance at the surface. Solar irradiance, also known as global horizontal irradiance (GHI), is defined as the amount of radiative energy from the sun striking a flat surface of a fixed area over a fixed time period and is generally recorded in units of W m^{-2}. GHI can be decomposed into two components, as shown in Eq. 2.5.

$$\mathrm{GHI} = \mathrm{DHI} + \mathrm{DNI}\cos(\theta_z) \tag{2.5}$$

The two components are direct normal irradiance (DNI), the radiation coming directly from the sun, and diffuse horizontal irradiance (DHI), the total radiation reflected by the sky and clouds (Eq. 2.5). DNI is multiplied by the solar zenith angle θ_z to determine the component striking a horizontal surface. The amount of solar irradiance at a particular location and time is subject to factors with varying degrees of predictability from directly computable based on simple geometry to highly uncertain. The top-of-atmosphere irradiance depends on the distance between Earth and the sun, which varies cyclically based on Earth's orbit. Diurnal effects are determined by calculating the solar zenith and azimuth angles for a location and time. At higher zenith angles, solar radiation has to travel a longer distance through the atmosphere, leading to a larger DHI component. Clouds absorb and reflect solar irradiance to varying degrees depending on their thickness, height, composition, and position relative to the sun. Aerosols directly absorb a fraction of solar radiation and can indirectly lead to the formation of clouds (Jimenez et al. 2016).

The best type of forecast guidance for solar irradiance forecasting depends heavily on the lead time (Diagne et al. 2013). On timescales of a few minutes to an hour, persistence and autoregressive models are hard to beat. From 1 to 6 hours, statistical models and advection of clouds from satellite data perform well. From 6 hours to multiple days, NWP models combined with a MOS-type

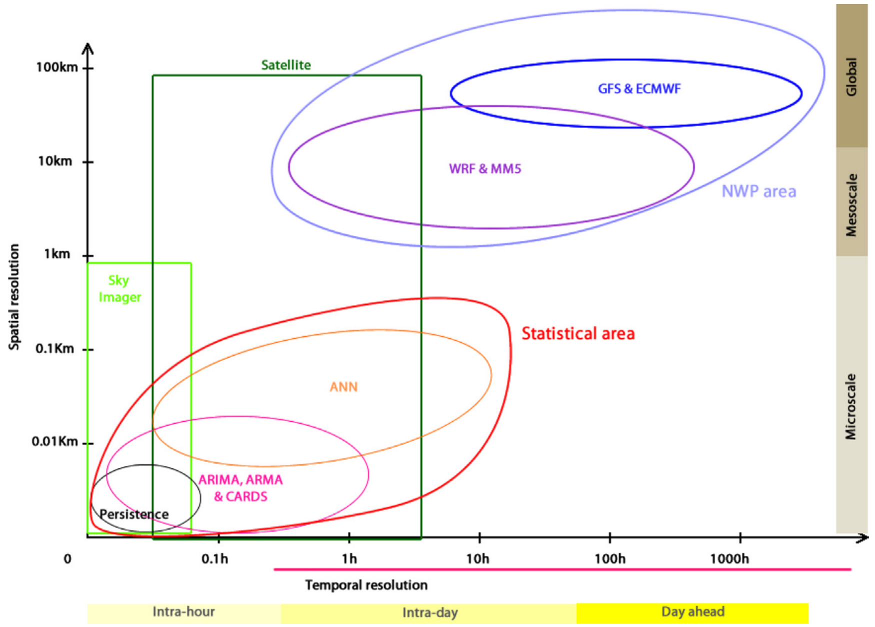


Figure 2.4: Appropriate solar irradiance forecasting models for different time and spatial scales from Diagne et al. (2013).

correction will outperform any purely statistical or extrapolation based methods (Diagne et al. 2013; Lorenz et al. 2009). For MOS-like systems, the modelers either spatially average and bias-correct solar irradiance before feeding it into a simulated PV system (Lorenz et al. 2009), or they predict solar power output directly using an archive of PV power output (Zamo et al. 2014). Lorenz et al. (2009) found that spatial averaging of NWP solar irradiance helped improve predictions by better accounting for cloud effects. Zamo et al. (2014) tested a wide range of machine learning models on predicting PV power output from a group of plants in France, and found that random forests produced the lowest errors and outperformed gradient boosting, linear regression, and support vector regression even after having them all optimized through grid search and cross-validation. While a wide range of NWP variables were provided, shortwave solar irradiance, sun angle, low level relative humidity, and low level cloud cover recorded high values of variable importance (Zamo et al. 2014). Since low-level clouds tend to be thicker and closer to the surface, they are more likely to block the sun than higher level clouds. The random forest in this case is properly ordering the degree of physical impact from each type of variable.

2.6 Forecast Evaluation

The goodness of a forecast is a function of its consistency, value, and quality (Murphy 1993). *Consistent* forecasts reflect the best judgement of the forecaster and are not skewed to to maximize a verification metric. Forecast *value* is a function of how much an end user benefits from the information in a forecast. Forecast *quality* is determined by how well forecasts correspond with observations. Consistency is determined by the forecaster or forecast system and is difficult to assess directly. Value is dependent on the end user and their circumstances and can vary greatly such that a lower quality forecast could potentially have more value. This dissertation focuses on evaluating different aspects of forecast quality. Forecast quality is described by several related attributes that can be expressed as scalar scores (Wilks 2011). Wilks (2011) and Murphy (1993) describe seven of the primary aspects of forecast quality that manifest themselves through different evaluation techniques. The attributes are described in Table 2.2. Each attribute is linked to an aspect of the joint probability distribution between forecasts and observations.

Most types of forecasts can be reduced to a deterministic prediction of whether or not a certain event will occur. Verification of binary discrete forecasts is performed with a binary contingency table (Table 2.3). The table represents

Table 2.2: Descriptions of the attributes of forecast quality based on Murphy (1993). The probability of a forecast value is $p(f)$ while the probability of an observed value is $p(x)$.

Attribute	Description	Distributions
Accuracy	Correspondence between individual forecasts and observations.	$p(f,x)$
Skill	Accuracy of forecast relative to accuracy of a reference forecast.	$p(f,x)$
Systemic bias	Difference between mean forecast and mean observation.	$p(f)$ & $p(x)$
Reliability	Correspondence between conditional mean observation and conditioning forecast, aggregated over all forecasts.	$p(x\|f)$ & $p(f)$
Resolution	Difference between conditional mean observation and and unconditional mean observation, aggregated over all forecasts.	$p(x\|f)$ & $p(f)$
Sharpness	Variability of forecast distribution.	$p(f)$
Discrimination	Correspondence between conditional mean forecast and conditioning observation, aggregated over all observations.	$p(f\|x)$ & $p(x)$

Table 2.3: Example of a binary contingency table.

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

Table 2.4: Verification scores derived from a binary contingency table.

Score Name	Equation
Percent Correct (PC)	$\frac{TP+TN}{TP+TN+FP+FN}$
Critical Success Index (CSI)	$\frac{TP}{TP+FP+FN}$
Probability of Detection (POD)	$\frac{TP}{TP+FN}$
Probability of False Detection (POFD)	$\frac{FP}{TN+FP}$
False Alarm Ratio (FAR)	$\frac{FP}{TP+FP}$
Success Ratio (SR)	$\frac{TP}{TP+FP}$
Frequency Bias (Bias)	$\frac{TP+FP}{TP+FN}$

the joint distribution of all possible forecast and observation pairs. Scores describing many properties of the forecasts can be calculated from the contingency table (Table 2.4). Since high impact weather events tend to be rare, there is more interest in predicting the positive events correctly. However, the negative events usually far outweigh positive events in frequency, so statistics that give equal weight to positive and negative events or only focus on negative events, such as Percent Correct and Probability of False Detection, will still return really high scores even if the positive events are poorly forecasted. Critical Success Index (CSI; Gilbert 1884) serves as a better measure of accuracy for rare events by ignoring true negatives entirely. Frequency bias determines if an event is being overforecasted (bias > 1) or underforecasted (bias < 1). False Alarm Ratio (FAR) is a measure of reliability for a positive event while Probability of Detection (POD) measures the ability of the forecast to discriminate between positive and negative events (Wilks 2011).

Probabilistic forecasts of binary events are evaluated using diagrams and scores that examine the different properties of forecast quality at each probability threshold. The accuracy of probability forecasts is evaluated with the Brier

Score (Brier 1950), which shows the squared difference between probability forecasts and the occurrence of the event. The Brier Score can be decomposed into three terms in Eq. 2.6:

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}), \tag{2.6}$$

where N is the number of forecasts, K is the number of probability bins, n_k is the number of forecasts in each bin, p_k is the forecast probability, $\bar{o}_k$ is the observed relative frequency for a given probability, and $\bar{o}$ is the climatological relative frequency (Murphy 1973). The first term indicates the reliability of the probabilities as the distance between the forecast probability and observed relative frequency, the second term determines the resolution as the difference between the observed and climatological relative frequencies, and the uncertainty reflects the underlying predictability of the problem. Rare events tend to have low uncertainty. Good probability forecasts minimize the reliability term while maximizing the resolution term. The attributes diagram (Hsu and Murphy 1986) is a graphical representation of the Brier Score decomposition. An example of an attributes diagram is shown in Fig. 2.5. The diagram expands on the reliability diagram, which plots the forecast probability versus the observed relative frequency, to include lines demarcating “No Skill” and “No Resolution”. The “No Skill” line marks the points where the reliability term equals the resolution term, which results in a Brier Skill Score of 0 (Hsu and Murphy 1986). The “No Resolution” line indicates probability forecasts that have the same observed relative frequency as the climatological probability. The attributes diagram can be used to assess whether the forecasts for a particular probability threshold are contributing resolution and positive skill. Forecasts

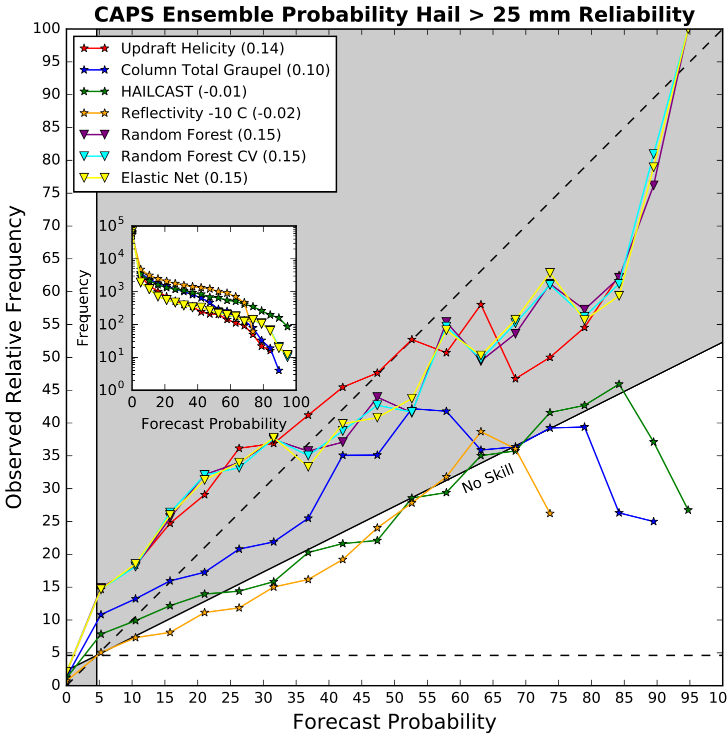


Figure 2.5: An example of an attributes diagram.

with negative Brier Skill Scores can still be useful if the slope of their reliability curve is positive, which indicates the potential for additional calibration at the expense of sharpness (Wilks 2011).

The ability of a probabilistic forecast to discriminate between two outcomes can be assessed using a Relative Operating Characteristic (ROC) diagram (Mason 1982). The diagram (Fig. 2.6) is a plot of Probability of False Detection on the x-axis versus Probability of Detection on the y-axis. A set of thresholds are chosen to make binary deterministic forecasts from probabilistic on continuous-valued forecasts, and binary contingency tables are constructed for each threshold. The POD and POFD are calculated at each threshold to form a curve. At the lowest threshold, all forecasts are yes forecasts, which results in a POD of 1 and a POFD of 0. At the highest threshold, the opposite is

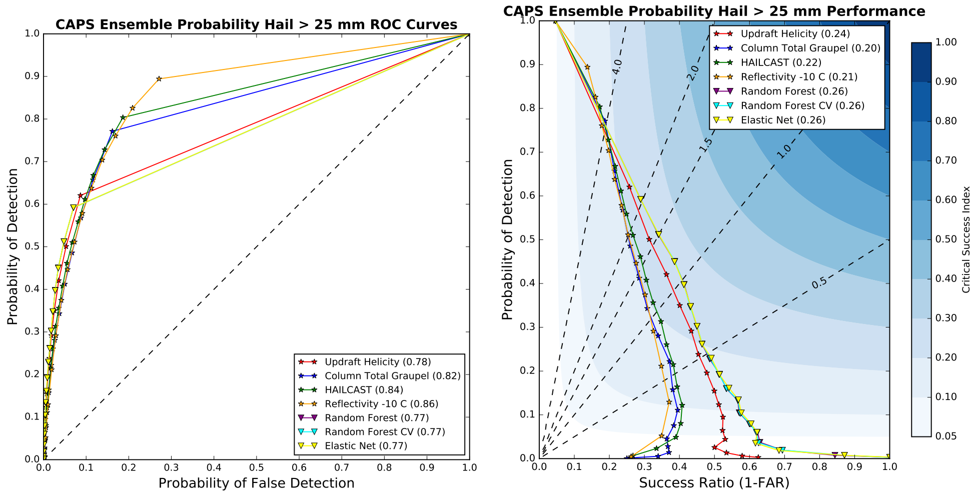


Figure 2.6: Examples of ROC (left) and performance (right) diagrams.

true, resulting in a POD of 0 and POFD of 1. If the POD equals the POFD, then the forecast is no better than a random or uniform forecast. Positive skill over climatology occurs when the POD exceeds the POFD. The area under the ROC Curve (AUC) is a summary metric for the skill across all thresholds with 1 indicating a perfect forecast and 0.5 equal to a random forecast. The ROC curve is insensitive to calibration and the underlying distribution of the event in question, which makes it a good estimator of potential skill but could lead to poor conclusions if calibration is important (Wilks 2011). Because ROC curves weigh positive and negative events equally, ROC curves may not be the best tool for evaluating rare event forecasts. The performance diagram (Roebber 2009), a variation on the ROC curve that replaces the probability of false detection with the false alarm ratio along the x-axis, is able to display measures of accuracy, bias, reliability, and discrimination all in one diagram (Fig. 2.6). It is more sensitive to the ability to predict positive events because all the statistics plotted ignore true negatives (Roebber 2009).

Evaluation of continuous-valued deterministic forecasts can also be framed in a distributions-oriented way. Accuracy is assessed through the mean squared error (MSE) and mean absolute error (MAE). The MSE is the mean of the squared difference between each forecast and observation pair while the MAE is the mean of the absolute value of the difference between each forecast and observation pair. MSE is differentiable for all possible error values, which makes it a popular choice for loss functions (Hastie et al. 2009), but the squared error penalizes large deviations heavily and makes the score more sensitive to outliers. MAE linearly weights errors, so it is viewed as a more robust score. Mean error (ME), or the mean of the differences between forecasts and observations, is a measure of systemic bias. Reliability and discrimination can be assessed by binning the continuous forecasts and examining the marginal distributions of the observations conditioned on the forecasts and forecasts conditioned on the observations, respectively. Sharpness is proportional to the variance of the forecasts.

One major goal of forecast evaluation is determining which forecast system produces the highest quality forecasts. In order to do so, multiple forecasting systems are evaluated over a large sample of cases using a common set of measures and are ranked. If two forecast systems have similar scores, then their differences and relative rankings may be due to random chance within the sample rather than forecast system design. Statistical hypothesis testing can help determine if two sets of forecasts originate from the same distribution or not. Traditional hypothesis testing requires the assumption of certain properties of the forecast distribution, such as normality. Resampling tests, however, are non-parametric and only require that each item in a sample be independent of the others (Efron and Tibshirani 1994).

Bootstrapping is a method for calculating confidence intervals for any arbitrary statistic by repeatedly resampling a dataset with replacement and calculating the statistic on each of the replicate samples. The percentiles of the bootstrap statistic distribution then are used as confidence intervals for that statistic. The width of the bootstrap confidence interval indicates the uncertainty of the statistic calculated on the dataset, which is a function of the variance and sample size of the dataset. The difference between two samples can be inferred as statistically significant if the bootstrap distribution of the difference in sample statistics does not overlap 0. If two bootstrap distributions do not overlap each other, statistical significance can be inferred; but if there is overlap, statistical significance may still be possible depending on the amount of overlap (Cumming and Finch 2005).

If bootstrap confidence intervals computed independently for each forecasting system overlap, then statistically significant differences in performance can still be inferred if the different models forecast the same cases. If the case indices are resampled the same way for each model, then the performance statistics calculated for each bootstrap replicate can be ranked. If one model is ranked higher than another model for most bootstrap samples, then there is more confidence that the difference in scores is statistically significant even if the actual difference is small. Small but consistent differences in performance are important in applications where many forecasts are used, and forecast errors lead to decisions that incur a certain cost. The cumulative cost reduction of a small but consistent forecast improvement could be very significant, particularly in a domain like electric utility forecasting where power purchasing decisions are made hourly.

Permutation tests can estimate the p-value of the difference in a statistic between two paired sets of data (Efron and Tibshirani 1994). Paired data means that each item in one sample is a paired with another item from the other sample, such as forecasts from two different models for the same event. In a permutation test, the difference in test statistics is calculated for the original samples. Then a null distribution of that statistic is computed by randomly shuffling paired items from one sample to the other and recalculating the statistic a large number of times under the null hypothesis that both samples come from the same population. The percentile of the original statistic in the null distribution is its p-value. In cases where the bootstrap confidence intervals of paired data overlap with each other, a permutation test can be used to determine if the difference is statistically significant.

Performing multiple statistical hypothesis tests on the same dataset increases the probability that at least one of the tests will falsely indicate statistical significance. This is known as the multiple comparisons problem and is an increasingly important issue as the dimensionality of datasets continues to increase (Jensen and Cohen 2000; Lindquist and Mejia 2015). The challenge is to find a significance threshold α that accounts for the increased chance of false alarms while still minimizing the probability of missing a result that is statistically significant. The Bonferroni correction (Dunn 1961) is a simple but conservative way to correct the p-values of individual hypothesis tests to maintain a desired Family-Wise Error Rate by dividing the original p-value by the number of hypothesis tests. The resulting individual p-values tend to be very small, which increases the likelihood of falsely failing to reject the null hypothesis. An alternative approach is to minimize the False Discovery Rate by determining the number of null hypotheses to reject based on the distribution of ranked p-values

(Benjamini and Hochberg 1995). First the p-values from each hypothesis in a group of size N are ranked from smallest to largest, and then the p-value of the false discovery rate p_{FDR} is calculated such that:

$$p_{FDR} = \max_{j=1,\ldots,N} \left\{ p_{(j)} : p_{(j)} \leq \frac{j}{N} \alpha_{\text{global}} \right\}, \tag{2.7}$$

where α_{global} is the desired global significance level (Wilks 2006, 2011). Null hypotheses with p-values below the global significance threshold but above p_{FDR} are not rejected. This test is less conservative than the Bonferroni correction and results in fewer false negatives at the expense of an expected number of false positives.

Chapter 3

Day-Ahead Hail Prediction Integrating Machine Learning with Storm-Scale Numerical Weather Models

Hail, or large spherical ice precipitation produced by thunderstorms, has caused billions of dollars in losses by damaging buildings, vehicles, and crops (Changnon 2009). Economic losses from hail have been increasing over the past two decades as populations have increased and cities have expanded in the hail-prone regions of the central United States (Changnon et al. 2000; Rosencrants and Ashley 2015). Some losses from hail could be mitigated with accurate forecasts of severe hail potential that give people and companies time to protect vehicles and property from an incoming hailstorm.

Forecasting hail size and location is a challenging problem for meteorologists due to major uncertainties in both the forecasting and observing processes. Unlike more traditional meteorological conditions such as temperature and rainfall, hail size is not measured directly by automated instruments. The primary source of empirical observations comes from humans estimating the largest size found at their location, and hail size estimated from radar is calibrated on those imperfect human observations. Within a storm, hail size can vary dramatically and is generally not spatially contiguous. Accurate hail forecasts require predictions about the characteristics of potential hail-producing storms and the environmental conditions surrounding them. Ensembles of numerical weather prediction models can estimate the range of possible atmospheric conditions and can partially resolve the individual storm cells that produce hail up to a day in advance (Clark et al. 2012b). Current numerical models do not produce explicit hail size forecasts. Hail potential can be inferred indirectly through proxy variables related to storm intensity (Clark et al. 2013), or more directly through a physical (Brimelow et al. 2006) or a machine learning model (Manzato 2013) approach linking atmospheric conditions to the largest possible hail size in a given area and time period. While previous studies have focused on predicting hail sizes over large areas and time period, this chapter investigates how the latest high-resolution numerical weather prediction model output can be integrated with machine learning models to predict hail potential over more specific areas and times. Because of the much larger data volumes associated with these models, I adapted advanced techniques from the image processing and machine learning fields to make hail predictions in an operational setting.

The purpose of this chapter is to describe and evaluate techniques for producing day-ahead, hourly forecasts of hail diameter using storm-scale numerical weather prediction models, image processing, and machine learning. The hypothesis is that the machine-learning-based techniques equal or exceed the performance of a physics-based hail size model. Forecasts from both machine-learning and physics-based techniques were generated during the 2014 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed Experimental Forecast Program (EFP) and were evaluated statistically and subjectively by teams of research and operational meteorologists (Clark et al. 2012b). Hail size forecasts are derived from each ensemble member by identifying forecast hailstorms, matching the forecast storms with observed hailstorms, extracting data within the storm areas, and then fitting a machine learning model between the atmospheric variables and the observed hail size. Forecasts are produced for whether or not any hail will occur, the maximum hail diameter produced from a particular storm, and the probability of hail at least 25.4 mm (1 inch) in diameter within 40 km of a point, which are the size criteria for severe hail and the spatial verification threshold used by the National Weather Service.

3.1 Hail Observations

Developing a machine learning model to predict hail requires a reliable estimate of hail spatial coverage and diameter. No automated network exists to detect hail at the ground, so hail size observations come from either storm spotter reports or estimates derived from NEXRAD radar. Reports of hail at least 1 inch in diameter are collected by the NOAA National Weather Service Storm Prediction Center (SPC). The database is extensive and publicly available, but it suffers from many limitations. The recorded hail diameters are often estimated by comparing the stone to analog objects, such as golf balls. This estimation technique results in unnatural peaks in the hail size distribution (Jewell and Brimelow 2009). The locations of hail in the dataset are also biased toward population centers and major highways.

Radar-estimated hail size offers a solution to the population bias issue plaguing hail reports. The NOAA NSSL Multi-Radar Multi-Sensor (MRMS) gridded Maximum Estimated Size of Hail (MESH), which derives a maximum hail size from gridded 3D radar reflectivity (Witt et al. 1998), is used as the best approximation for the observed hail size. A multi-year comparison of MRMS MESH to storm reports found that MESH was unbiased and had superior spatial coverage to hail reports (Cintineo et al. 2012). The native MESH data were interpolated to the model domain using cubic spline interpolation.

3.2 Storm-Scale Ensemble

Output from the Center for Analysis and Prediction of Storms (CAPS) Storm-Storm Scale Ensemble Forecast (SSEF) system (Kong 2014), which was run in conjunction with the NOAA Hazardous Weather Testbed Experimental Forecast Program, is used as the forecast input into the machine learning models. The SSEF consists of an ensemble of Weather Research and Forecasting (WRF) Advanced Research WRF models with randomly perturbed initial and boundary conditions. In addition, each ensemble member used a different combination of microphysics (physics describing how water changes phase and grows into precipitation), planetary boundary layer (atmosphere near the surface), and land surface model (vegetation and soil processes) parameterization schemes in order to increase the diversity of model solutions. Each SSEF run was initialized at 00 UTC and produced hourly output during the period from late April to early June. The 2013 SSEF was used to train and validate the machine learning models while the 2014 SSEF was used for testing. The 2013 SSEF consisted of 30 model runs from 26 April to 7 June 2013, and the 2014 SSEF consisted of 12 model runs between 15 May and 6 June 2014. The 18 to 30 hour forecasts valid from 18 to 6 UTC are evaluated as they cover the time frame when hailstorms are most likely and contain storms that were not present when the SSEF initiated.

3.3 Machine Learning Framework

The procedure for extracting storm information from NWP models and inputting it into machine learning models to produce hail size forecasts is described in Fig. 3.1.

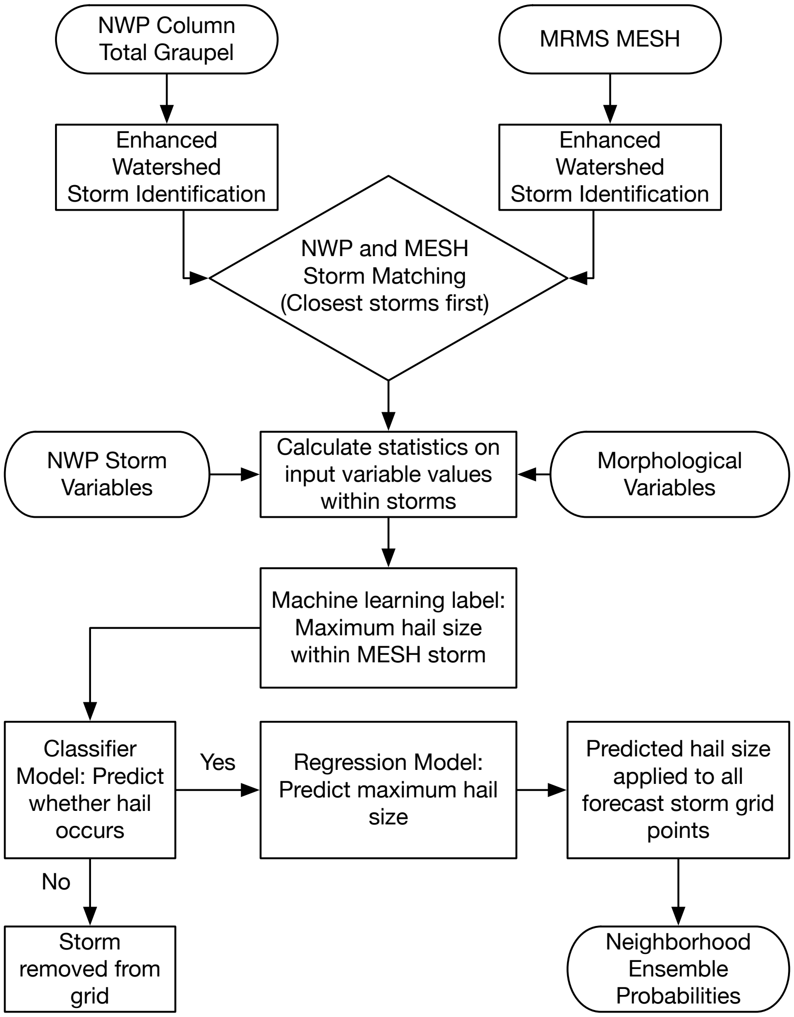
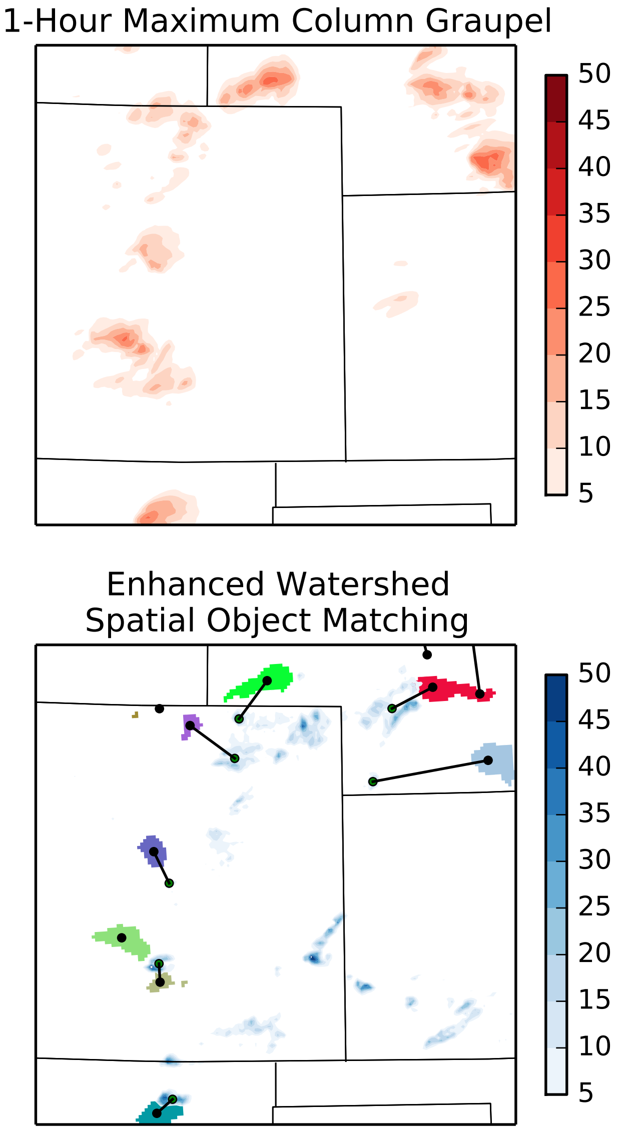

NWP Column Total Graupel
MRMS MESH
Enhanced Watershed Storm Identification
Enhanced Watershed Storm Identification
NWP and MESH Storm Matching (Closest storms first)
NWP Storm Variables
Calculate statistics on input variable values within storms
Morphological Variables
Machine learning label: Maximum hail size within MESH storm
Classifier Model: Predict whether hail occurs
Yes
Regression Model: Predict maximum hail size
Predicted hail size applied to all forecast storm grid points
No
Storm removed from grid
Neighborhood Ensemble Probabilities

Figure 3.1: A summary of the procedure for predicting hail size using machine learning models.

3.3.1 Hailstorm Identification and Matching

Hail size prediction first requires determining the areas in which hail is likely to occur. Model atmospheric conditions related to hail should only occur in the areas where the model produces ice-containing storms, so identifying likely storm areas in the model both reduces the noise in the training data and greatly reduces the required computational power. To find ice-containing storms, I examine the 1-hour maximum column total graupel field, which indicates the maximum value over the previous hour of the total mass of spherical ice particles in a column of air. For object identification, I use the enhanced watershed technique (Lakshmanan et al. 2009). As with the traditional watershed, local maxima in the column total graupel field are first identified, and then objects are grown from the maxima in discrete steps until stopping criteria are met. While the traditional watershed uses a global lower threshold or maximum number of steps as its stopping criteria, the enhanced watershed also includes an area criterion and buffer zones around local maxima. Prior to applying the enhanced watershed to the data, a Gaussian filter was applied to each grid in order to increase spatial correlations and generate smoother objects.

The enhanced watershed is applied to both the model column total graupel fields and the observed MESH field. The enhanced watershed was manually tuned to capture a wide range of hail swath intensities while keeping neighboring swaths as separate objects. Forecast and observed hailstorm objects are matched iteratively based on their Euclidean centroid distance. The closest objects are matched first, then the next closest, and so on until all unpaired objects under 200 km apart are matched with one other object. Since each observed hail object can only be matched with one forecast hail object, some storms near isolated hail observations do not get matched. An example of the enhanced watershed and object matching being applied is shown in Fig. 3.2.


1-Hour Maximum Column Graupel

Enhanced Watershed
Spatial Object Matching

Figure 3.2: In the first panel, the 1-hour maximum column-summed graupel from a member of the SSEF at 22 UTC on 6 June 2014 is shown. The second panel shows the hailstorm objects extracted from the column graupel grid by the enhanced watershed technique in solid colors. The connecting lines indicate the closest matches between the forecasted hailstorms and observed MESH (blue contours) objects.

Once storms are identified and matched, statistics describing different properties of the storm and atmosphere are extracted from each hailstorm object. These statistics include the mean, standard deviation, minimum, and maximum of WRF output variables describing the strength of the storm as well as the conditions of the storm environment (Table 3.1). The forecast label is the maximum hail size within the matched MESH object, or 0 if no match was found.

3.3.2 Hail Classification and Size Regression

Machine learning models first determine if a specific forecast storm will produce any hail, and given that the storm does produce hail, what size the hail will be. A classification model was trained on all cases to produce a binary prediction of whether or not the storm would produce hail, and a regression model was trained on only the storms that were matched with an observed hail event. Three machine learning models are tested: random forest, gradient boosting regression trees, and a combination of a logistic classification model and Ridge regression. All methods were implemented using the Python *scikit-learn* library (Pedregosa et al. 2011). Random forests (Breiman 2001a) are ensembles of decision trees that use bootstrap resampling of the training data and random subsampling of input variables to increase the diversity of the member decision trees and improve predictive accuracy. For this experiment, a 100-tree random forest with the default parameters for *scikit-learn* was used. Gradient boosting regression trees (Friedman 2001) are a stagewise, additive ensemble of decision trees that are iteratively trained to predict the residuals of the summed predictions from all of the previous trees. The contribution of each tree to the final

Table 3.1: Input variables for the machine learning models from the SSEF ensemble members. Storm variables (S) describe conditions within the storm and environment variables (E) describe the surrounding atmosphere.

Variable	Description	Units
Max Updraft Speed (S)	Upward vertical wind speed	m s^{-1}
Max Downdraft Speed (S)	Downward vertical wind speed	m s^{-1}
Max Updraft Helicity (S)	Proxy for updraft intensity	m^2 s^{-2}
Radar Reflectivity (S)	Simulated radar intensity	dBZ
Max Column Graupel (S)	Total mass of ice particles	kg m^{-2}
0-5 km Total Graupel (S)	Mass of ice particles	kg m^{-2}
Storm Height (S)	Height of top of storm	m
Bunker's Storm Motion (S)	Storm speed and direction	m s^{-1}
Mean Layer CAPE (E)	Mean instability	J kg^{-1}
Most Unstable CAPE (E)	Max possible instability	J kg^{-1}
Mean Layer CIN (E)	Mean Inhibition	J kg^{-1}
Most Unstable CIN (E)	Lowest possible inhibition	J kg^{-1}
Lifted Condensation Level (E)	Distance to cloud base	m
Precipitable Water (E)	Water contained in air column	mm
0-6 km Wind Shear (E)	Magnitude of wind difference	m s^{-1}
0-3 km Storm-Rel. Helicity (E)	Horizontal rotation	m^2 s^{-2}
0-3 km Lapse Rate (E)	Vertical temperature change	K km^{-1}
850 mb Specific Humidity (E)	Water vapor amount	g kg^{-1}

prediction is regulated by a learning rate that scales the prediction from each tree. The gradient boosted regression tree models in this experiment used 1000 trees, a learning rate of 0.05, and a maximum tree depth of 5. Both methods have produced strong predictive performance in many domains and both can be analyzed using variable importance measures and partial dependence plots. A logistic regression is a linear classifier that translates input parameters into a probability through a logistic function (Wilks 2011). Ridge regression is a form of linear regression with a penalty term to restrict the size of the coefficients and make the regression more robust (Hoerl and Kennard 1970). When the predicted hail sizes were applied to the original forecast grid, the storms producing no hail were removed from the grid, and the predicted size values were applied to the grid points within the area covered by each forecast hail storm.

3.3.3 HAILCAST

HAILCAST is a one-dimensional, physics-based coupled cloud and hail model (Brimelow et al. 2002). The technique has been further refined to run during the integration of a storm-scale numerical model (Adams-Selin et al. 2014) and has been released publicly in WRF version 3.6. In addition to being run during the 2014 EFP, HAILCAST has been incorporated into the operational Air Force Weather Agency storm-scale ensemble. HAILCAST is run at each SSEF member grid point with an updraft speed at least 10 m s^{-1}. The maximum HAILCAST hail size within each forecast hailstorm object was used as the comparison prediction with the machine learning methods because it provided the most analogous estimate to the observed maximum hail size.

3.3.4 Neighborhood Ensemble Probability

The machine learning methods produce a calibrated hail size forecast for each ensemble member and each time step. These machine learning forecasts do not cover the full range of possible hail sizes at every grid point because the SSEF contains spatial and temporal errors in storm placement and intensity and does not fully approximate internal storm dynamics as well as the processes that govern precipitation formation and thermodynamic changes associated with them. These physics errors results in modeled storms that do not form, move, and intensify at the same rate as the real ones. One approach commonly used to account for this spatial error is the neighborhood ensemble probability method (Schwartz et al. 2010). Conditional probabilities of severe hail are calculated by counting the number of grid points in a local, circular neighborhood in which severe hail occurs and dividing by the number of grid points in which any hail occurs. The probability from all ensemble members are averaged together, and a Gaussian filter is applied to smooth the edges of the non-zero probabilities. Since each model forecast has been bias-corrected by the machine learning regressions, the resulting probabilities should also be unbiased. The size of the neighborhood can be adjusted to capture uncertainties at varying scales. Weather forecasters prefer spatially smooth probabilities as they more closely match human forecasts. The drawbacks of neighborhood ensemble probabilities are that they weaken probability gradients and can understate the threat of single isolated storms while highlighting clusters of more widespread marginal storms.

3.4 Results

I statistically validated the hail size and probability forecasts based on 12 hail days from 15 May to 6 June 2014. The predicted hail sizes were compared with the maximum hail sizes within each matched observed hailstorm object. The probability forecasts were compared at each grid point with whether or not hail at least 25.4 mm in diameter was observed within 40 km of that point, which are the evaluation criteria used by the SPC.

3.4.1 Hail Size Forecasts

The machine learning and HAILCAST size forecasts showed skill in predicting hail sizes up to 60 mm in diameter, which account for the bulk of all hail events. Both tree-based methods predicted that most severe hail would be between 25 and 60 mm, and most of their predictions were close to those values. Observed hail over 60 mm was also predicted to be within the 25 to 60 mm range (Fig. 3.3). While Ridge regression and HAILCAST predicted hail sizes over the full range of observed values, both methods tended to overpredict the maximum hail diameter, especially HAILCAST.

Examining the errors for each ensemble member reveals some links between the error characteristics and the microphysics parameterization scheme used by each member (Fig. 3.4). A lack of overlap of two 95% confidence intervals indicates that their distributions share a probability density of less than 5%, so the difference in errors can be considered statistically significant at the 5% level (Cumming and Finch 2005). The bootstrap confidence intervals for HAILCAST do not overlap with those of any of the machine learning methods for all but

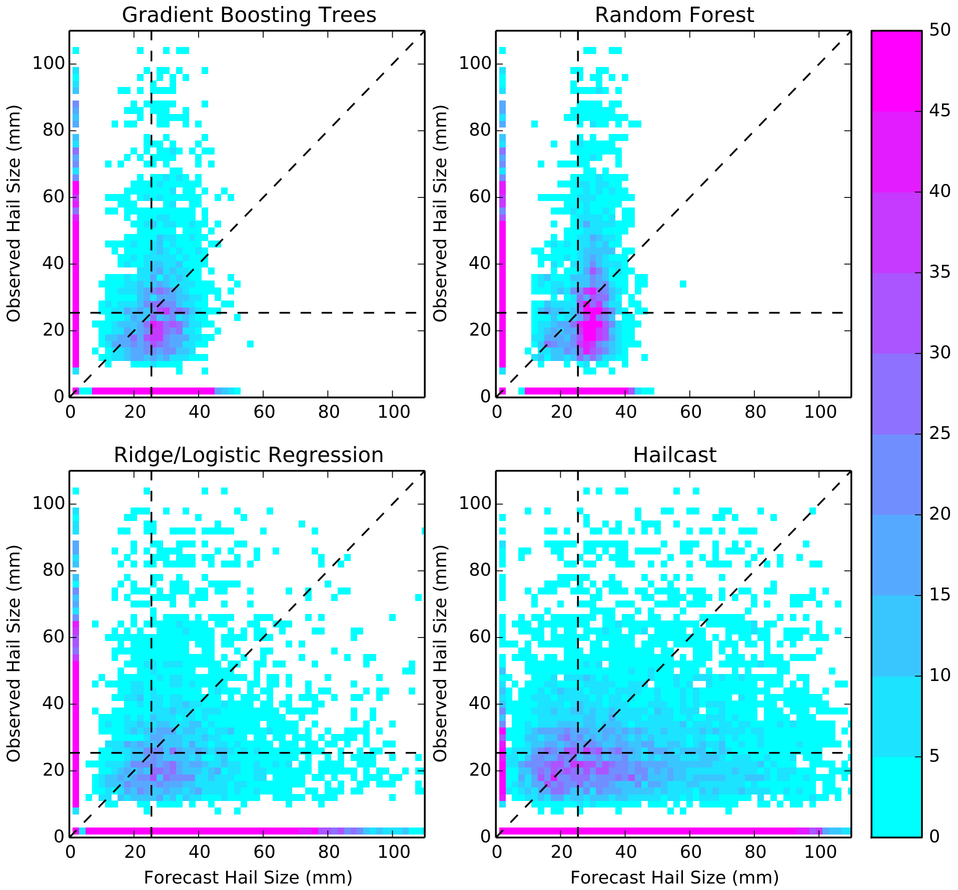


Figure 3.3: Heatmaps of the distributions of forecast errors for each hail size model.

one member. The error was greatest in ensemble members using the Thompson microphysics scheme. The Thompson scheme assumes a relatively larger graupel density compared to the other schemes, which HAILCAST used as the basis for growing its hailstones. The Milbrandt and Yau (MY) scheme has separate graupel and hail densities, and HAILCAST performed best in the members using that scheme and overlapped in confidence intervals with the machine learning models. For the Thompson members, the confidence intervals of the three machine learning methods overlapped with each other. The Gradient Boosting and Random Forest confidence intervals contained errors lower than

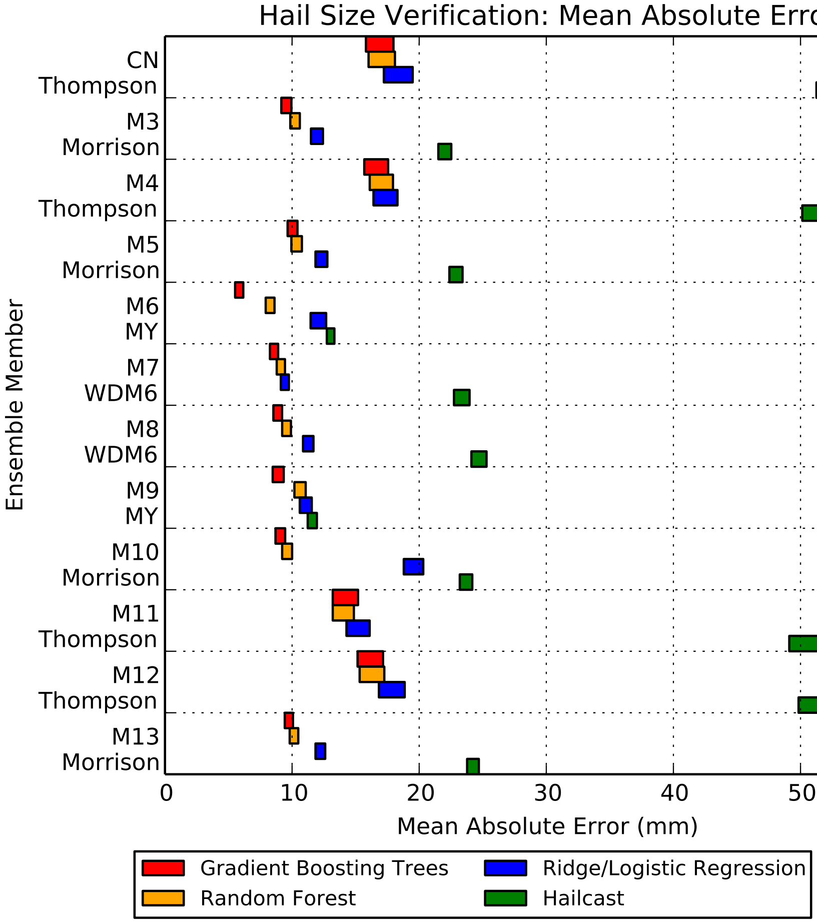


Figure 3.4: Comparison of the 95% bootstrap confidence intervals of the forecast hail size mean absolute error by model type and ensemble member. The microphysics scheme used in each ensemble member is indicated below the name of the member.

linear regression for the Morrison and MY members. The WDM6 members also reported low size errors but little difference among the predictions from the different

The hail occurrence predictions also showed similar skill among all machine learning methods and ensemble members . A performance diagram (Roebber 2009) displays the relationships among four binary contingency table scores (Fig. 3.5). The machine learning methods had similar success ratios, but there was a wider variance in the percentage of hailstorms detected. HAILCAST recorded higher POD and higher CSIs for most members. Some of the performance issues stem from the enhanced watershed parameters fitting storms from some models better than others due to differences in microphysics.

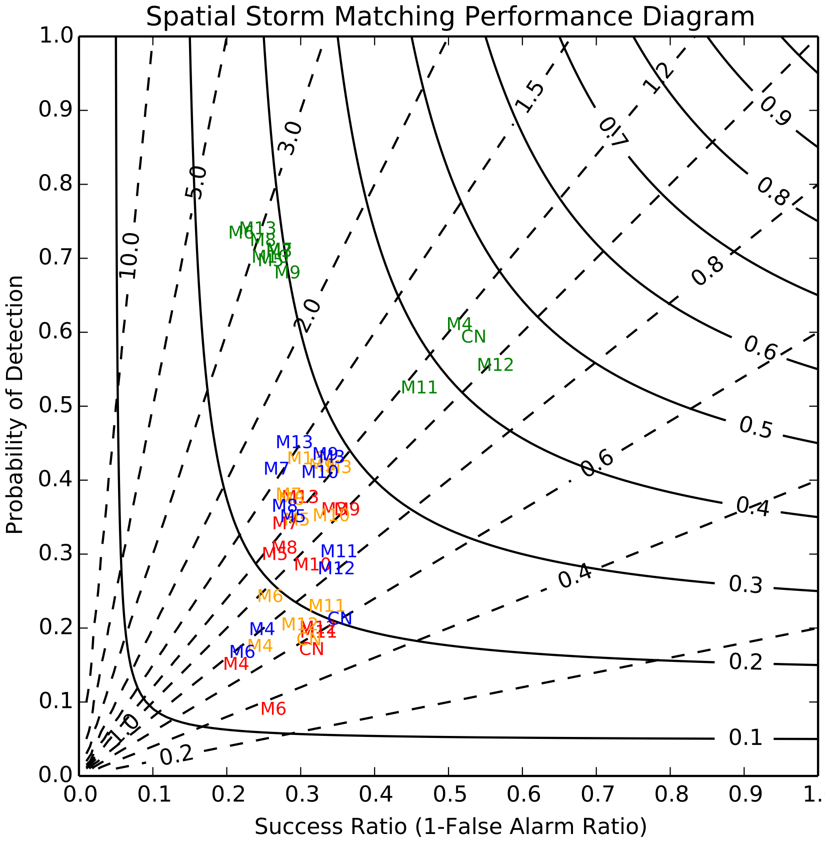


Figure 3.5: A performance diagram measures the quality of each model and ensemble member pairing to match forecast and observed hail storms spatially. The solid contours indicate the critical success index, and the dashed contours indicate frequency bias. Points along the diagonal are unbiased. Models with better accuracy appear closer to the upper right corner of the diagram. In keeping with the same color scheme as Fig. 3.4, the red points are Gradient Boosting Trees, the orange points are Random Forests, the blue points are Linear Regression, and the green points are HAILCAST.

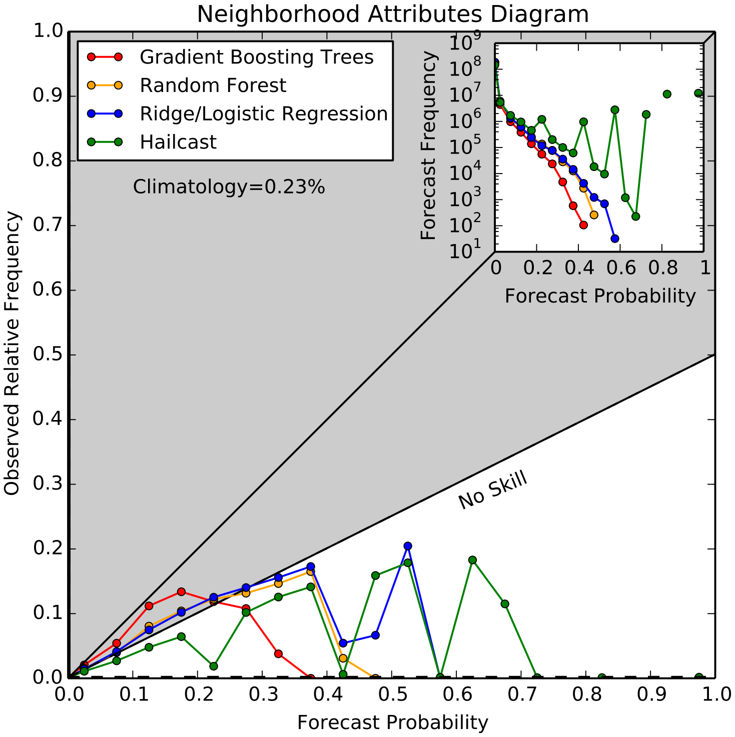


Figure 3.6: Attributes diagram that compares the forecast probabilities of each model with their corresponding observed relative frequencies. Points in the gray area have positive Brier skill scores, and points outside the gray area have negative Brier skill scores. The inset indicates the observed frequency of each probability forecast.

3.4.2 Neighborhood Probability Forecasts

Since the machine learning approaches produced hail forecasts with little bias, the resulting neighborhood probabilities tended to be more reliable, or occurring at the frequency given by the probability, than the corresponding HAILCAST forecasts. For probabilities ranging from 0 to 20%, gradient boosting trees are nearly reliable and the other methods are slightly overconfident while HAILCAST is more overconfident (Fig. 3.6). At higher probabilities, there was overconfidence from all methods. From subjective verification of the different hail forecasts, this overconfidence is linked to a spatial displacement of the highest neighborhood probabilities away from where severe hail fell at a particular time.

3.4.3 Case Studies

The hail event with the largest number of storm reports and greatest amount of property damage during the experiment occurred on 3 June 2014 in Nebraska. Multiple rounds of storms produced wind-driven baseball to softball sized hail that left large dents and holes in cars, crops, and the sides and roofs of houses. Each model generated a neighborhood probability prediction for each hour from 18 to 00 UTC. The maximum 1-hour probabilities during that time period are displayed in Fig. 3.7. All models encompassed the full observed area of 25 mm or greater hail with nonzero probabilities and have their highest confidence in eastern Nebraska, where the largest hail was observed. All models also displayed enhanced probabilities in western Nebraska where isolated storms also produced severe hail. Random forest produced the subjectively best forecast of the machine learning methods because its maximum overlapped the 75 mm hail most closely and because it had relatively lower probabilities for the western Nebraska storms. HAILCAST produced the most confident forecast, but it had high probabilities well outside the area where 25 mm hail was observed.

A more marginal but widespread hail event occurred on 21 May 2014 in Colorado, Kansas, Oklahoma, and Texas. An isolated hailstorm dropped severe hail and caused flooding in downtown Denver, and additional storms dropped hail across eastern Colorado. The ensemble means of the hail size forecasts are shown in Fig. 3.8. HAILCAST and ridge regression generally overestimated the

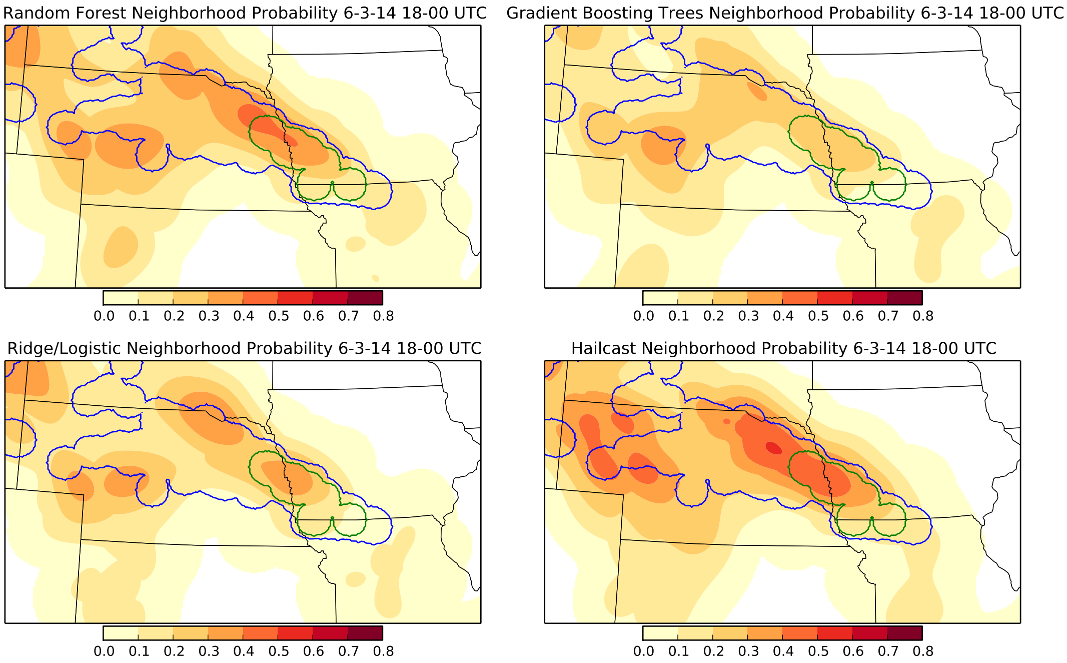

Random Forest Neighborhood Probability 6-3-14 18-00 UTC
Gradient Boosting Trees Neighborhood Probability 6-3-14 18-00 UTC
Ridge/Logistic Neighborhood Probability 6-3-14 18-00 UTC
Hailcast Neighborhood Probability 6-3-14 18-00 UTC
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8

Figure 3.7: Maximum neighborhood ensemble probabilities between 18 and 00 UTC on 3 June 2014. The blue contour indicates the areas that were within 40 km of 25 mm diameter hail, and the green contour indicates the same distance from 75 mm diameter hail.

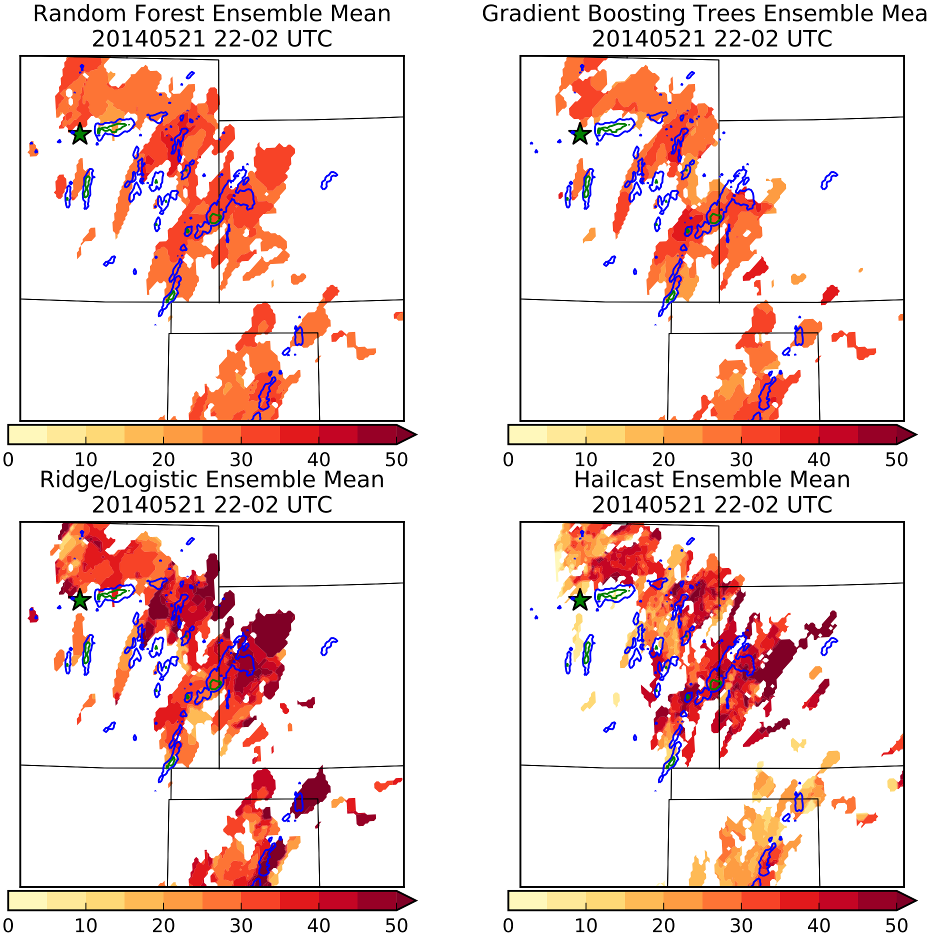


Figure 3.8: Ensemble mean hail sizes from each model for 21 May 2014 from 22 to 02 UTC. Blue contours indicate observed hail sizes of at least 5 mm and green contours indicate hail sizes of at least 25 mm.

maximum hail sizes for the day with widespread areas of over 50 mm hail. Random forest and gradient boosting produced hail sizes closer to what occurred, and gradient boosting also had a wider range of hail sizes than random forest. The most intense portions of the forecast hail swaths were shifted northeast of the observed hail swaths, so while the general character of the event is correctly forecast, downtown Denver was forecast to receive no hail in 3 of the 4 models. The neighborhood probabilities in Fig. 3.9 account for this spatial error and show non-zero probabilities over Denver. The random forest neighborhood

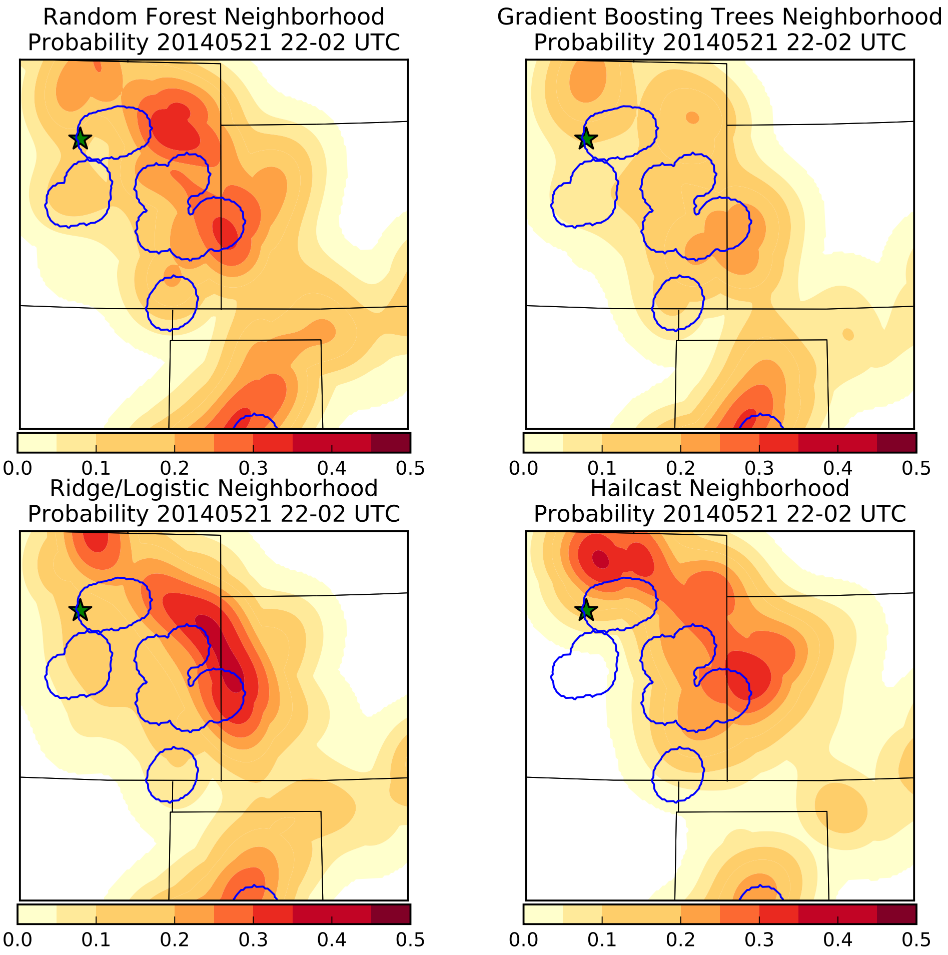

Random Forest Neighborhood
Probability 20140521 22-02 UTC
Gradient Boosting Trees Neighborhood
Probability 20140521 22-02 UTC
Ridge/Logistic Neighborhood
Probability 20140521 22-02 UTC
Hailcast Neighborhood
Probability 20140521 22-02 UTC
0.0
0.1
0.2
0.3
0.4
0.5

Figure 3.9: Neighborhood ensemble probability of severe hail from each model for 21 May 2014 from 22 to 02 UTC. Blue contours indicate hail sizes of at least 25 mm and green contours indicate hail sizes of at least 75 mm. The location of Denver is shown with a green star.

probabilities capture the Colorado hail the best by showing two areas of high hail potential and by having non-zero probabilities of hail over Denver.

3.5 Discussion

Generating and validating daily hail forecasts with a group of experienced meteorologists provided insights about the good qualities of the forecasts and what needed improvements. The machine-learning neighborhood probabilities were useful because the bias-correction reduced the false alarm area compared to HAILCAST. The probability forecasts were closer to the best forecast from a trained meteorologist given the same information. Further improvement to machine learning model performance is constrained by the model storm information. The storm representation can be improved with better resolution, model physics, and initial conditions, but it will always contain uncertainties and errors because we cannot fully observe the atmosphere, the physical models contain approximations, and computational power is limited. While the different machine learning models were not able to predict hail above 60 mm in diameter, this was largely because there was very little training data at these sizes. HAILCAST, on the other hand, predicted hail over 60 mm almost every day during the experiment and was not trusted by the meteorologists because of that issue. This was also the first operational test for both models, and the forecaster feedback has been valuable for introducing improvements to both systems. Overall, the lower size errors and greater reliability of the machine learning models show that they are the superior method for predicting severe hail. However, none of the methods evaluated show good enough performance to predict 50 mm hail consistently. More optimizations need to be performed on both HAILCAST and the machine learning models to capture these events more reliably. This is addressed in the following chapter.

3.6 Conclusions

Hail is a dangerous severe weather phenomenon that causes increasingly extensive economic damage each year. Improving hail prediction with more accurate information about expected hail locations and intensity will allow people to mitigate some of the potential impact of hail. I have demonstrated in an operational setting a hail prediction system that applies machine learning and image processing techniques to storm-scale numerical model ensembles. The approach shows accuracy in predicting hail location and in discriminating its severity with lead times of up to a day in advance of a hailstorm. The machine learning approaches demonstrated some advantages over physics-based hail size calculations. Improvements to the numerical models and machine learning approaches should lead to increasingly accurate hail size and location forecasts.

Chapter 4

Track-Based Day-Ahead Hail Forecasting with Machine Learning

The purpose of this chapter is to describe track-based day-ahead probabilistic hail forecasts generated from machine learning models and compare them with existing storm surrogate and sounding-based hail forecasting methods to determine what information and value the machine learning models add to the forecasting process. The evaluation tests the hypotheses that a machine learning model will produce accurate and reliable hail forecasts, the machine learning forecasts can detect more hail storms and produce fewer false alarms than other hail size methods, and machine learning model performance is more consistent across different NWP model configurations than other hail size diagnostics. The methods introduced here extend the methods described in Chapter 3 by incorporating forecast and observational uncertainty into the forecast model along with improved tracking and matching of forecast and observed hailstorms.

4.1 Methods

4.1.1 Data

Convection-allowing model ensemble output was drawn from two systems with very different design choices. One set of ensemble forecasts came from the

Table 4.1: CAPS ensemble member physics parameterizations for 2014 and 2015. The planetary boundary layer (PBL) schemes tested are the Mellor-Yamada-Janjic (MYJ; Mellor and Yamada 1982), Yonsei University (YSU; Hong et al. 2006), Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi and Niino 2004), and Quasi-Normal Scale Elimination (QNSE; Sukoriansky et al. 2005).

Member	2014 Microphysics	2015 Microphysics	2014 PBL	2015 PBL
CN	Thompson	Thompson	MYJ	MYJ
M3	Morrison	P3	YSU	MYNN
M4	Thompson	MY	QNSE	YSU
M5	Morrison	Morrison	MYNN	MYNN
M6	MY	MY	MYJ	MYJ
M7	WDM6	P3	YSU	YSU
M8	WDM6	P3	QNSE	MYJ
M9	MY	MY	MYNN	MYNN
M10	Morrison	Morrison	YSU	YSU
M11	Thompson	Thompson	YSU	YSU
M12	Thompson	Thompson	MYNN	MYNN
M13	Morrison	Morrison	QNSE	MYJ

Center for Analysis and Prediction of Storms (CAPS) Storm-Scale Ensemble Forecast (SSEF), which consists of 12 WRF-ARW models with varied combinations of microphysics and planetary boundary layer (PBL) parameterizations and perturbed initial and boundary conditions. The 2014 SSEF used 4 km horizontal grid spacing while the 2015 SSEF was reduced to 3 km grid spacing. Both versions of the SSEF were initialized with a 3DVAR data assimilation process that included radar data and started running at 0000 UTC with hourly output for 60 hours. Individual member parameterization configurations are listed in Table 4.1. The SSEF uses the Thompson (Thompson et al. 2008), Morrison (Morrison and Gettelman 2008), Milbrandt and Yau (MY; Milbrandt and Yau 2005), WRF Double Moment 6-class (WDM6; Lim and Hong 2010), and Predicted Particle Properties (P3; Morrison and Milbrandt 2015; Morrison et al. 2015) microphysics schemes. The other set of ensemble forecasts came from the

National Center for Atmospheric Research (NCAR) Ensemble (Schwartz et al. 2015), which consists of 10 3-km grid spacing WRF members initialized from the DART Ensemble Kalman Filter data assimilation system. All members use the Thompson microphysics scheme (Thompson et al. 2008) and the Mellor-Yamada-Janjic (MYJ; Mellor and Yamada 1982) PBL scheme. The ensemble began running daily in April 2015 and will continue running for a full year.

Observations of hail size come from the NOAA National Severe Storms Laboratory Multi-Radar Multi-Sensor (MRMS) radar mosaic (Zhang et al. 2011). MRMS merges radar reflectivity from multiple radars onto a $0.01° \times 0.01°$ uniform grid with 2-minute updates and performs a series of quality control procedures. The MRMS Maximum Expected Size of Hail (MESH) (Witt et al. 1998; Cintineo et al. 2012) product is used for the hail size observations. MESH is a power law relationship between the Severe Hail Index, which is a product derived from radar reflectivity values above the melting level, and hail reports from 9 hail events in 1992 in Oklahoma and Florida. The MESH relationship was calibrated such that the MESH value should exceed 75% of hail reports for a given Severe Hail Index value. Because MRMS provides more complete information about the full depth of a storm than a single radar, and because MESH indirectly accounts for melting effects through melting layer height information and calibration to hail reports, MRMS MESH shows low bias and good spatial coverage when compared with high resolution hail reports (Cintineo et al. 2012).

4.1.2 Storm Object Identification, Tracking, and Matching

Object-based forecasting methods provide the dual advantages of identifying relevant areas while also greatly reducing the amount of data needing to be

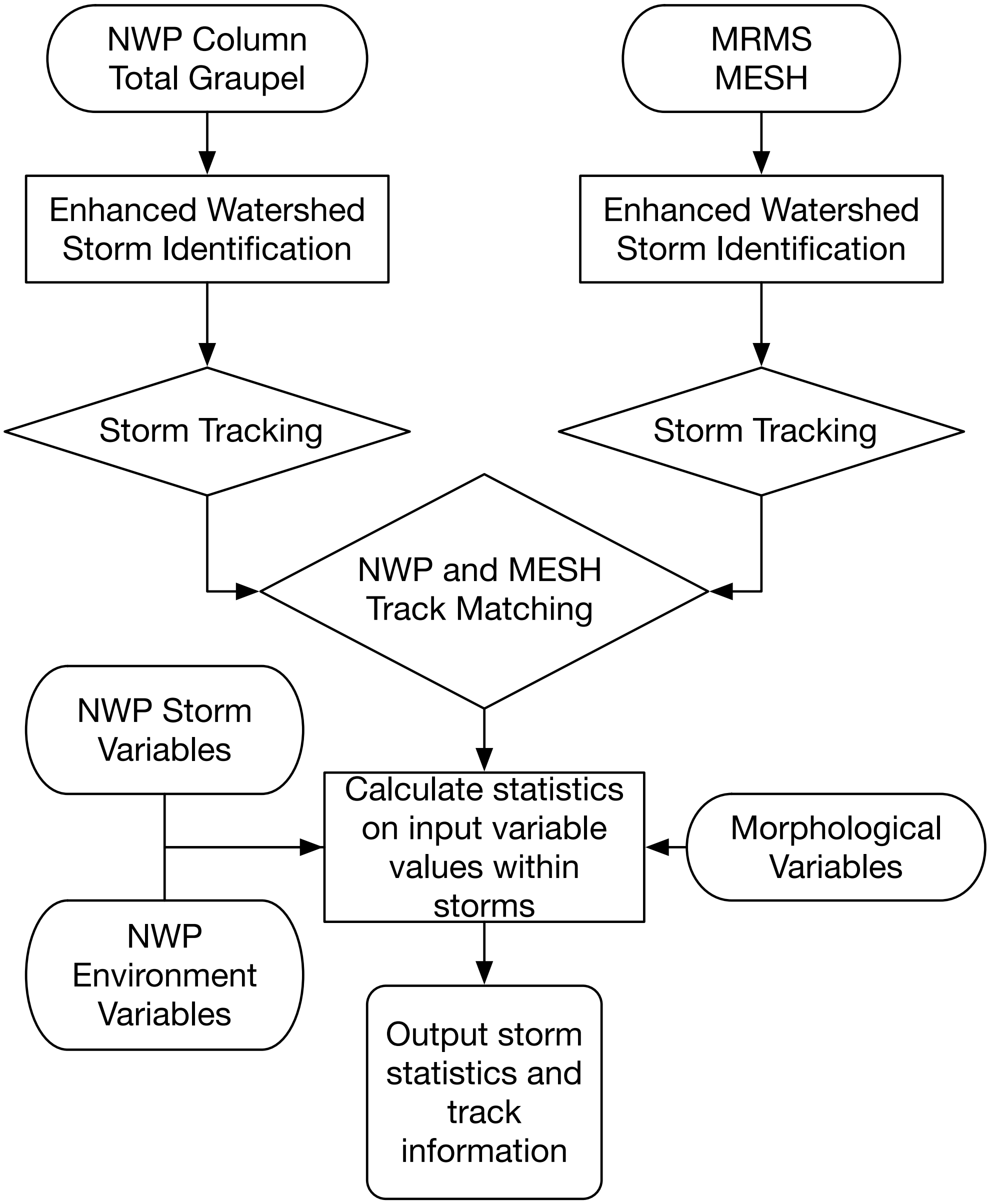

NWP Column Total Graupel
MRMS MESH
Enhanced Watershed Storm Identification
Enhanced Watershed Storm Identification
Storm Tracking
Storm Tracking
NWP and MESH Track Matching
NWP Storm Variables
Calculate statistics on input variable values within storms
Morphological Variables
NWP Environment Variables
Output storm statistics and track information

Figure 4.1: Diagram of how hail forecast and observation data are pre-processed before being used in the machine learning model. Storm tracking, track matching, and separate processing of storm and environment variables have been added to the pre-processing procedure in Fig. 3.1.

processed. The storm data processing procedure is summarized in Fig. 4.1. The enhanced watershed method (Lakshmanan et al. 2009) identifies potential hailstorm objects from the storm proxy field by identifying local maxima and then growing the objects until they meet area and intensity criteria. Hourly-maximum column-integrated graupel mass was used as the hailstorm proxy as it identifies any storm containing a significant number of graupel particles, including both ordinary thunderstorms and supercells. A minimum area of 16 grid points and a maximum area of 100 grid points was used to isolate individual storm cells while filtering storms that only lasted for a small amount of time. The minimum area was chosen to correspond to the minimum resolvable area for a single storm cell, and maximum area roughly corresponds to the area that a single storm could traverse within an hour.

Once storm objects are identified at every time step in the model run and in the observations, the objects are linked together into tracks. By grouping storms into tracks, the data processing algorithm identifies temporal trends within the storms and captures the full life cycle of a storm. A constrained version of the Hungarian method (Munkres 1957), a globally optimal matching algorithm, linked storm objects into tracks. The Hungarian method forms the basis of the TITAN storm tracking algorithm (Dixon and Wiener 1993). Unlike radar-based storm-tracking approaches, which typically receive 5-minute updates, CAMs produce hourly maximum output each hour. The hourly-maximum objects tend to be elongated along the axis of motion, so traditional Euclidean centroid distance metrics can fail in situations where individual storms are propagating parallel to each other in a line. To correct for this issue, a spatial cross-correlation motion estimation method (Lakshmanan and Smith 2010) translates the storm centroid before the matching algorithm is applied. The cross-correlation search area is constrained by the storm dimensions to minimize motion estimation error. The maximum centroid distance for tracking is 24 km, which was empirically found to connect storms appropriately while minimizing the issue of tracks jumping to adjacent storms.

Matching forecast and observed storm tracks requires a distance metric that can account for both spatial and temporal differences among tracks and can fairly compare tracks with different time durations. This track matching distance function used for this experiment is

$$D_{tm} = 0.5\frac{d_s}{160} + 0.3\frac{t_s}{3} + 0.1\frac{t_d}{16} + 0.1\frac{a_d}{200} \tag{4.1}$$

Eq. 4.1 contains a weighted combination of the distance between the starting points of each track (d_s), the difference in start times (t_s), the difference in durations (t_d), and the difference in mean areas (a_d). Tighter maximum distance constraints, which are the denominators for each term in Eq. 4.1, for start location difference (160 km) and start time difference (3 hours) were used to limit the search area, while duration (16 hours) and area (200 km^2) were used to break ties between storm objects and encourage more similar storms to be matched together. The distances have units of km, time and duration differences use hours, and area differences use km.^{2} If any of the components exceeded the maximum value, the track pairing was then excluded from consideration. Each forecast track was paired with the closest observed track that met all the distance criteria, so multiple forecast tracks could be matched with the same observed track.

4.1.3 Machine Learning Procedure

After storm tracking and matching has occurred, forecast inputs are extracted from each storm object. This approach extracts statistics about the distribution of meteorological variable values from pixels within the extent of a storm object. These statistics are the mean, standard deviation, minimum, maximum, percentiles, and skew. The full list of input variables for the CAPS ensemble is listed in Table 4.2, and the full list for the NCAR ensemble is in Table 4.3. The CAPS and NCAR ensembles use different post-processing systems and archive differing amounts of data, so some variables available in one system were not available in the other.

The machine learning hail forecast procedure is summarized in Fig. 4.2. Hail occurrence and the distribution of hail sizes are extracted from the MESH tracks and are used as the output labels for the machine learning models. If a forecast storm track is matched with a MESH track, then hail is assumed to have occurred. Forecast storm tracks with no matching MESH track are false alarms, and unmatched MESH tracks are misses. The hail size distribution within a MESH object is modeled as a gamma distribution fit to the MESH object values using maximum likelihood estimation. The gamma distribution probability density function takes the form:

$$f(x) = \frac{(x/\beta)^{\alpha-1}\exp(-x/\beta)}{\beta\Gamma(\alpha)} - x_0, \quad \alpha, \beta > 0 \tag{4.2}$$

where α is the shape parameter, x_0 is the location parameter, and β is the scale parameter. An example of a MESH object and the gamma distribution fit to its hail size distribution is shown in Fig. 4.3. The shape parameter affects the

Table 4.2: Input variables for the CAPS ensemble machine learning models. The mean, maximum, minimum, median, standard deviation, skewness, 10th percentile, and 90th percentile of the grid point values within the boundaries of each storm object were calculated for the Storm Proxy and Environment variables. CAPE is Convective Available Potential Energy, CIN is Convective Inhibition, ML is the Mean Layer, MU is the Most Unstable layer, and LCL is the Lifted Condensation Level.

Storm Proxy	**Environment**	**Morphological**
Column Total Graupel	MLCAPE	Area
2-5 km Updraft Helicity	MLCIN	Eccentricity
Reflectivity -10°C	MUCAPE	Major Axis Length
Updraft Speed	MUCIN	Minor Axis Length
Downdraft Speed	LCL Height	Orientation
Echo Top Height	0-3 km Storm Rel. Helicity	Extent
Precipitation	0-6 km Wind Shear	
Precipitable Water	500 mb Temperature	
Bunkers U	700 mb Temperature	
Bunkers V	2 m Dewpoint	
	2 m Temperature	
	850 mb Specific Humidity	
	0-3 km Lapse Rate	
	700-500 mb Lapse Rate	
	10 m U-Wind	
	10 m V-Wind	
	700 mb U-Wind	
	700 mb V-Wind	
Location		
Forecast Hour		
Valid Hour UTC		
Current Duration		
Total Duration		
W.-E. Storm Motion		
S.-N. Storm Motion		

Table 4.3: Input variables for the NCAR ensemble machine learning models. The mean, maximum, minimum, median, standard deviation, skewness, 10th percentile, and 90th percentile of the grid point values within the boundaries of each storm object were calculated for the Storm Proxy and Environment variables.

Storm Proxy	**Environment**	**Morphological**
Column Total Graupel	SBCAPE	Area
2-5 km Updraft Helicity	SBCIN	Eccentricity
Composite Reflectivity	MUCAPE	Major Axis Length
Updraft Speed	Precipitable Water	Minor Axis Length
Downdraft Speed	LCL Height	Orientation
Thompson Hail Size Max.	0-3 km S.R. Helicity	Extent
Thompson Hail Size Sfc.	0-6 km Wind Shear	
0-3 km Updraft Helicity	0-1 km Wind Shear	
2-5 km Min. Updraft Helicity		
10 m Wind Speed		
0-1 km Vorticity		
Location		
Forecast Hour		
Valid Hour UTC		
Current Duration		
Total Duration		
W.-E. Storm Motion		
S.-N. Storm Motion		

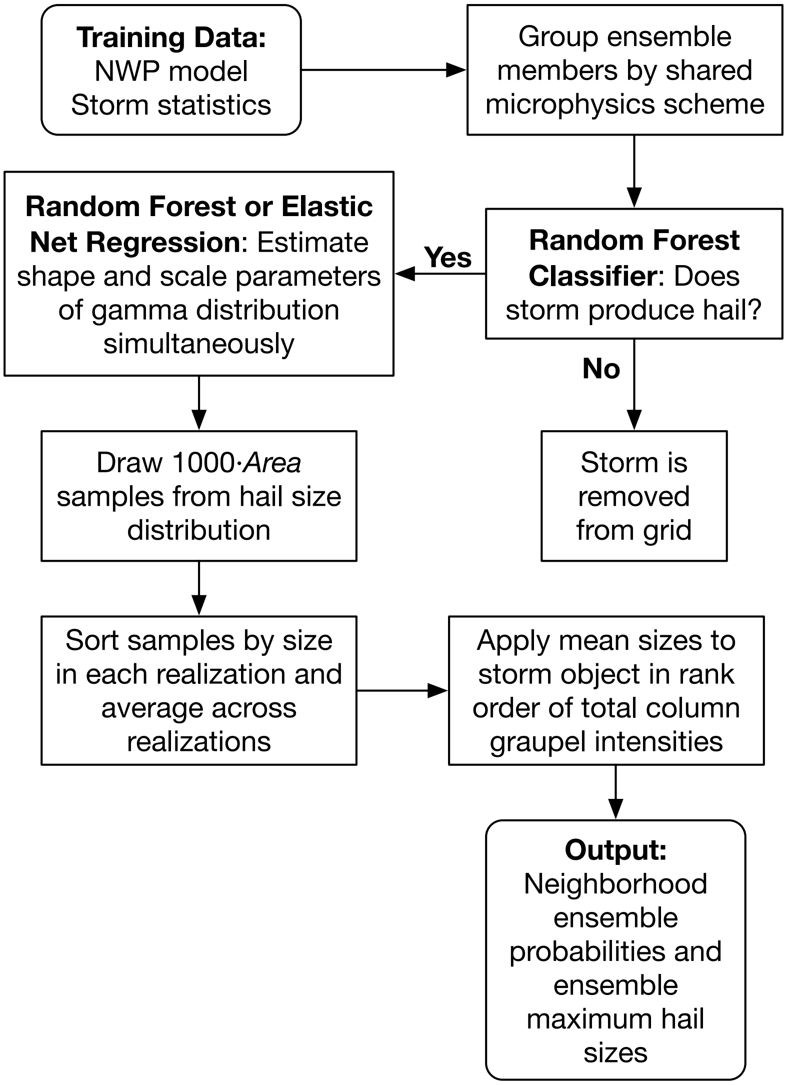

Training Data:
NWP model
Storm statistics
Group ensemble members by shared microphysics scheme
Random Forest Classifier: Does storm produce hail?
Yes
Random Forest or Elastic Net Regression: Estimate shape and scale parameters of gamma distribution simultaneously
No
Storm is removed from grid
Draw 1000·*Area* samples from hail size distribution
Sort samples by size in each realization and average across realizations
Apply mean sizes to storm object in rank order of total column graupel intensities
Output: Neighborhood ensemble probabilities and ensemble maximum hail sizes

Figure 4.2: Machine learning hail size forecasting procedure diagram.

skewness of the distribution with small shape parameter values causing more skew and larger values leading to less skew. The location parameter determines the minimum value of the distribution and was fixed at 6 mm. The scale parameter stretches or squeezes the gamma distribution and has the same units as the quantity being modeled (Wilks 2011). The MESH size distributions exhibit an inverse log-linear relationship between the shape and scale parameters.

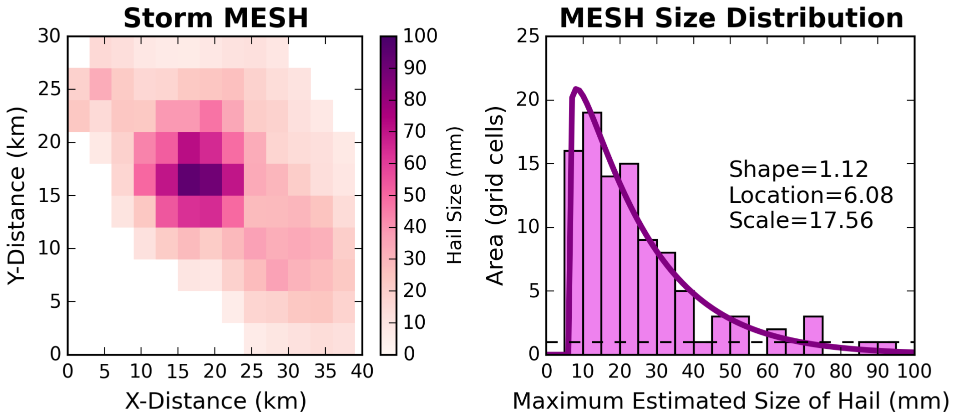


Figure 4.3: An example MESH object and the resulting discrete and parametric MESH size distributions.

I use machine learning models from the *scikit-learn* package version 0.16.1 (Pedregosa et al. 2011). A stacked model approach is used with a classifier model predicting whether or not hail occurs, and a regression model that predicts the parameters of the gamma distribution. A random forest (Breiman 2001a) classifier predicts whether hail will occur. The classifier random forest weights each training example by the inverse of its class frequency, has 500 trees, 10 minimum samples at the leaf node, and uses a grid search with cross-validation to determine the maximum number of features sampled from square root of the number of features, 20, 30, and 50 features.

Multitask learning models (Caruana 1997) predict the shape and scale parameters of the hail size distribution simultaneously. During training, the models choose weights and parameters to minimize the total error over all predicted values instead of fitting separate models for each value. Multitask learning provides additional information about the fitting process and maintains correlations among the predicted values. The hail size distribution parameters are log-transformed and normalized before fitting to capture the log-linear relationships and reduce bias in the error metric from fitting variables with differing ranges of values. Both the random forest and elastic net regression (Zou and Hastie 2005) support multitask learning and are used in this experiment. A default random forest, called "Random Forest," with 500 trees, minimum samples at the split node of 10, and sampling square root of the total number of features is used. An optimized random forest, called "Random Forest CV," with 500 trees, and cross validated grid searching of maximum features from square root, 30, 50, and 100 features, and minimum samples at a splitting node from 5, 10, and 20 samples. The elastic net determines the ratio between the ridge and lasso terms from a validation set and normalizes the values of the input features.

After the machine learning models estimate the probability of any hail and the MESH probability distribution, grid point hail size distribution estimates are performed through a sampling and sorting procedure (Fig. 4.2). For each hailstorm object with a probability of hail occurrence at least 50 %, 1000 random samples are drawn from the predicted gamma distribution for each grid point (Fig 4.2). The samples are then sorted in the dimension of the area of the object. The mean and percentiles are then calculated over the samples and are applied to the prediction grid. Convection-Allowing Model ensemble post-processing products such as neighborhood probabilities and ensemble maximum fields can then be derived from these grids and compared with other hail size and storm surrogate forecasts.

4.1.4 Evaluation

The machine learning models are trained in a way to maximize training data while preserving the independence of the evaluation data. For the CAPS ensemble, each machine learning model is trained on all ensemble members sharing the same microphysics parameterization scheme. The 2014 CAPS ensemble output were used to train the machine learning models. Because the 2015 ensemble used the P3 microphysics scheme in place of WDM6, the P3 members were evaluated using models trained on the Milbrandt and Yau members. The testing period included all runs from 12 May to 5 June 2015. Runs with missing MESH data were excluded from the evaluation.

Training and evaluation for the NCAR ensemble is performed cyclically with a new round of training performed every 2 weeks. Since each member uses the same parameterizations and is equally likely, all members are pooled into the training data for the machine learning models. The forecasting period runs from 15 May through 30 July 2015.

Storm-surrogate neighborhood ensemble probabilities (Sobash et al. 2011) are created from the machine learning models and are compared with storm-surrogate ensemble probabilities derived from thresholding the storm proxy variables. The 3 km grid for each ensemble member is subsampled into a 42 km grid, and each point on the grid is marked with a 1 if at least 1 grid point within a 42 km radius exceeds a specified intensity threshold. A Gaussian filter is then applied to the coarse grid and spreads the probability mass to surrounding grid points to reflect the estimated spatial uncertainty. Larger standard deviations for the Gaussian filter result in a greater probability of detection but also decrease the sharpness and increase the false alarm ratio. Storm-surrogate neighborhood probabilities are also calculated for HAILCAST and storm surrogate variables, including reflectivity at the -10°C level, updraft helicity and column total graupel. Updraft helicity thresholds of 75 and 150 m^2 s^{-2}, reflectivity threshold of 60 and 60 dBZ, and column total graupel thresholds of 25 and 50 kg m^{-2} were used to discriminate 25 and 50 mm diameter hail, respectively.

The statistical significance of the differences in verification scores is assessed using paired bootstrap confidence intervals and permutation tests. The bootstrap (Efron and Tibshirani 1994) is a non-parametric method for calculating the uncertainty of a statistic by repeatedly calculating that statistic on resampled replicates of the original sample. Confidence intervals for the statistic are calculated by finding the percentiles of the bootstrapped statistic distribution that correspond to the upper and lower limits of the confidence interval. Because each model performs forecasts for the same events, the event identities are resampled the same way for each forecast method. Once the verification statistics of interest are calculated on each bootstrap sample, the hail forecasting methods are ranked by their score, and the frequency of each rank for each model is calculated. By counting the frequency of each rank, the consistency of the model performance relative to another model can be quantified, even if the differences in scores are small. If there is no overlap in the rankings, then the difference in rankings can be considered statistically significant. For the hail forecast statistics, 1000 bootstrap samples are used to capture the required amount of precision for a 95% confidence interval in a reasonable amount of computational time.

If there is overlap, then a permutation test needs to be performed between the two overlapping models to determine if the p-value of the difference between their scores is statistically significant. In a permutation test, the difference in a statistic is calculated between two paired samples, such as forecasts from two models. Based on the exchangeability principle (Wilks 2011), a null hypothesis distribution of this difference is then created by randomly switching some of the forecast values between the two models and then recalculating the difference in statistics 1000 times. The percentile of the original difference relative to the null hypothesis distribution is the p-value. If the p-value is less than the chosen significance threshold of 5 %, then the difference in models is statistically significant. Because performing multiple comparisons increases the chance that some of the null hypotheses are falsely rejected, the significance threshold is adjusted to retain a group false discovery rate of 5 % using the p-value ranking approach (Benjamini and Hochberg 1995; Wilks 2006).

4.2 Results

4.2.1 Hail Object Forecast Evaluation

Probabilistic forecasts of hail occurrence were evaluated for each member of the CAPS and NCAR ensembles. This evaluation determines how well the random forest discriminates between forecast storms matched with MESH tracks and those that are not. Fig. 4.4 shows the ROC curve and reliability diagram for the CAPS ensemble members. The ensemble members cluster by microphysics scheme as Thompson members have lower AUCs than other members, and P3 members have higher probabilities of detection at low probability thresholds. On the reliability diagram, all members display a consistent under forecasting bias but produce sharp probability forecasts. At low forecast probabilities, the Thompson members have a higher observed relative frequency than other models, resulting in a lower Brier Skill Score. In the NCAR Ensemble (Fig. 4.5), the random forest shows good discrimination skill with AUCs of 0.70. The

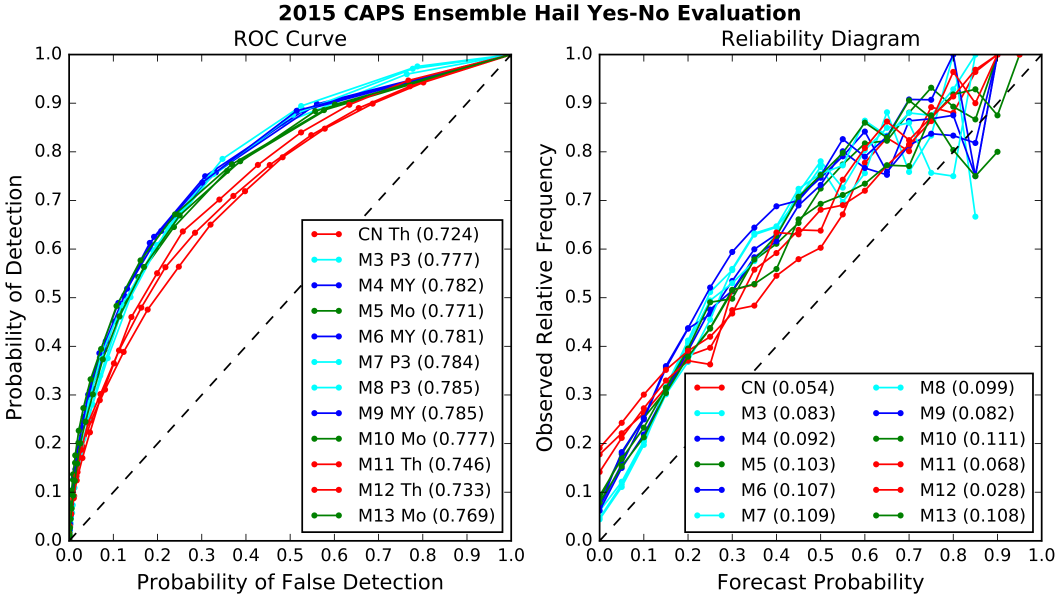


Figure 4.4: ROC Curve and reliability diagram of random forest hail occurrence forecasts from each member of the 2015 CAPS Ensemble. Members are colored by microphysics scheme. The AUC for each member is indicated in the ROC Curve legend, and the Brier Skill Score for each member is in the reliability diagram legend.

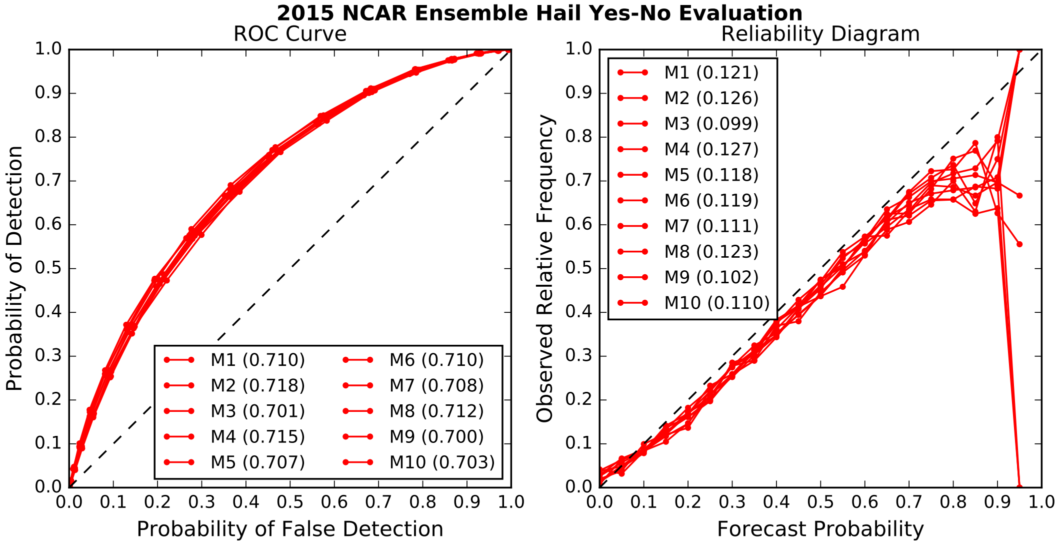


Figure 4.5: ROC Curve and reliability diagram of random forest hail occurrence forecasts from each member of the 2015 NCAR Ensemble. The AUC for each member is indicated in the ROC Curve legend, and the Brier Skill Score for each member is in the reliability diagram legend.

reliability diagram shows that the probability forecasts are reliable but slightly overconfident.

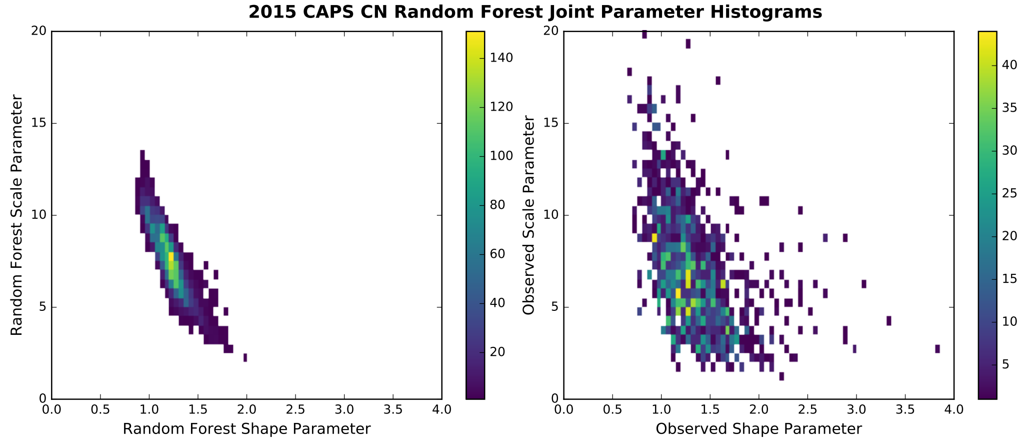


Figure 4.6: Joint distribution of the shape and scale parameters for the CAPS ensemble control member random forest forecast and observations. The other CAPS ensemble members exhibited similar relationships.

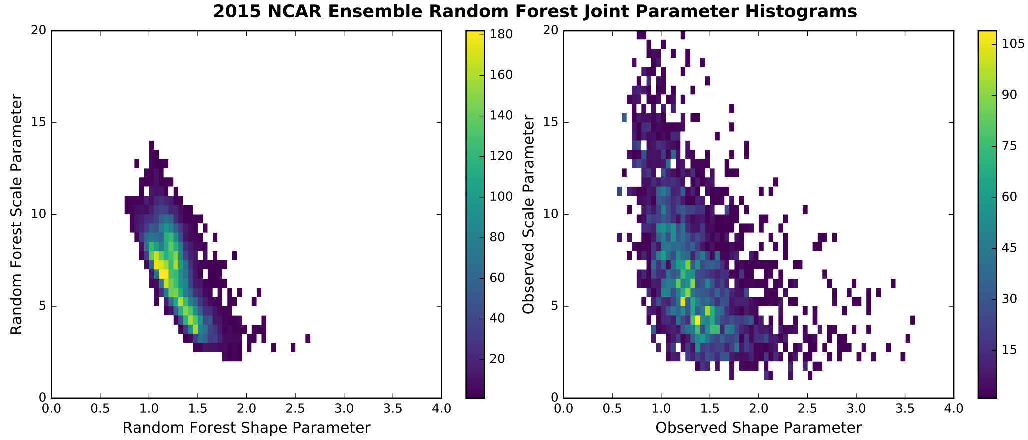


Figure 4.7: Joint distribution of the shape and scale parameters for the CAPS ensemble random forest forecast and observations.

The joint distributions of the shape and scale parameters for the CAPS ensemble control member and the NCAR ensemble are shown in Fig. 4.6 and Fig. 4.7, respectively. The CAPS and NCAR ensemble joint distributions capture the roughly log-linear relationship between the observed shape and scale parameters. In small hail cases with a small scale parameter, the distribution of hail sizes will exhibit a distribution closer to Gaussian, but for large hail cases with a large scale parameter, only a small area will contain large hail, so the distribution tends to be closer to exponential and have a smaller shape parameter. The random forest is able to capture this relationship due to multitask learning preserving the pairing and the log transform of the labels enabling a linear combination of trees to maintain the log-linear relationship. Neither random forest fully captures the sharpness of the observed distributions, but the shape is closely replicated.

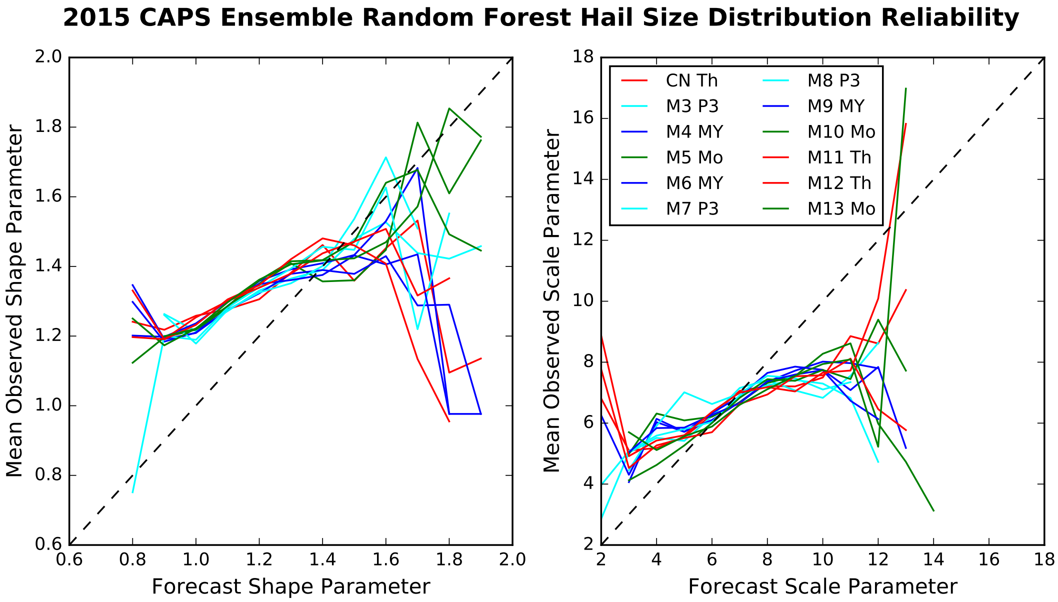


Figure 4.8: Reliability diagrams showing the CAPS Ensemble mean random forest observed hail size distribution parameters for each forecast parameter value.

Reliability diagrams for the shape and scale parameters in the CAPS and NCAR ensembles show how closely the forecast values correspond to observed values on average (Fig. 4.8 and Fig. 4.9). The CAPS ensemble exhibits poor

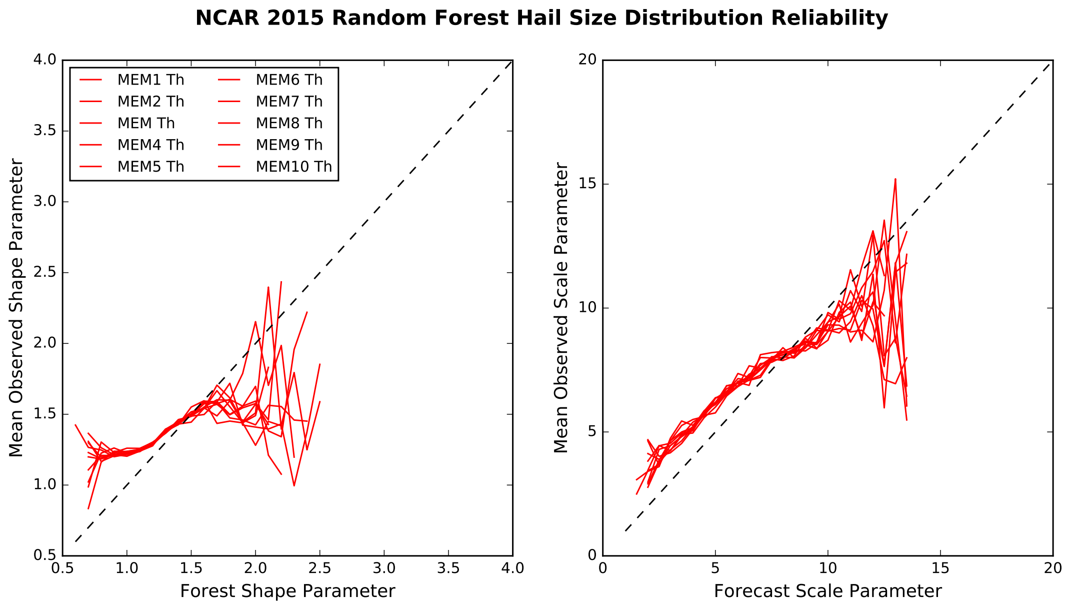
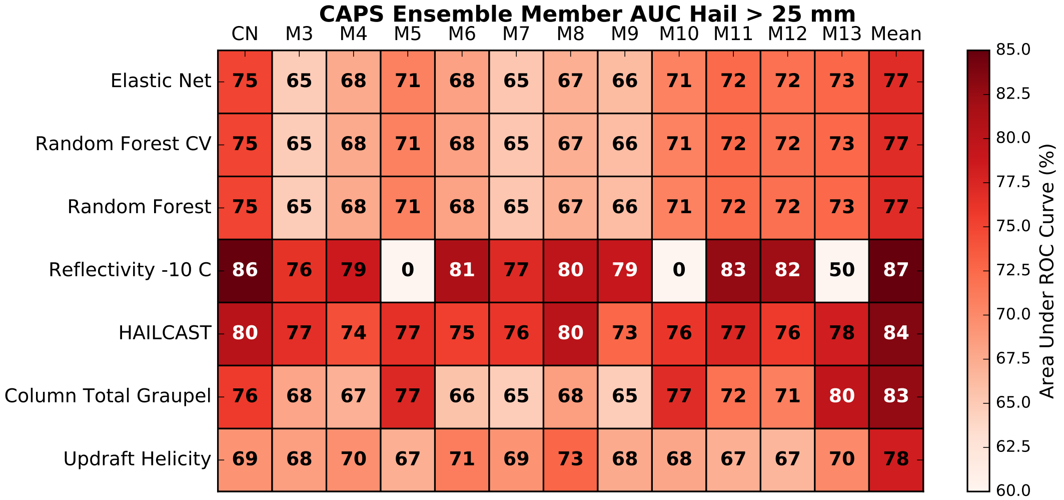


Figure 4.9: Reliability diagrams showing the NCAR ensemble mean random forest observed hail size distribution parameters for each forecast parameter value.

calibration with an underforecasting bias for small values of the shape and scale parameter and an overforecasting bias for large values. All of the ensemble members show similar issues, and there does not appear to be any clustering by microphysics scheme. The NCAR ensemble reliability curves exhibit better calibration than the CAPS ensemble but still show biases at extreme parameter values.

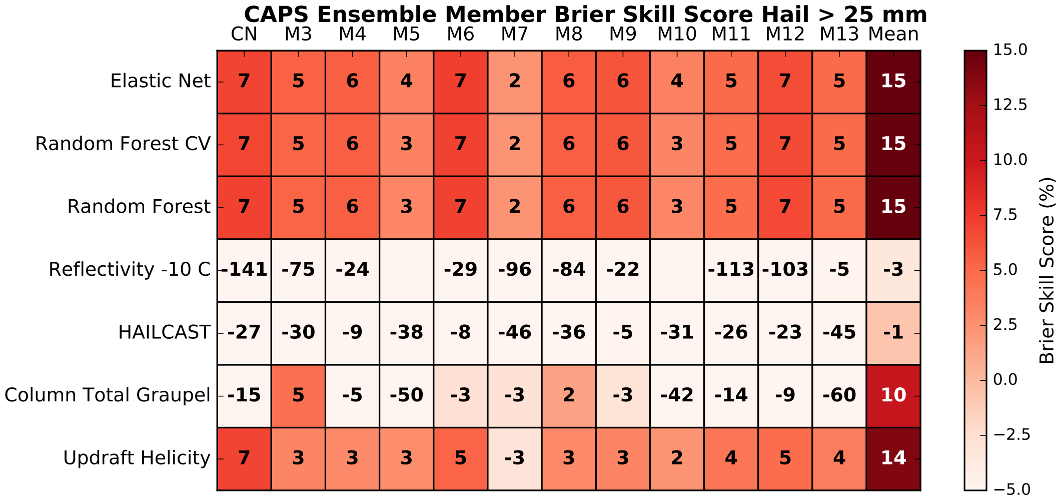
4.2.2 Full Period Forecast Evaluation

Storm surrogate probabilities were calculated over the 24-hour period from 12 UTC to 12 UTC for the machine learning models and raw storm surrogate variables. These probabilities are analogous to Storm Prediction Center Convective Outlooks and allow for a more direct comparison between the two types



	CN	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	Mean
Elastic Net	75	65	68	71	68	65	67	66	71	72	72	73	77
Random Forest CV	75	65	68	71	68	65	67	66	71	72	72	73	77
Random Forest	75	65	68	71	68	65	67	66	71	72	72	73	77
Reflectivity -10 C	86	76	79	0	81	77	80	79	0	83	82	50	87
HAILCAST	80	77	74	77	75	76	80	73	76	77	76	78	84
Column Total Graupel	76	68	67	77	66	65	68	65	77	72	71	80	83
Updraft Helicity	69	68	70	67	71	69	73	68	68	67	67	70	78

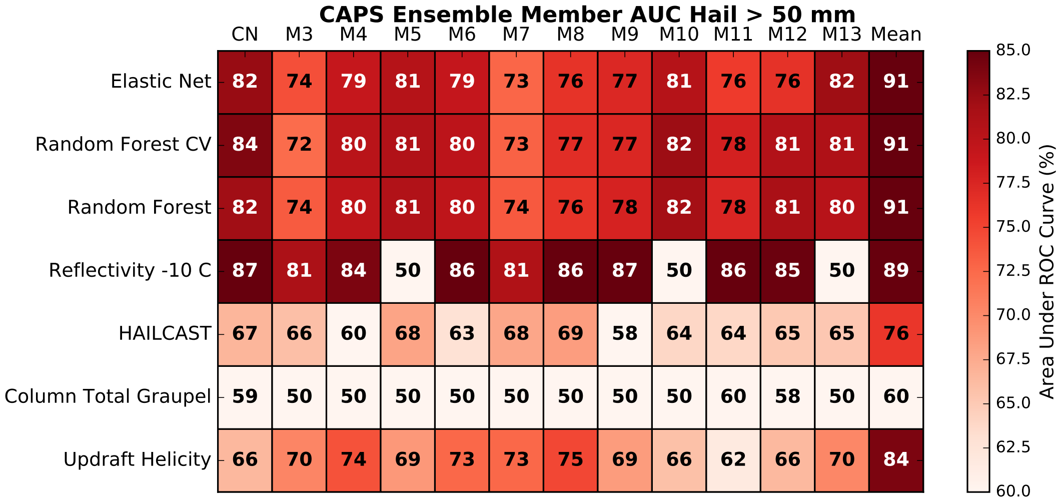
Figure 4.10: Area Under the ROC Curve (AUC) for storm surrogate probabilities of 25 mm hail for each member of the 2015 CAPS ensemble and the ensemble mean. The values have been multiplied by 100 to improve readability.



	CN	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	Mean
Elastic Net	7	5	6	4	7	2	6	6	4	5	7	5	15
Random Forest CV	7	5	6	3	7	2	6	6	3	5	7	5	15
Random Forest	7	5	6	3	7	2	6	6	3	5	7	5	15
Reflectivity -10 C	-141	-75	-24		-29	-96	-84	-22		-113	-103	-5	-3
HAILCAST	-27	-30	-9	-38	-8	-46	-36	-5	-31	-26	-23	-45	-1
Column Total Graupel	-15	5	-5	-50	-3	-3	2	-3	-42	-14	-9	-60	10
Updraft Helicity	7	3	3	3	5	-3	3	3	2	4	5	4	14

Figure 4.11: Brier Skill Score for storm surrogate probabilities of 25 mm hail for each member of the 2015 CAPS ensemble and the ensemble mean. Blank spots indicate that none of the forecasts from that member exceeded the intensity threshold. The values have been multiplied by 100 to improve readability.

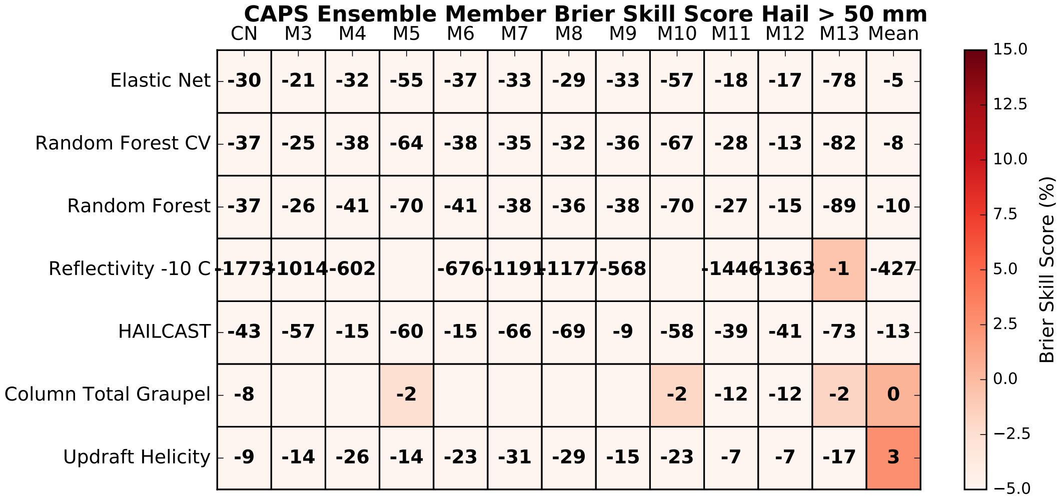
of forecast methods. The probabilities are first calculated for each member and then averaged. The CAPS ensemble members show a lot of variability in both AUC (Fig. 4.10) and Brier Skill Score (BSS; Fig. 4.11). The machine learning methods have identical AUCs and all perform worse than the storm-surrogate methods, but the opposite is true for BSS. The variables with the best discrimination are the least reliable and vice versa. The ensemble mean outperforms all individual members.



	CN	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	Mean
Elastic Net	82	74	79	81	79	73	76	77	81	76	76	82	91
Random Forest CV	84	72	80	81	80	73	77	77	82	78	81	81	91
Random Forest	82	74	80	81	80	74	76	78	82	78	81	80	91
Reflectivity -10 C	87	81	84	50	86	81	86	87	50	86	85	50	89
HAILCAST	67	66	60	68	63	68	69	58	64	64	65	65	76
Column Total Graupel	59	50	50	50	50	50	50	50	50	60	58	50	60
Updraft Helicity	66	70	74	69	73	73	75	69	66	62	66	70	84

Figure 4.12: AUC for storm surrogate probabilities of 50 mm hail for each member of the 2015 CAPS ensemble and the ensemble mean. The values have been multiplied by 100 to improve readability.

At the 50 mm threshold, the machine learning methods exhibit more variability among each other and among members (Fig. 4.12). The machine learning methods consistently maintain higher AUC values than the storm surrogates with the exception of Reflectivity at the -10C level. All of the members and methods except ensemble mean updraft helicity have negative BSS (Fig. 4.13). The machine learning methods achieve higher BSSs than the other storm surrogate variables. Some of the ensemble members have members whose radar



	CN	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	Mean
Elastic Net	-30	-21	-32	-55	-37	-33	-29	-33	-57	-18	-17	-78	-5
Random Forest CV	-37	-25	-38	-64	-38	-35	-32	-36	-67	-28	-13	-82	-8
Random Forest	-37	-26	-41	-70	-41	-38	-36	-38	-70	-27	-15	-89	-10
Reflectivity -10 C	-1773	-1014	-602		-676	-1191	-1177	-568		-1446	-1363	-1	-427
HAILCAST	-43	-57	-15	-60	-15	-66	-69	-9	-58	-39	-41	-73	-13
Column Total Graupel	-8			-2					-2	-12	-12	-2	0
Updraft Helicity	-9	-14	-26	-14	-23	-31	-29	-15	-23	-7	-7	-17	3

Figure 4.13: Brier Skill Score for storm surrogate probabilities of 50 mm hail for each member of the 2015 CAPS ensemble and the ensemble mean. Blank spots indicate that none of the forecasts from that member exceeded the intensity threshold. The values have been multiplied by 100 to improve readability.

reflectivity does not exceed 60 dBZ. Both of these members use the Morrison microphysics scheme, which tends to produce more widespread, less intense convection than other microphysics parameterizations. These microphysics differences also result in lower BSS for column total graupel.

The ensemble mean ROC curves at the 25 mm (Fig. 4.14) and 50 mm (Fig. 4.14) show that the biggest differences among approaches come with the minimum probability threshold and what level of probability of detection it provides. At 25 mm, the ROC curves all follow the same general trajectory but have different starting points. Reflectivity, HAILCAST, and column total graupel all have higher probabilities of detection (POD) but also higher probabilities of false detection (POFD). The machine learning models have nearly identical values. At 50 mm, radar reflectivity has the largest ROC area but also has much higher POFD than the other methods (Fig. 4.15). The machine learning methods and updraft helicity have lower POFD and higher POD than HAILCAST and column total graupel.

The performance diagram highlights the impact of false alarms more strongly than the ROC curve by using False Alarm Ratio (FAR) instead of POFD on the x-axis. Because the number of true negatives far outweighs the number of false alarms, the POFD tends to be low and less sensitive to false alarms. FAR, on the other hand, is the ratio of false alarms to total forecasts, so it is much more sensitive to changes in the number of false alarms. For the 25 mm threshold, the machine learning methods all have lower FAR than every other model (Fig. 4.16). The FAR for updraft helicity is nearly as low as the machine learning models. For 50 mm, all models have a much higher FAR, but their relative rankings stay the same (Fig. 4.17). At higher probability thresholds, the machine learning models have much lower FAR compared with the other methods but are only able to detect a small percentage of events.

The attributes diagrams indicate the reliability and sharpness of each method. The machine learning methods and updraft helicity are all reliable at the 25 mm threshold (Fig. 4.18) while the other storm surrogate methods are all overconfident. The overconfidence improves their AUC because they detect more events at a given probability threshold but hurts their reliability. All of the methods are sharp but HAILCAST has the largest number of forecasts at high probabilities. Every method is overconfident and less sharp at the 50 mm threshold (Fig. 4.19). The machine learning methods have the best combination of reliability and sharpness while reflectivity is extremely overconfident and column graupel is not very sharp. HAILCAST is more overconfident than the machine learning models but still produces sharp forecasts.

The sensitivity of the verification scores to sample variability is assessed by a bootstrap resampling of the forecast-observation pairs and permutation tests for models with adjacent ranks. With the lack of overlap in rankings, the storm surrogate methods have AUC values that are significantly higher at the 5 % level at 25 mm (Fig. 4.22). The machine learning methods show high overlap in their rankings and have the same AUC (Fig. 4.22), which is corroborated by the permutation test p-values (Fig. 4.22). The machine learning methods all significantly outperform every method at the 50 mm threshold (Fig. 4.22) but are statistically indistinguishable (Fig. 4.22). The machine learning methods had the statistically significant highest BSS at 25 mm (Fig. 4.23). The 50 mm BSS rankings showed small amounts of overlap between adjacent models in the rankings (Fig. 4.23). Updraft Helicity and Column Total Graupel were more reliable than any of the machine learning methods.

The spatial distribution of probabilistic hail forecasts from the CAPS ensemble is shown in Fig. 4.24. The 10 % probability threshold was chosen to show the maximum spatial extent of the forecasts at a threshold commonly used by forecasters to assess hail risk. The Random Forest and Updraft Helicity methods capture the two observed maxima in 25 mm hail frequency in the Texas Panhandle and northeast Colorado. The Random Forest underestimates the eastward extent of hail frequency while Updraft Helicity captures the full area better but with a slight eastward bias. The frequency maxima for

HAILCAST and Column Total Graupel are displaced far eastward into western Missouri and eastern Texas. HAILCAST and Column Total Graupel also have large numbers of hail forecasts along the Gulf coast whereas the Random Forest and Updraft Helicity have none in that area. These biases are more apparent in the maps of false positives (Fig. 4.25) and false negatives (Fig. 4.26). The Random Forest and Updraft Helicity false positives are concentrated in the areas near the observed hail maxima, and the misses are primarily found along the Gulf Coast. The HAILCAST and Column Total Graupel false positives are found more along the Gulf coast and in Missouri while significant numbers of false negatives can be found along the eastern edge of the Rocky Mountains. All of the methods have false negatives in northeast New Mexico and central Wyoming, which may be a result of underlying model bias for a particular case or subset of cases. The Random Forest does the best at capturing the extent of the 50 mm reports (Fig. 4.24) while the other methods exhibit eastward biases in their maxima.

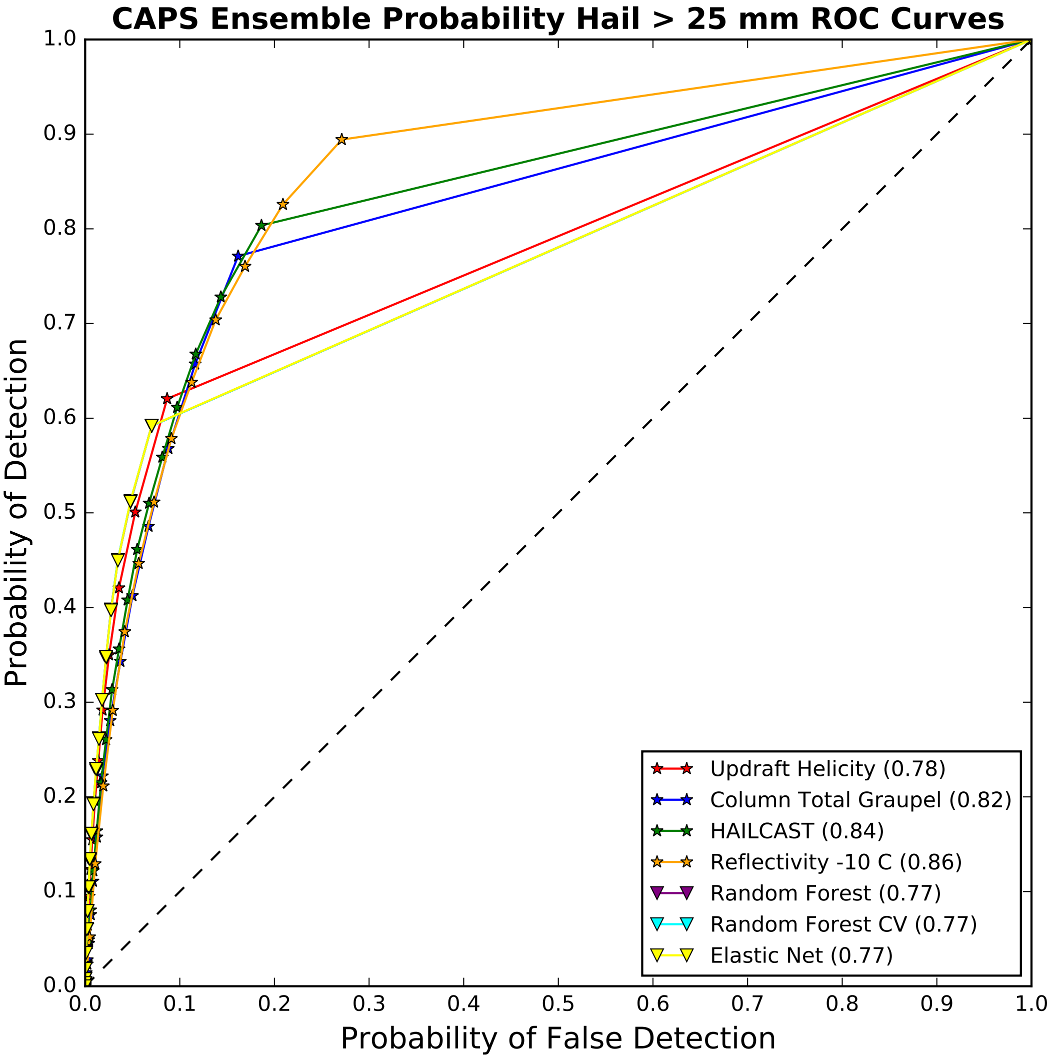


Figure 4.14: ROC curves for each CAPS ensemble mean storm-surrogate probability of 25 mm hail. The AUC for each curve is in the legend.

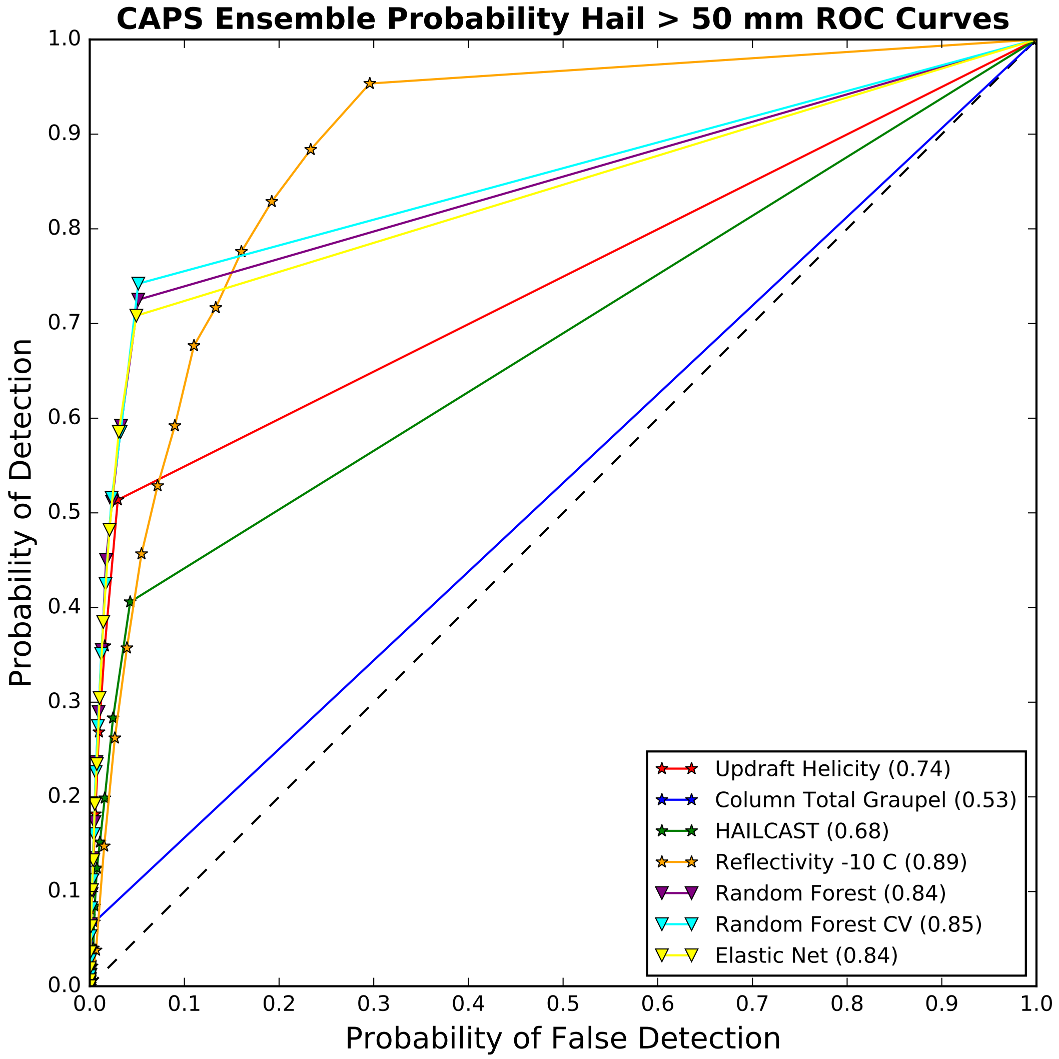


Figure 4.15: ROC curves for each CAPS ensemble mean storm-surrogate probability of 50 mm hail. The AUC for each curve is in the legend.

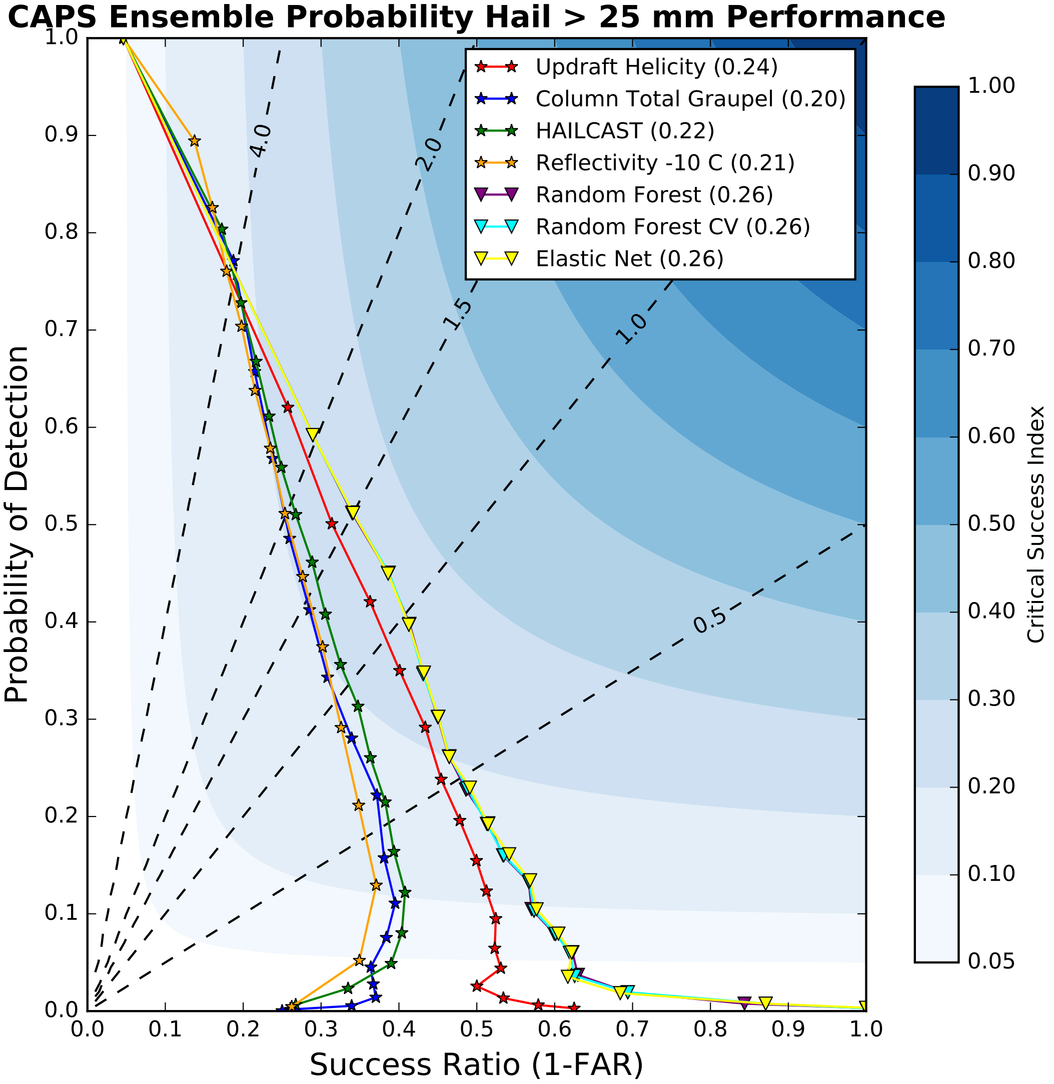


Figure 4.16: Performance curves for each CAPS ensemble mean storm-surrogate probability of 25 mm hail. The maximum CSI for each curve is in the legend.

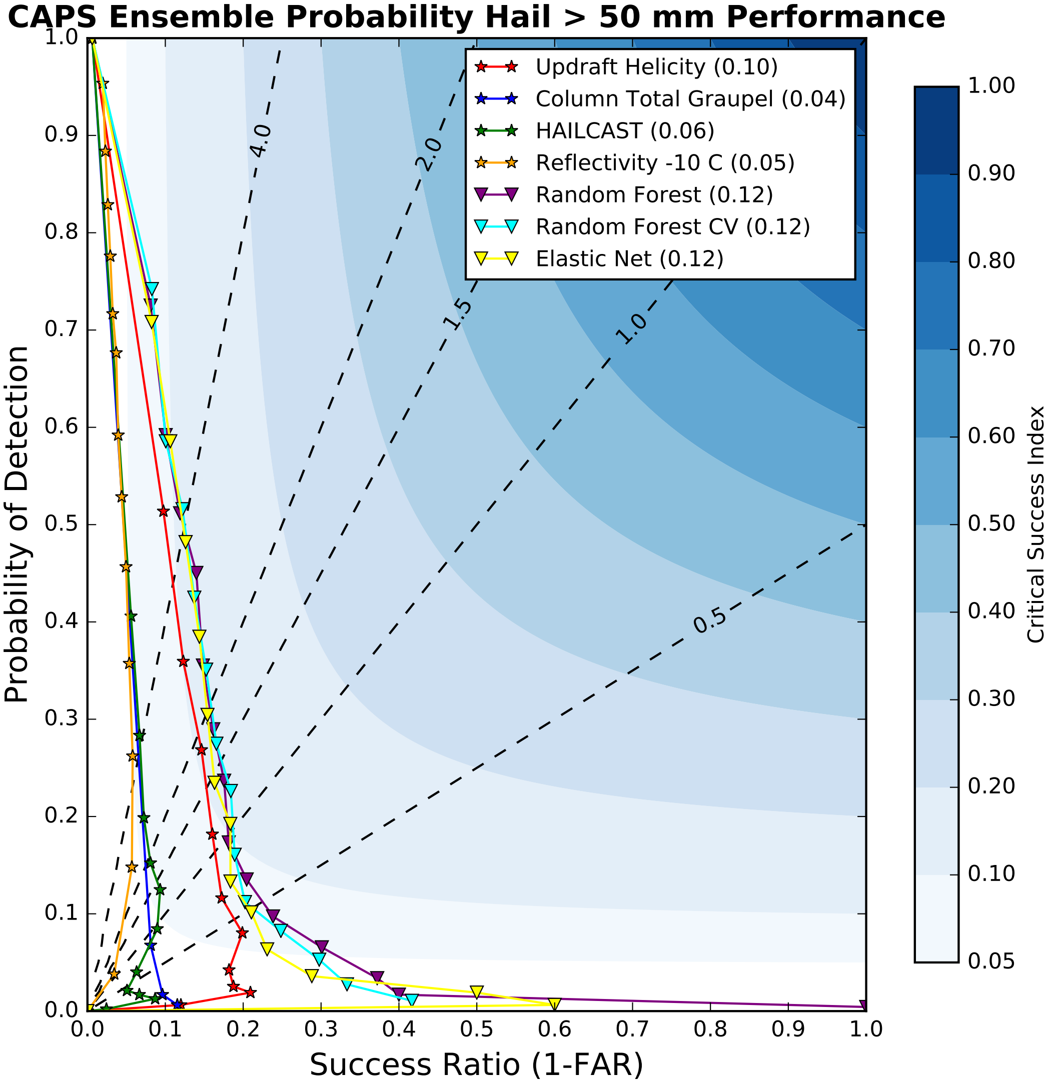


Figure 4.17: Performance curves for each CAPS ensemble mean storm-surrogate probability of 50 mm hail. The maximum CSI for each curve is in the legend.

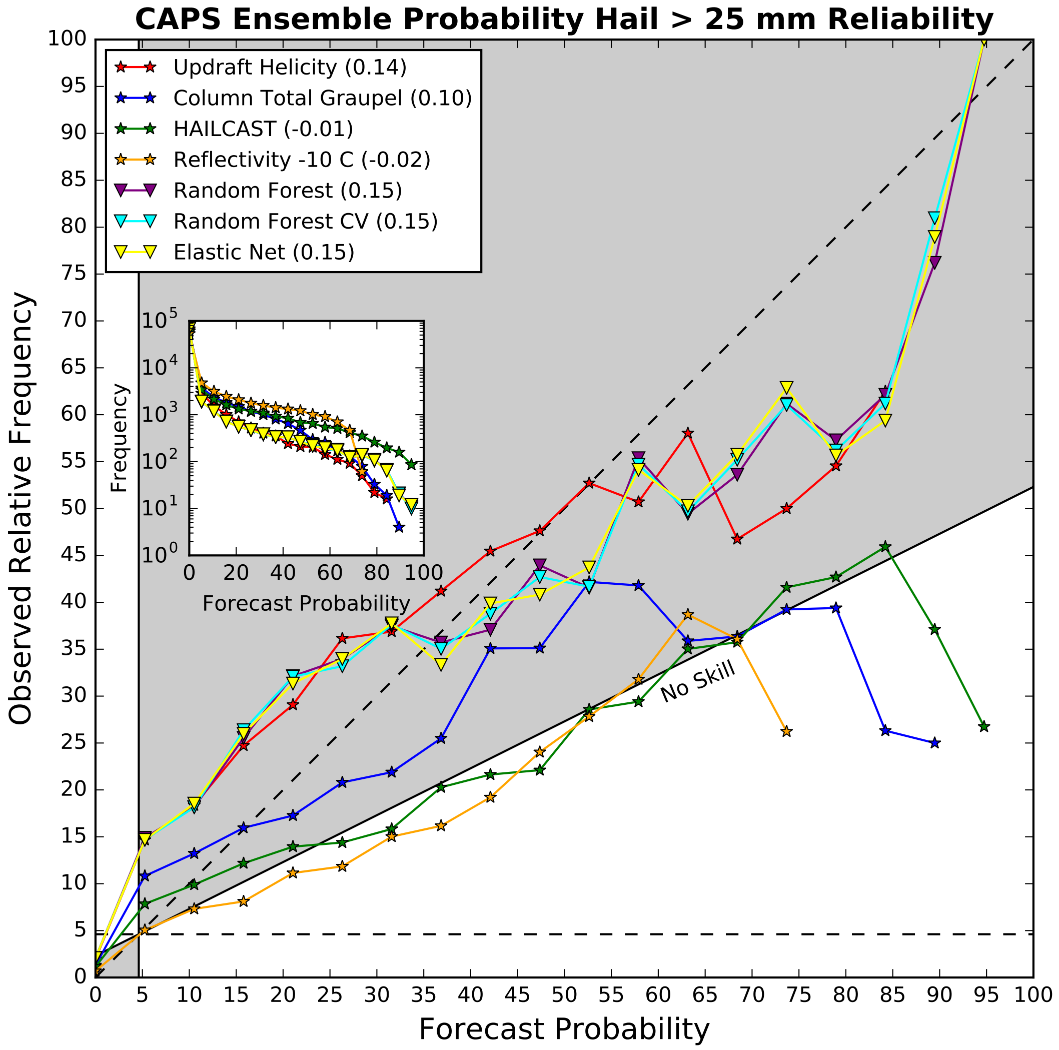


Figure 4.18: Reliability curves for each CAPS ensemble mean storm-surrogate probability of 25 mm hail. The BSS for each curve is in the legend.

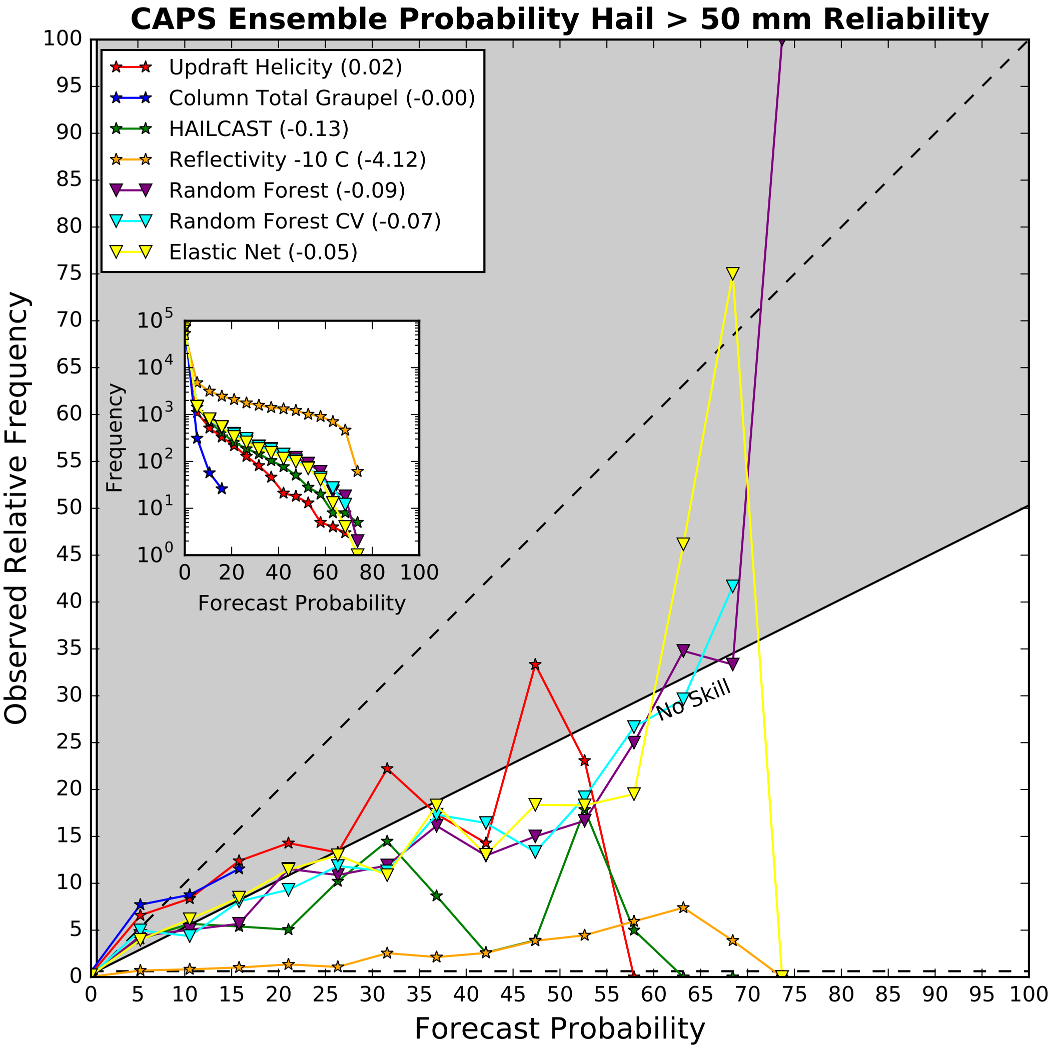


Figure 4.19: Reliability curves for each CAPS ensemble mean storm-surrogate probability of 50 mm hail. The BSS for each curve is in the legend.

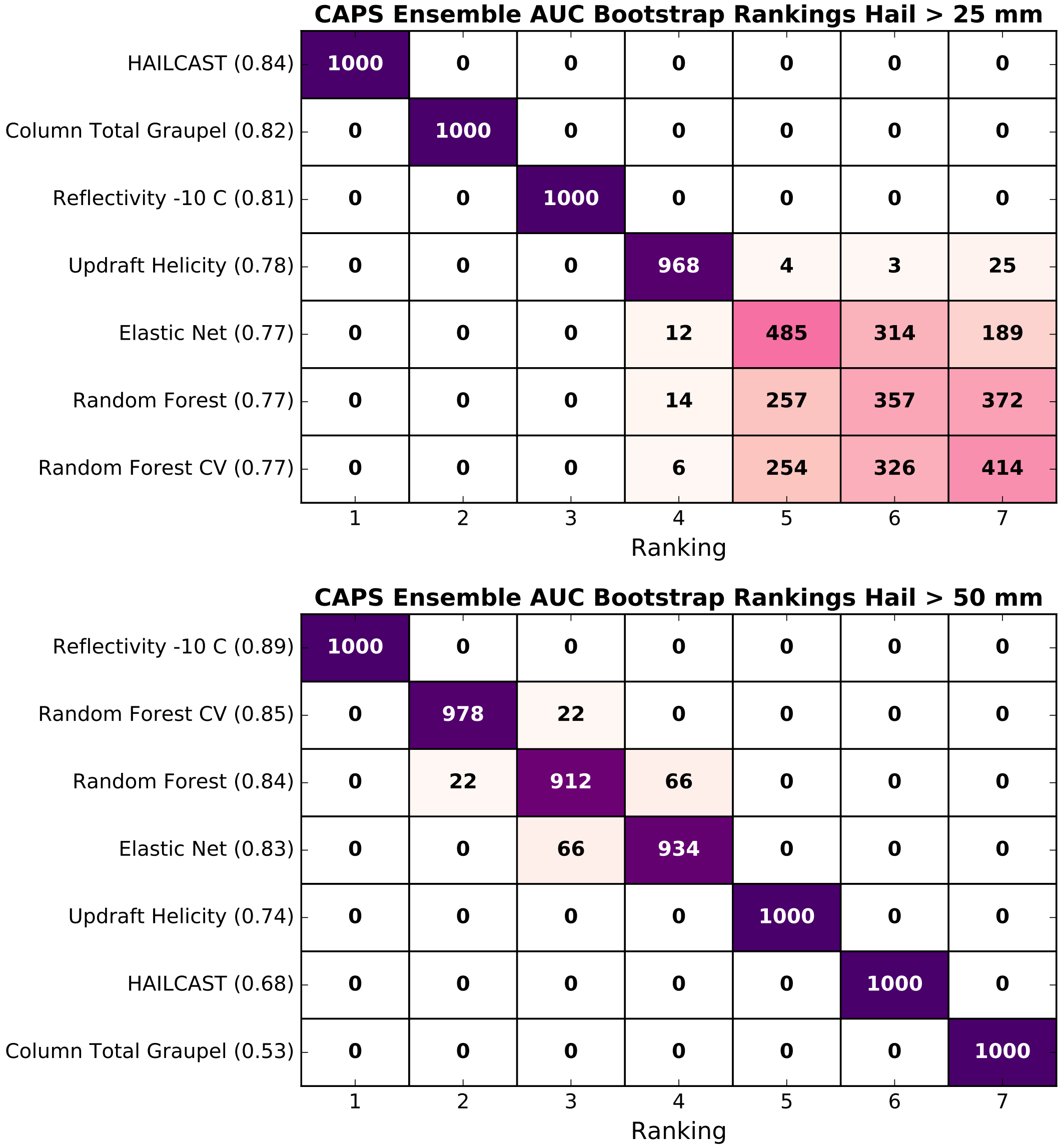
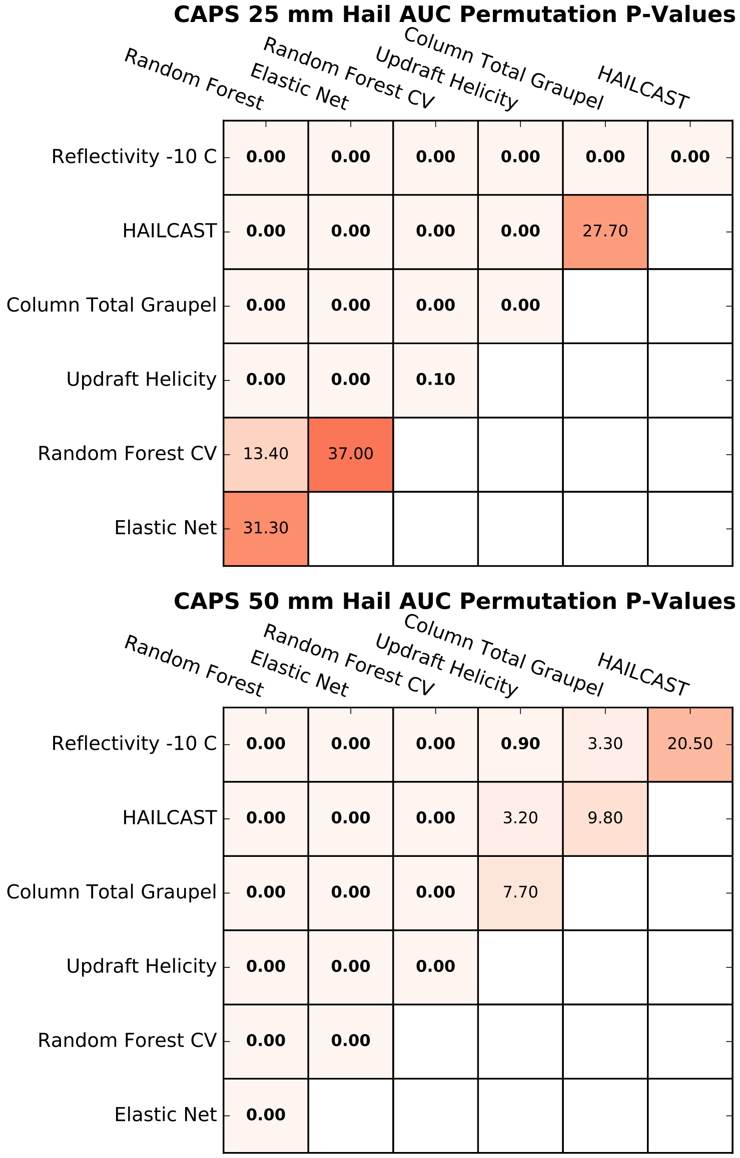


Figure 4.20: Frequency of CAPS Ensemble AUC rankings in a paired bootstrap comparison of the different methods for probability of 25 and 50 mm hail. The mean AUC for each method is shown in parentheses.



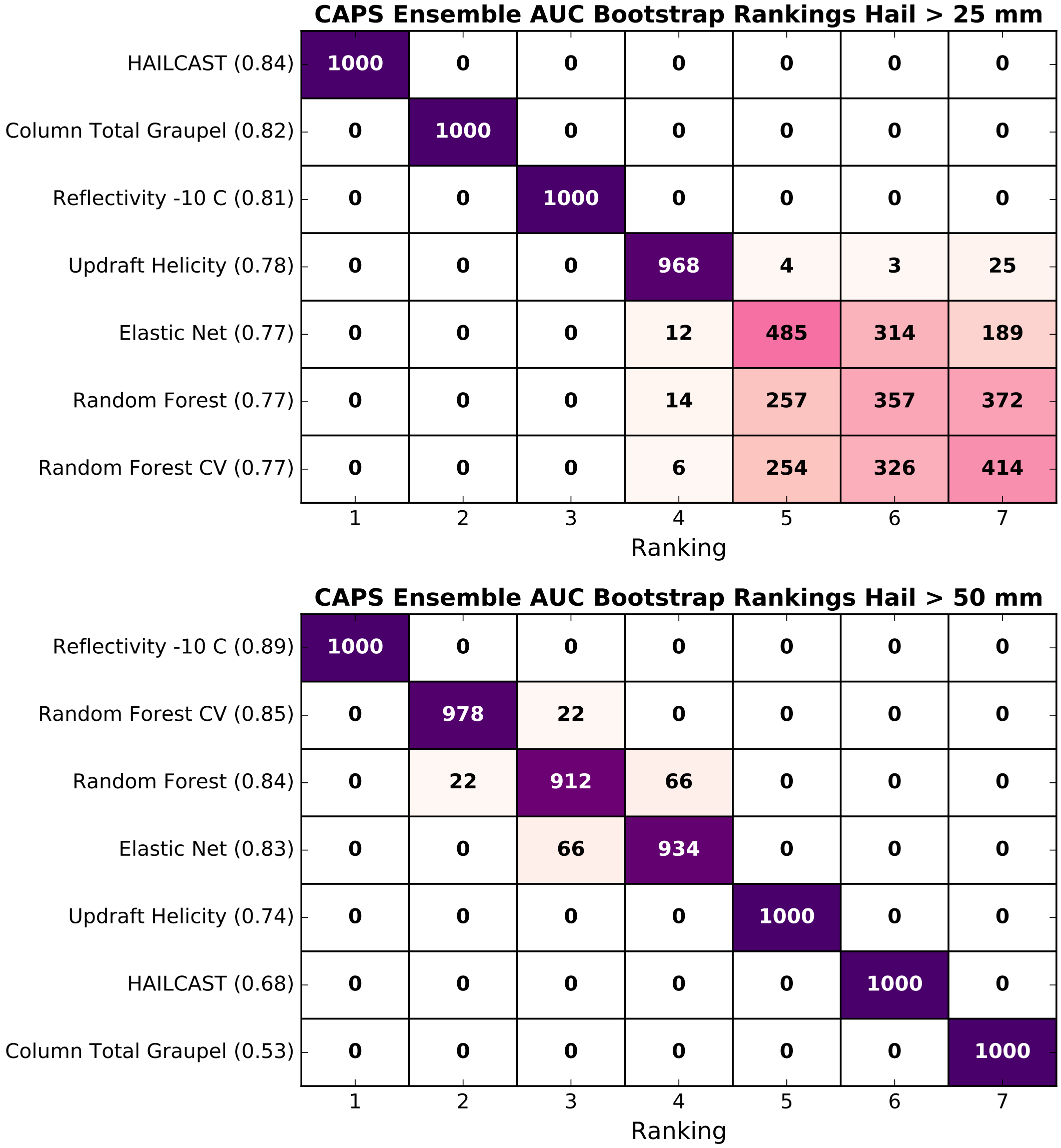
CAPS 25 mm Hail AUC Permutation P-Values

	Random Forest	Elastic Net	Random Forest CV	Updraft Helicity	Column Total Graupel	HAILCAST
Reflectivity -10 C	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
HAILCAST	**0.00**	**0.00**	**0.00**	**0.00**	27.70	
Column Total Graupel	**0.00**	**0.00**	**0.00**	**0.00**		
Updraft Helicity	**0.00**	**0.00**	**0.10**			
Random Forest CV	13.40	37.00				
Elastic Net	31.30					

CAPS 50 mm Hail AUC Permutation P-Values

	Random Forest	Elastic Net	Random Forest CV	Updraft Helicity	Column Total Graupel	HAILCAST
Reflectivity -10 C	**0.00**	**0.00**	**0.00**	**0.90**	3.30	20.50
HAILCAST	**0.00**	**0.00**	**0.00**	3.20	9.80	
Column Total Graupel	**0.00**	**0.00**	**0.00**	7.70		
Updraft Helicity	**0.00**	**0.00**	**0.00**			
Random Forest CV	**0.00**	**0.00**				
Elastic Net	**0.00**					

Figure 4.21: The p-values (multiplied by 100) from permutation tests for the difference in AUC between hail forecast models at 25 and 50 mm thresholds. P-values in bold are statistically significant based on the false discovery method with a rate of 0.05 (α=0.03). Darker reds are associated with larger p-values.



CAPS Ensemble AUC Bootstrap Rankings Hail > 25 mm

	1	2	3	4	5	6	7
HAILCAST (0.84)	1000	0	0	0	0	0	0
Column Total Graupel (0.82)	0	1000	0	0	0	0	0
Reflectivity -10 C (0.81)	0	0	1000	0	0	0	0
Updraft Helicity (0.78)	0	0	0	968	4	3	25
Elastic Net (0.77)	0	0	0	12	485	314	189
Random Forest (0.77)	0	0	0	14	257	357	372
Random Forest CV (0.77)	0	0	0	6	254	326	414

Ranking

CAPS Ensemble AUC Bootstrap Rankings Hail > 50 mm

	1	2	3	4	5	6	7
Reflectivity -10 C (0.89)	1000	0	0	0	0	0	0
Random Forest CV (0.85)	0	978	22	0	0	0	0
Random Forest (0.84)	0	22	912	66	0	0	0
Elastic Net (0.83)	0	0	66	934	0	0	0
Updraft Helicity (0.74)	0	0	0	0	1000	0	0
HAILCAST (0.68)	0	0	0	0	0	1000	0
Column Total Graupel (0.53)	0	0	0	0	0	0	1000

Ranking

Figure 4.22: Frequency of CAPS Ensemble AUC rankings in a paired bootstrap comparison of the different methods for probability of 25 and 50 mm hail. The mean AUC for each method is shown in parentheses.

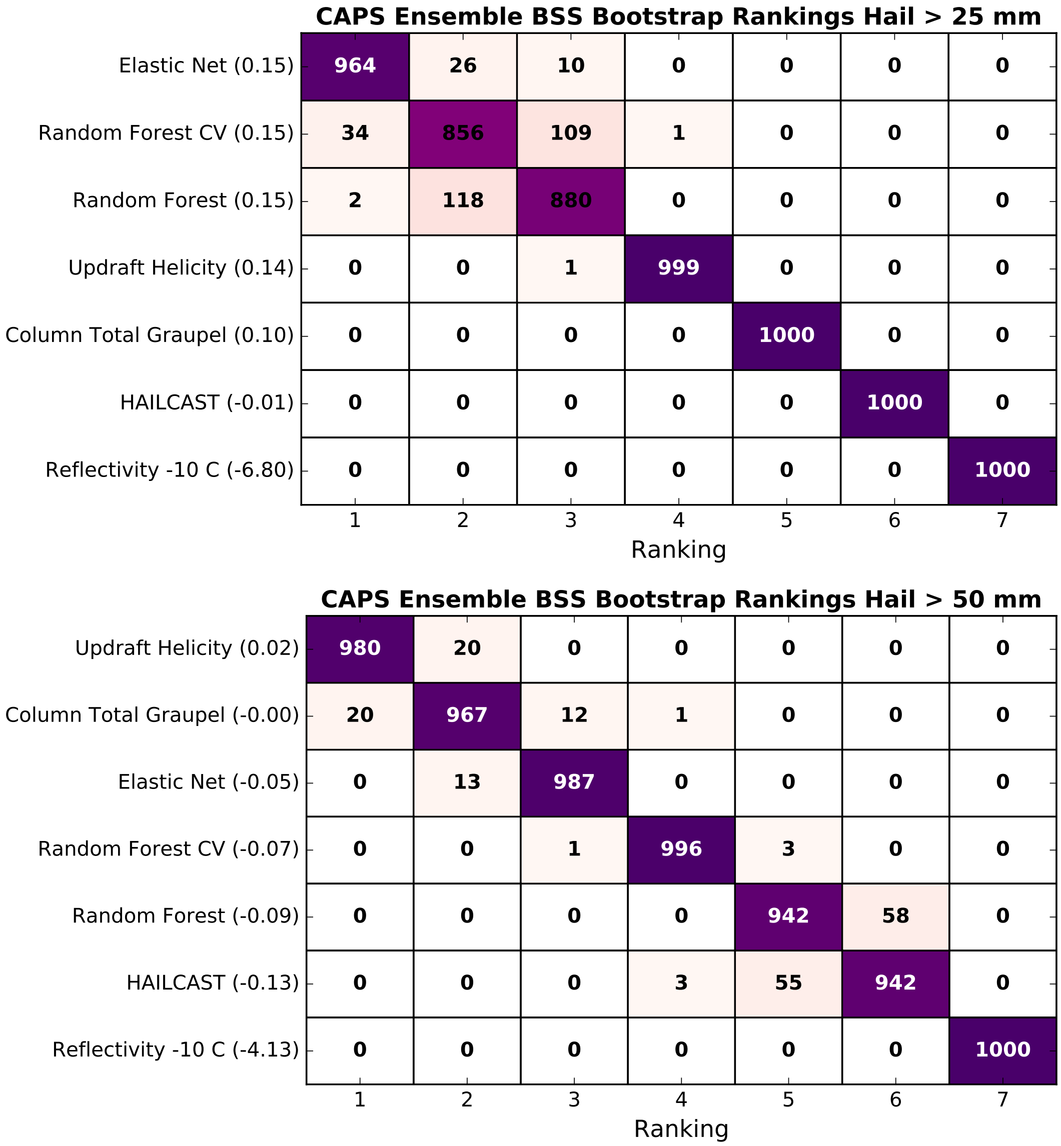


Figure 4.23: Frequency of BSS rankings in a paired bootstrap comparison of the different methods for probability of 25 and 50 mm hail. The mean BSS for each method is shown in parentheses. All differences in BSS were statistically significant at the 5 % threshold based on permutation tests.

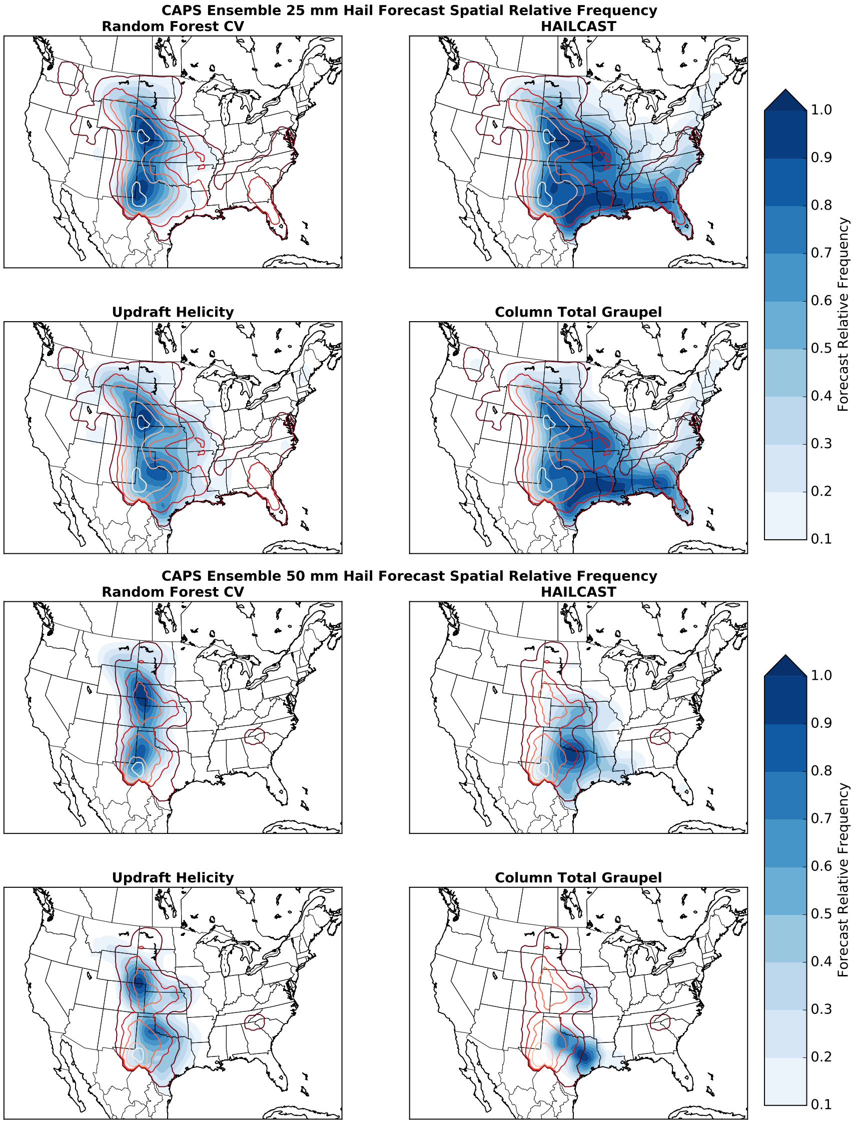


Figure 4.24: CAPS Ensemble spatial distributions of 10% hail forecasts from select models at the 25 and 50 mm hail thresholds. The blue filled contours are forecast relative frequencies, and the red contours are observed relative frequencies.

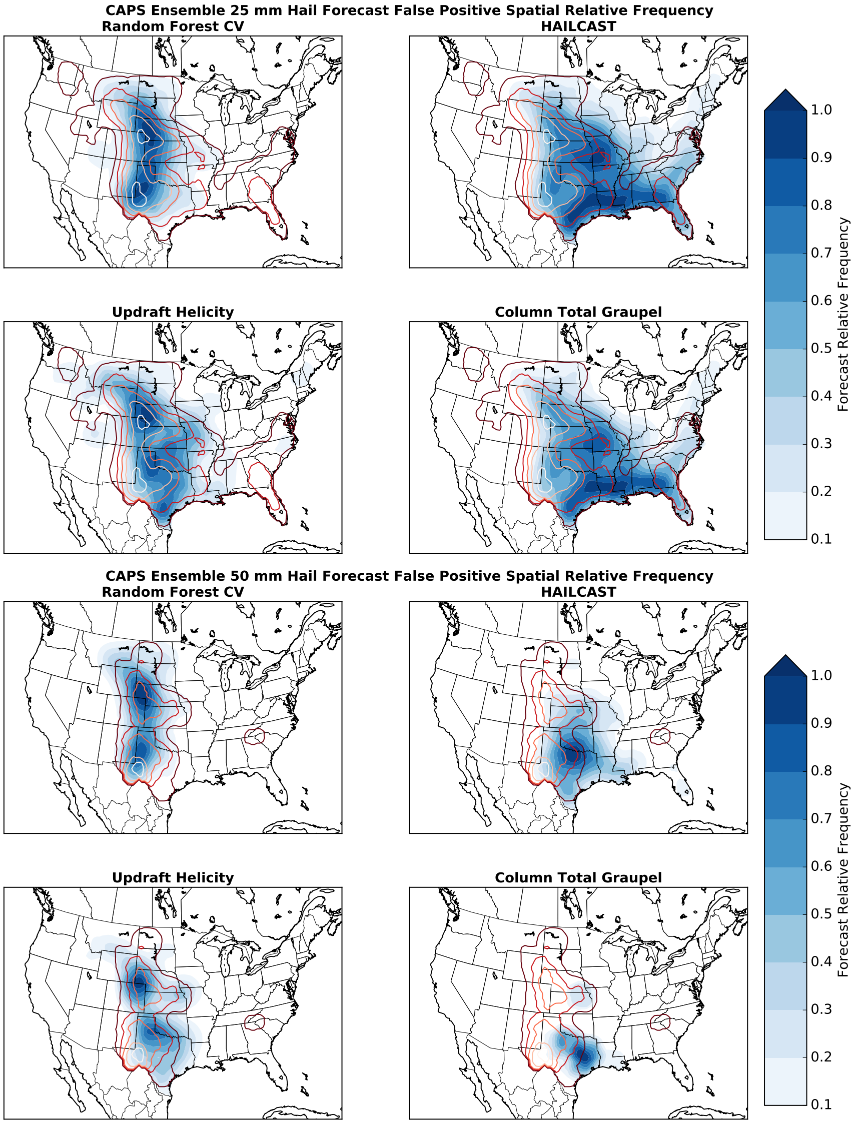

CAPS Ensemble 25 mm Hail Forecast False Positive Spatial Relative Frequency
Random Forest CV
HAILCAST
Updraft Helicity
Column Total Graupel
CAPS Ensemble 50 mm Hail Forecast False Positive Spatial Relative Frequency
Random Forest CV
HAILCAST
Updraft Helicity
Column Total Graupel
Forecast Relative Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Forecast Relative Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Figure 4.25: CAPS Ensemble spatial distributions of 10% hail forecast false positives from select models at the 25 and 50 mm hail thresholds. The blue filled contours are forecast relative frequencies, and the red contours are observed relative frequencies.

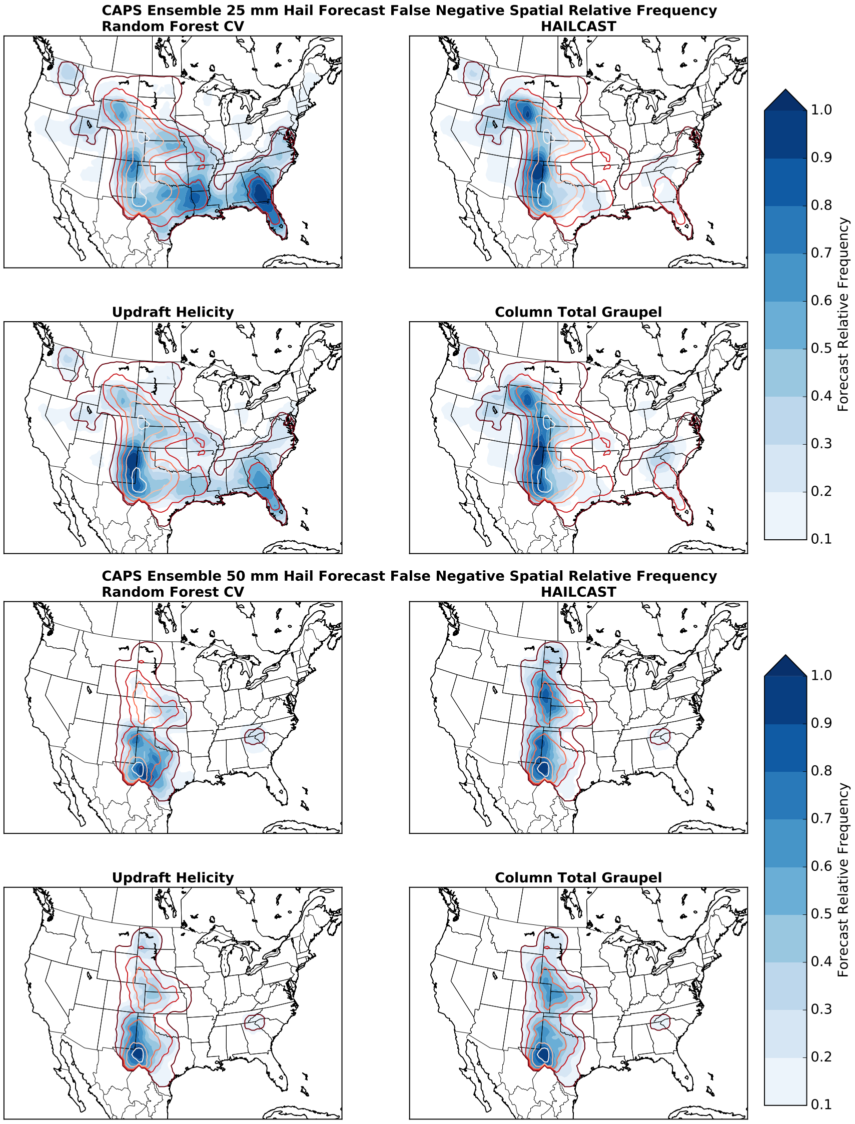

CAPS Ensemble 25 mm Hail Forecast False Negative Spatial Relative Frequency
Random Forest CV
HAILCAST
Updraft Helicity
Column Total Graupel
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Forecast Relative Frequency
CAPS Ensemble 50 mm Hail Forecast False Negative Spatial Relative Frequency
Random Forest CV
HAILCAST
Updraft Helicity
Column Total Graupel
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Forecast Relative Frequency

Figure 4.26: CAPS Ensemble spatial distributions of 10% hail forecast false negatives from select models at the 25 and 50 mm hail thresholds. The blue filled contours are forecast relative frequencies, and the red contours are observed relative frequencies.

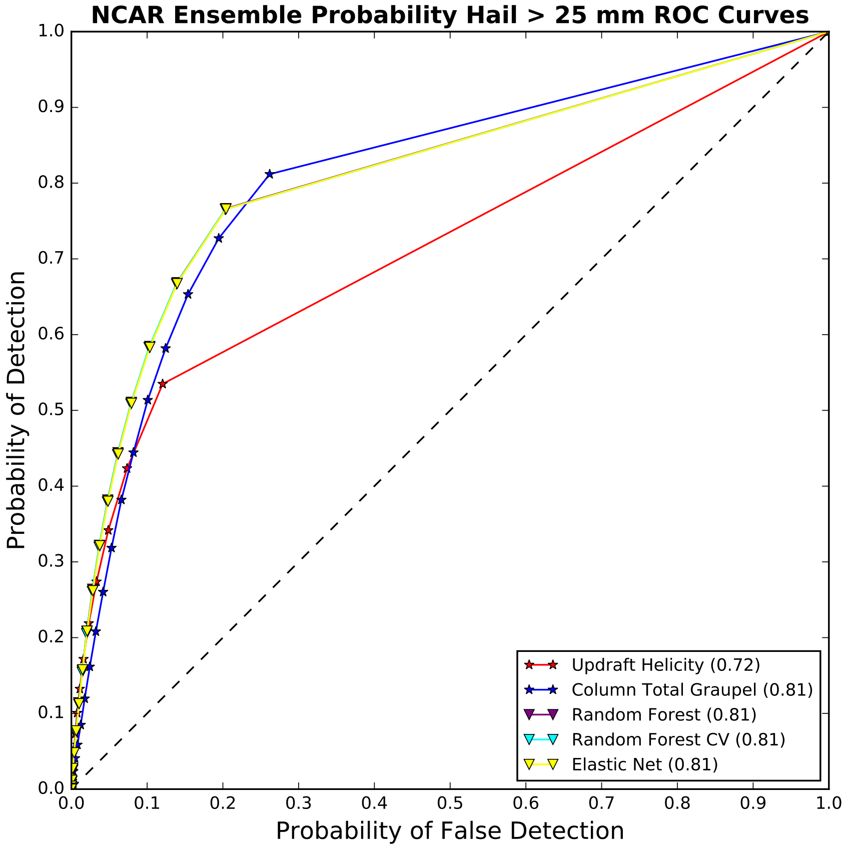


Figure 4.27: ROC curves for each NCAR ensemble mean storm-surrogate probability of 25 mm hail. The AUC for each curve is in the legend.

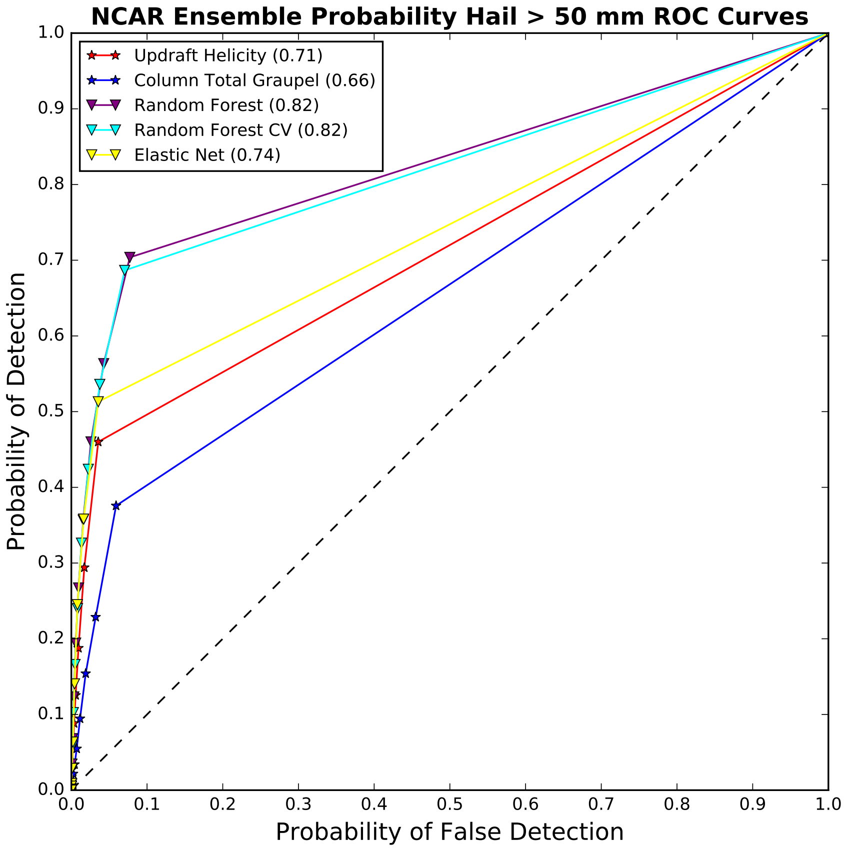


Figure 4.28: ROC curves for each NCAR ensemble mean storm-surrogate probability of 50 mm hail. The AUC for each curve is in the legend.

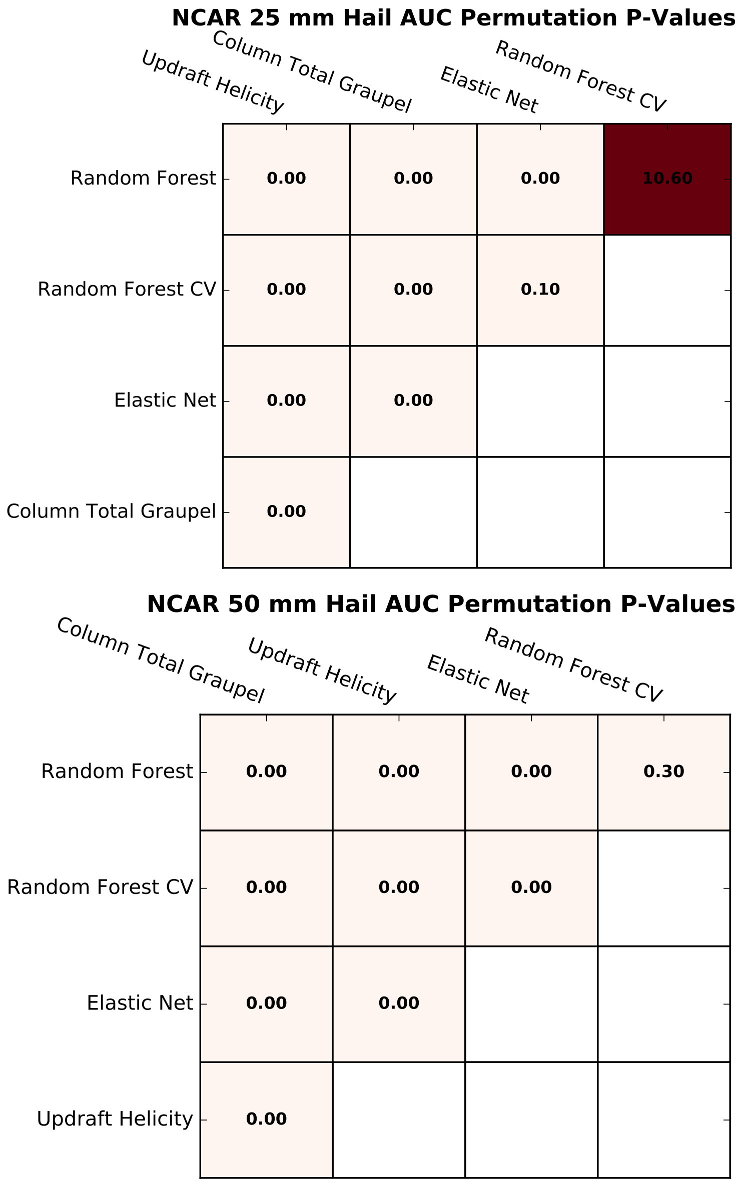
The NCAR ensemble verification results show a much clearer separation between the storm surrogates and machine learning hail forecast methods. The machine learning models at the 25 mm threshold are able to attain slightly lower POD than column total graupel while reducing the POFD (Fig. 4.27). Column total graupel also has a higher probability of detection than updraft helicity. The different machine learning methods differentiate themselves at the 50 mm threshold with the cross-validated Random Forest having the highest POD, followed by default random forest, updraft helicity, elastic net, and column total graupel (Fig. 4.28). All of the methods show discrimination skill at both thresholds, but the machine learning models are able to provide additional improvement in the detection of extreme events. Permutation tests applied to the AUC values indicate that all methods have statistically significantly different AUC values at the 5 % level except for 25 mm Random Forest and Random Forest CV forecasts (Fig. 4.29).

The NCAR ensemble performance diagrams highlight the difference in false alarms for each method at high probability thresholds. Updraft helicity has a lower POD than total graupel for 25 mm hail but also has a lower FAR for most probability thresholds (Fig. 4.30). The machine learning methods are able to maintain a lower FAR than updraft helicity for all thresholds. The differences in FAR decreases at the 50 mm threshold but the random forest models maintain a lower FAR and higher POD consistently (Fig. 4.31). The elastic net underperforms the random forest models at this threshold.

In addition to being better discriminators for the NCAR ensemble, the machine learning methods also produce more reliable probabilities. While an overconfidence bias is apparent at both hail size thresholds, the machine learning methods have higher BSSs at 25 (Fig. 4.32) and 50 mm (Fig. 4.33). Both updraft helicity and the machine learning models are underconfident at low probability thresholds and then trend toward overconfidence at higher thresholds while still showing skill. The elastic net produces the most reliable probabilities at 50 mm at the expense of some sharpness compared with updraft helicity (Fig. 4.33). The random forest models are overconfident at 50 mm. Permutation tests of the Brier Score for each method show that all the differences in scores are statistically significant at the 5 % level.

The spatial distribution of hail forecasts from the NCAR ensemble is shown in Fig. 4.34. Most of the observed hail reports occur along the High Plains east of the Rocky Mountains with a smaller secondary maximum in the Southeast. For 25 mm hail, the machine learning methods capture the full extent of the High Plains hail observations while slightly underforecasting hail in the Southeast. Updraft helicity captures the northern hail observations well but underforecasts the frequency of hail in west Texas and eastern New Mexico as well as the Southeast. Column Total Graupel greatly overforecasts hail frequency in the Southeast and underforecasts hail in the High Plains with an eastward bias. The storms that produce hail in the Southeast in May through July tend to be pulse type thunderstorms, which lack rotating updrafts, so updraft helicity will not detect them. An eastward bulge in eastern Nebraska from overnight MCSs is also visible in both the forecast and observed relative frequencies from all models. The machine learning models overforecast the frequency of hail in the

High Plains while Column Total Graupel does the same over the eastern half of the US (Fig. 4.35). The machine learning methods and updraft helicity have the most misses along the Gulf Coast (Fig. 4.36). For 50 mm hail, the Random Forest best captures the full extent of the hail events in the High Plains while the Elastic Net is more concentrated in central Kansas (Fig. 4.34). Updraft helicity is more concentrated in the northern Plains while Column total Graupel is more biased toward Iowa. The false positives are in similar locations (Fig. 4.35) while the misses for all methods are mainly along the eastern edge of the Rockies.



NCAR 25 mm Hail AUC Permutation P-Values

	Updraft Helicity	Column Total Graupel	Elastic Net	Random Forest CV
Random Forest	0.00	0.00	0.00	10.60
Random Forest CV	0.00	0.00	0.10	
Elastic Net	0.00	0.00		
Column Total Graupel	0.00			

NCAR 50 mm Hail AUC Permutation P-Values

	Column Total Graupel	Updraft Helicity	Elastic Net	Random Forest CV
Random Forest	0.00	0.00	0.00	0.30
Random Forest CV	0.00	0.00	0.00	
Elastic Net	0.00	0.00		
Updraft Helicity	0.00			

Figure 4.29: The p-values (multiplied by 100) from permutation tests for the difference in AUC between NCAR ensemble hail forecast models at the 25 and 50 mm thresholds. Statistically significant p-values are based on the false discovery method with a rate of 0.05 (α=0.03). Darker reds are associated with larger p-values.

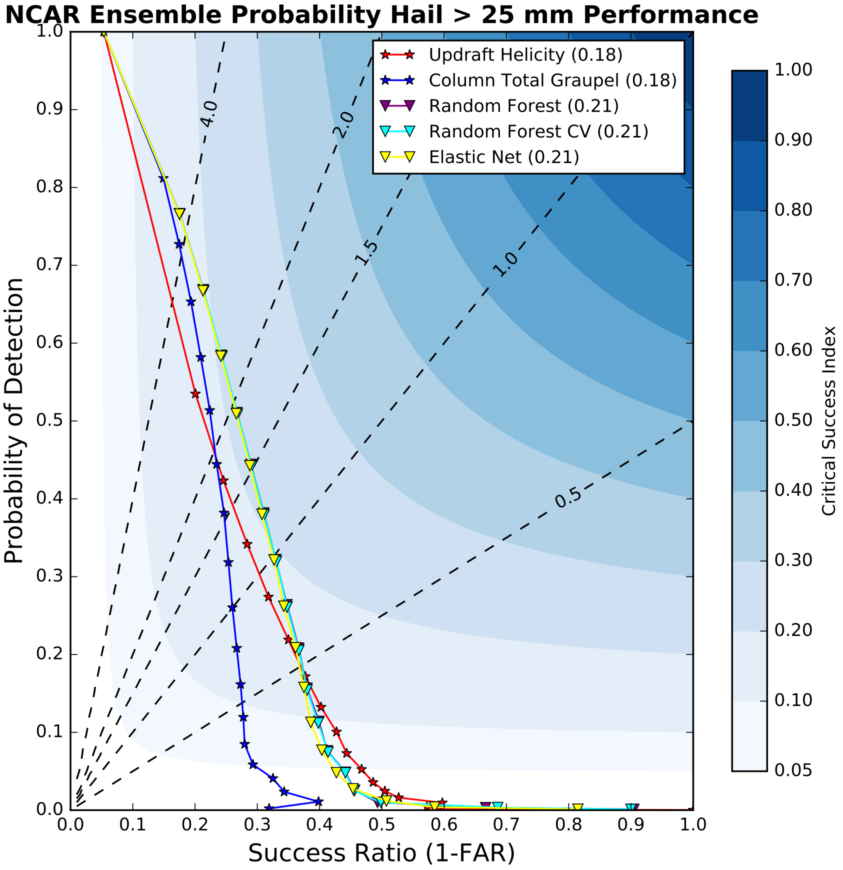


Figure 4.30: Performance curves for each NCAR ensemble mean storm-surrogate probability of 25 mm hail. The maximum CSI for each curve is in the legend.

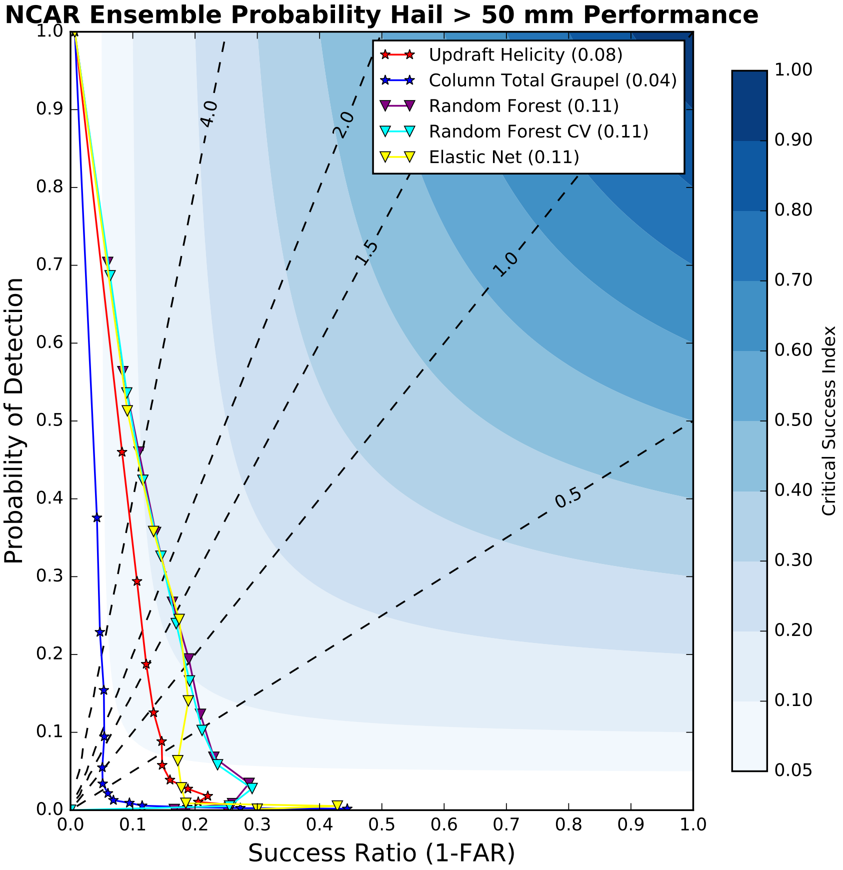


Figure 4.31: Performance curves for each NCAR ensemble mean storm-surrogate probability of 50 mm hail. The maximum CSI for each curve is in the legend.

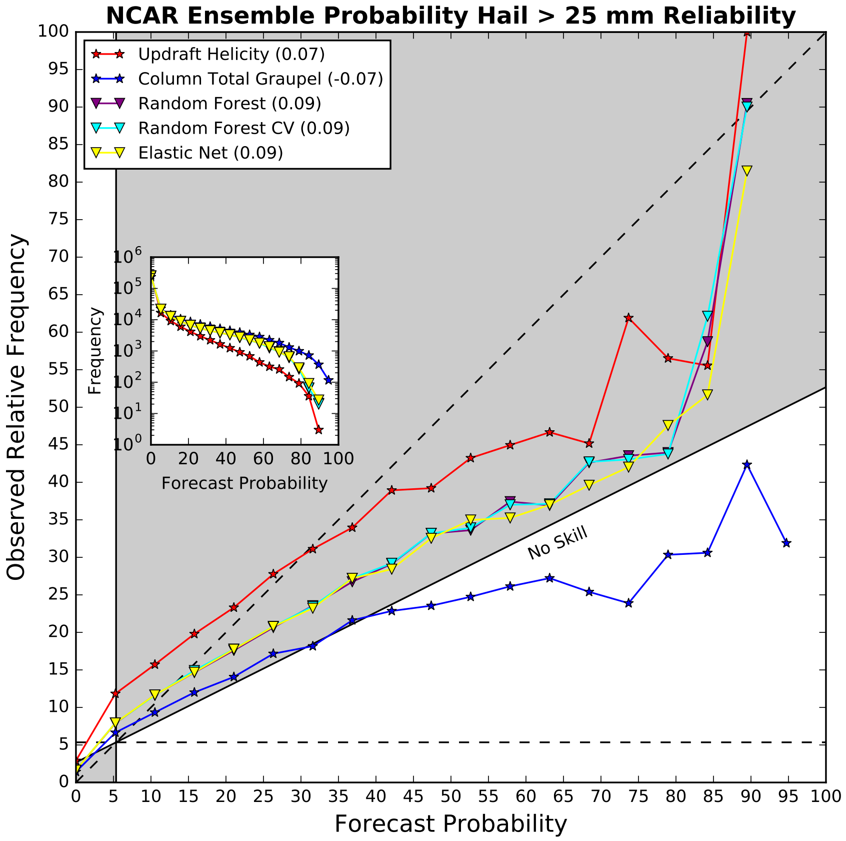


Figure 4.32: Reliability curves for each NCAR ensemble mean storm-surrogate probability of 25 mm hail on an attributes diagram. The maximum BSS for each curve is in the legend.

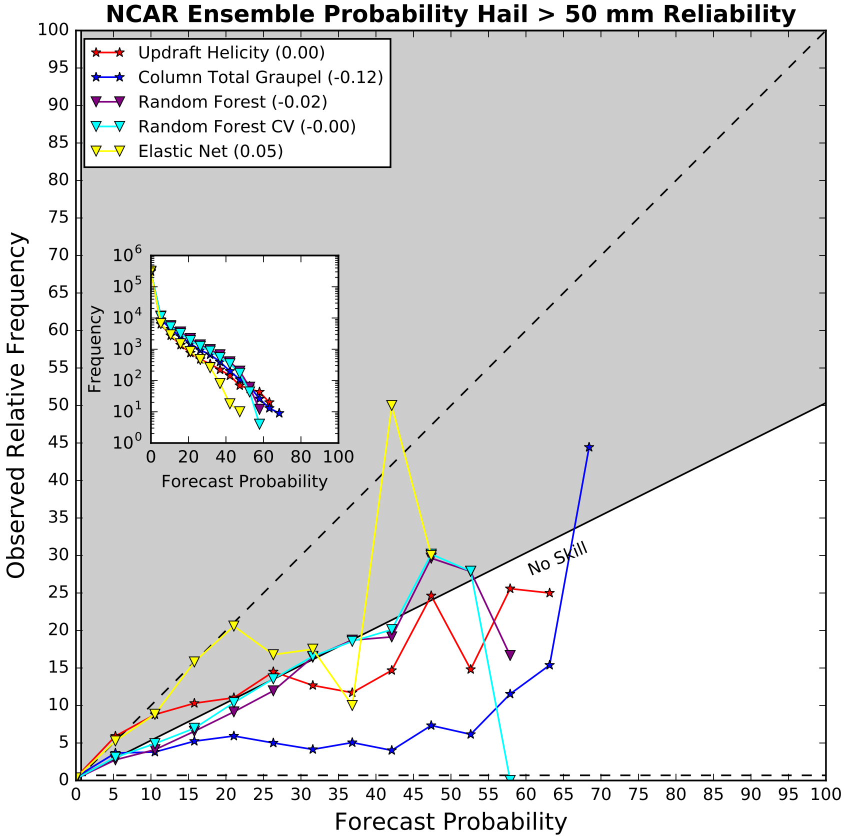


Figure 4.33: Reliability curves for each NCAR ensemble mean storm-surrogate probability of 50 mm hail on an attributes diagram. The maximum BSS for each curve is in the legend.

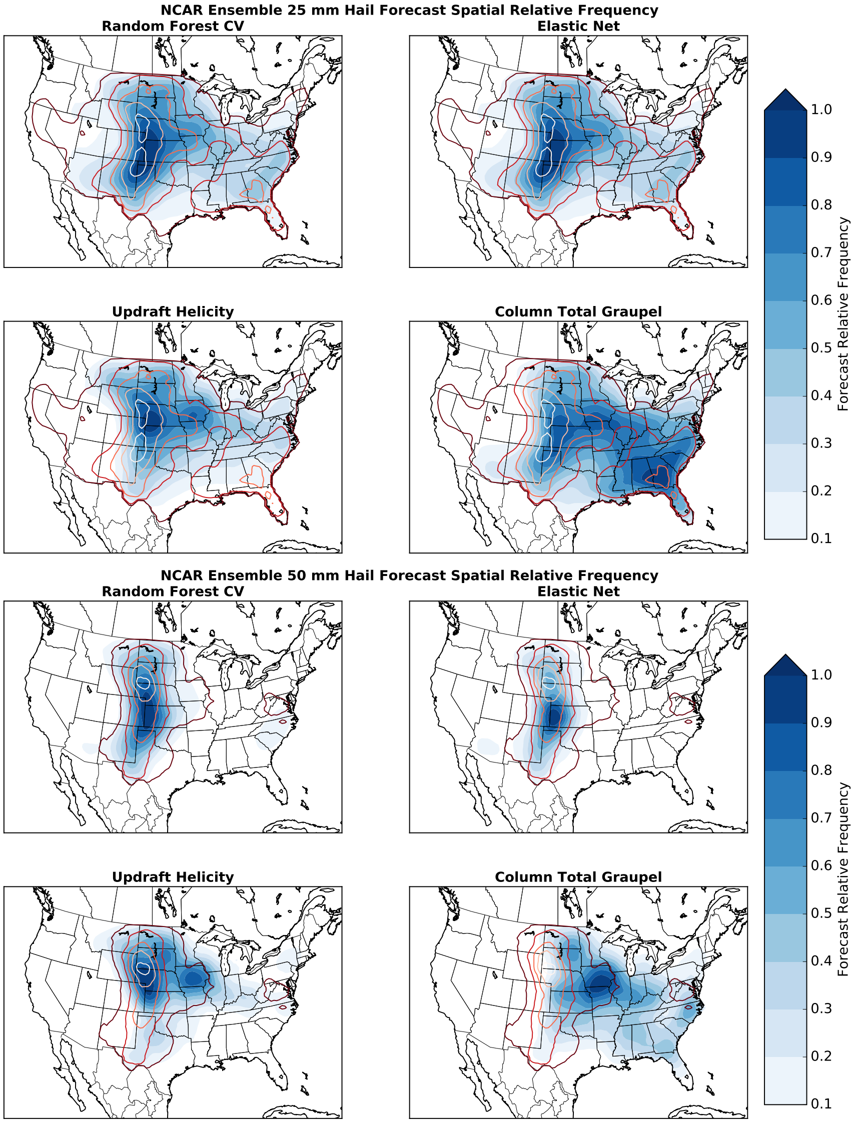
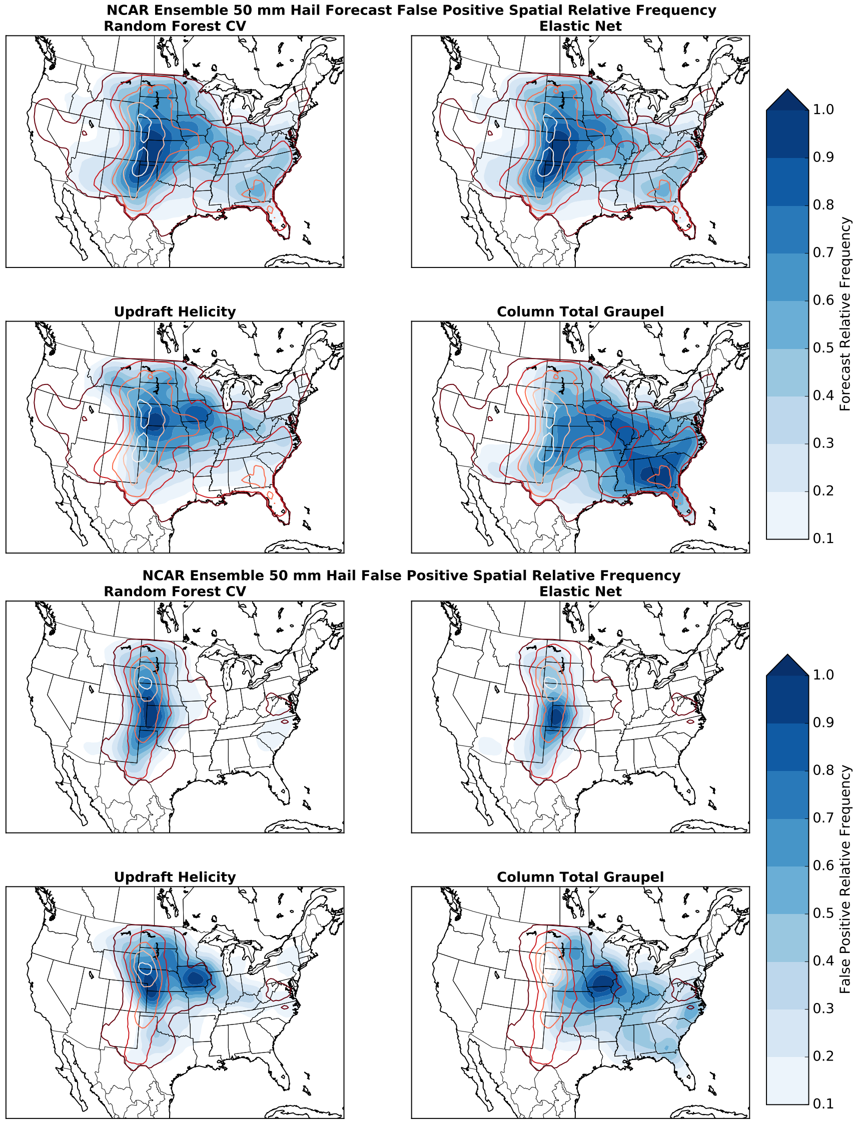


Figure 4.34: Spatial relative frequencies of 25 and 50 mm hail forecasts from the NCAR ensemble in comparison to the spatial relative frequencies of the observations.



NCAR Ensemble 50 mm Hail Forecast False Positive Spatial Relative Frequency

Random Forest CV

Elastic Net

Updraft Helicity

Column Total Graupel

NCAR Ensemble 50 mm Hail False Positive Spatial Relative Frequency

Random Forest CV

Elastic Net

Updraft Helicity

Column Total Graupel

Forecast Relative Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
False Positive Relative Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4.35: Spatial relative frequencies of 25 and 50 mm hail false positive forecasts from the NCAR ensemble in comparison to the spatial relative frequencies of the observations.

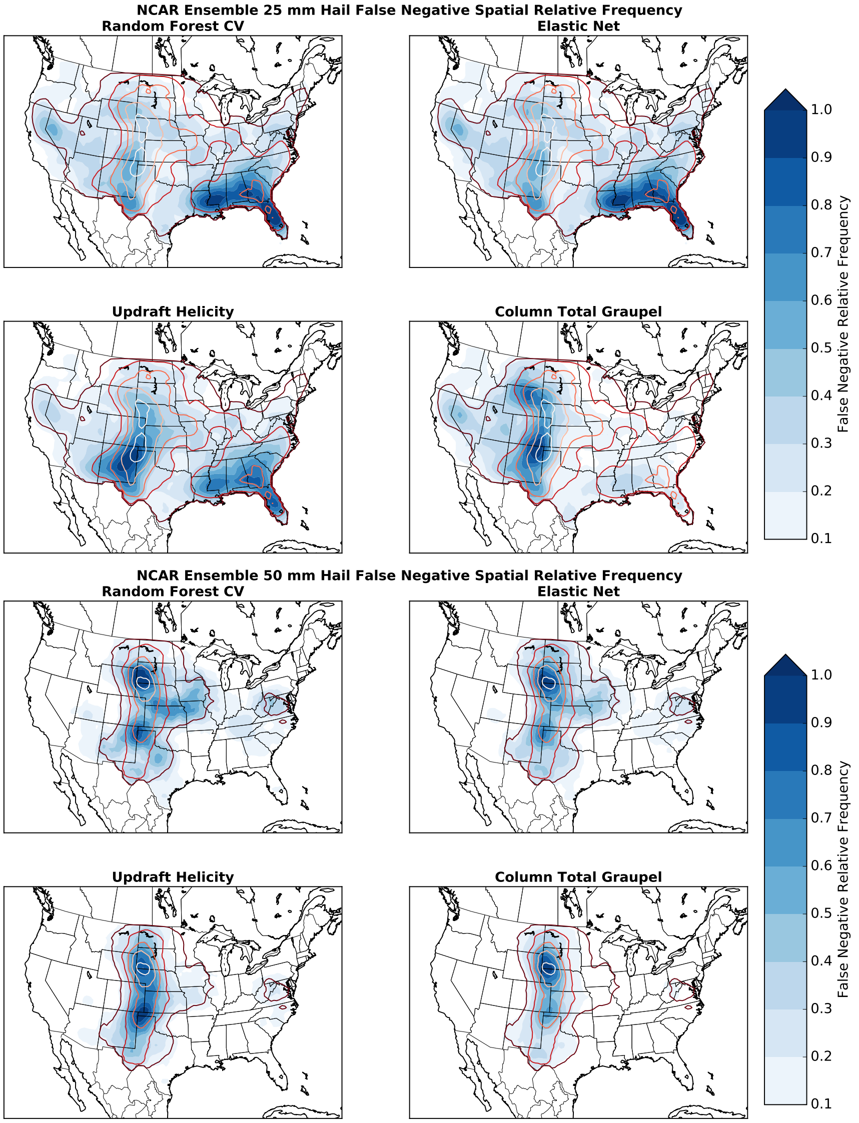

NCAR Ensemble 25 mm Hail False Negative Spatial Relative Frequency
Random Forest CV
Elastic Net
Updraft Helicity
Column Total Graupel
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
False Negative Relative Frequency
NCAR Ensemble 50 mm Hail False Negative Spatial Relative Frequency
Random Forest CV
Elastic Net
Updraft Helicity
Column Total Graupel
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
False Negative Relative Frequency

Figure 4.36: Spatial relative frequencies of 25 and 50 mm hail false negative forecasts from the NCAR ensemble in comparison to the spatial relative frequencies of the observations.

4.2.3 Forecast Examples

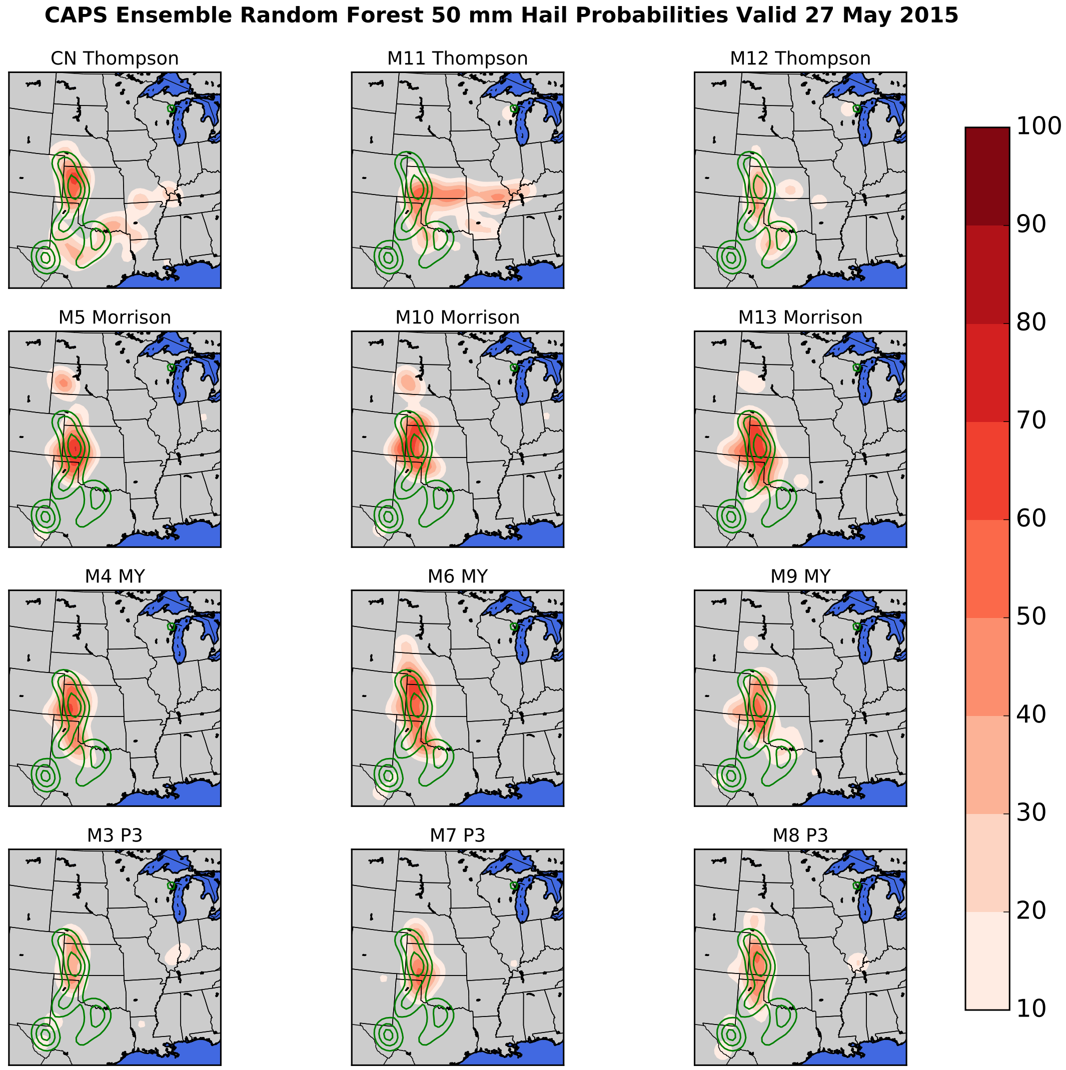


Figure 4.37: Surrogate severe probabilities of hail at least 50 mm in diameter for the period from 12 UTC 27 May 2015 to 12 UTC 28 May 2015. The green contours indicate the practically perfect probabilities of MESH over 50 mm.

Individual forecasts examples show the highlights and failure modes of the different hail forecasting methods. On 27 May 2015, significant hail was reported from isolated supercells in the high Plains from northeast Colorado through west Texas. The random forest hail model applied to the CAPS ensemble captured the extent of this event very well (Fig. 4.37) while HAILCAST (Fig. 4.38) only captured part of the area affected and placed higher hail probabilities on an MCS in Louisiana and Mississippi that produced a large number of wind reports but no significant hail. Ensemble member forecasts tended to cluster by microphysics scheme. The Thompson random forest forecasts tended to have probabilities extend through eastern Kansas and north Texas while the other members confined their probabilities to the High Plains. The HAILCAST members showed a much higher degree of variance than the random forest members and tended to place the highest probabilities in Louisiana. The MY members had much lower probabilities compared to the members from the other schemes.

The CAPS ensemble machine learning model forecasts capture the high end areas of the event well but do not perform as well at the lower hail threshold while the storm surrogate methods generally have the opposite problem (Fig. 4.39 and 4.40. The machine learning models capture the severe hail in Wisconsin and Michigan but miss the event in Alabama while the other methods capture all of the hail areas but have a major false alarm high confidence area in Louisiana. The machine learning models capture the 50 mm hail areas well and the axis along which they occur while HAILCAST and updraft helicity underestimate the extent.

The NCAR ensemble machine learning models do a much better job of capturing both the 25 and 50 mm hail events. The high probability area in the High Plains is much more closely aligned with the reports, and all the hail areas in Alabama are captured (Fig. 4.41). Updraft helicity and column total graupel also do a good job capturing the High Plains but are under and over confident with the eastern threat area. At 50 mm, Random Forest CV has the sharpest machine learning probabilities and covers all observed 50 mm hail, unlike the other two machine learning methods (Fig. 4.42). Column total graupel has a large false alarm area to the east while updraft helicity is less intense there. All do a good job with the main threat area.

On 4 June 2015, a major severe weather event occurred along the Front Range in Colorado and in western Kansas. Significant hail was observed in western Kansas and in northern Colorado. A supercell near Longmont, Colorado produced a westward-moving EF3 tornado, which was witnessed by the author, and large amounts of hail and rain. All methods from the CAPS ensemble captured the extent of the 25 mm hail really well (Fig. 4.43) although they showed an eastward bias in their maximum probability axes. The machine learning methods captured the 50 mm hail probabilities better than any of the storm surrogate methods, which either missed or had low confidence in the storms along the mountains (Fig. 4.44). The NCAR ensemble also performed well at 25 mm (Fig. 4.45) but had lower confidence at 50 mm (Fig. 4.46). The Random Forest CV captured all of the hail areas while the other machine learning models missed the storms in eastern Kansas. Having a more diverse set of ensemble members appeared the help the CAPS ensemble perform better overall on this case.

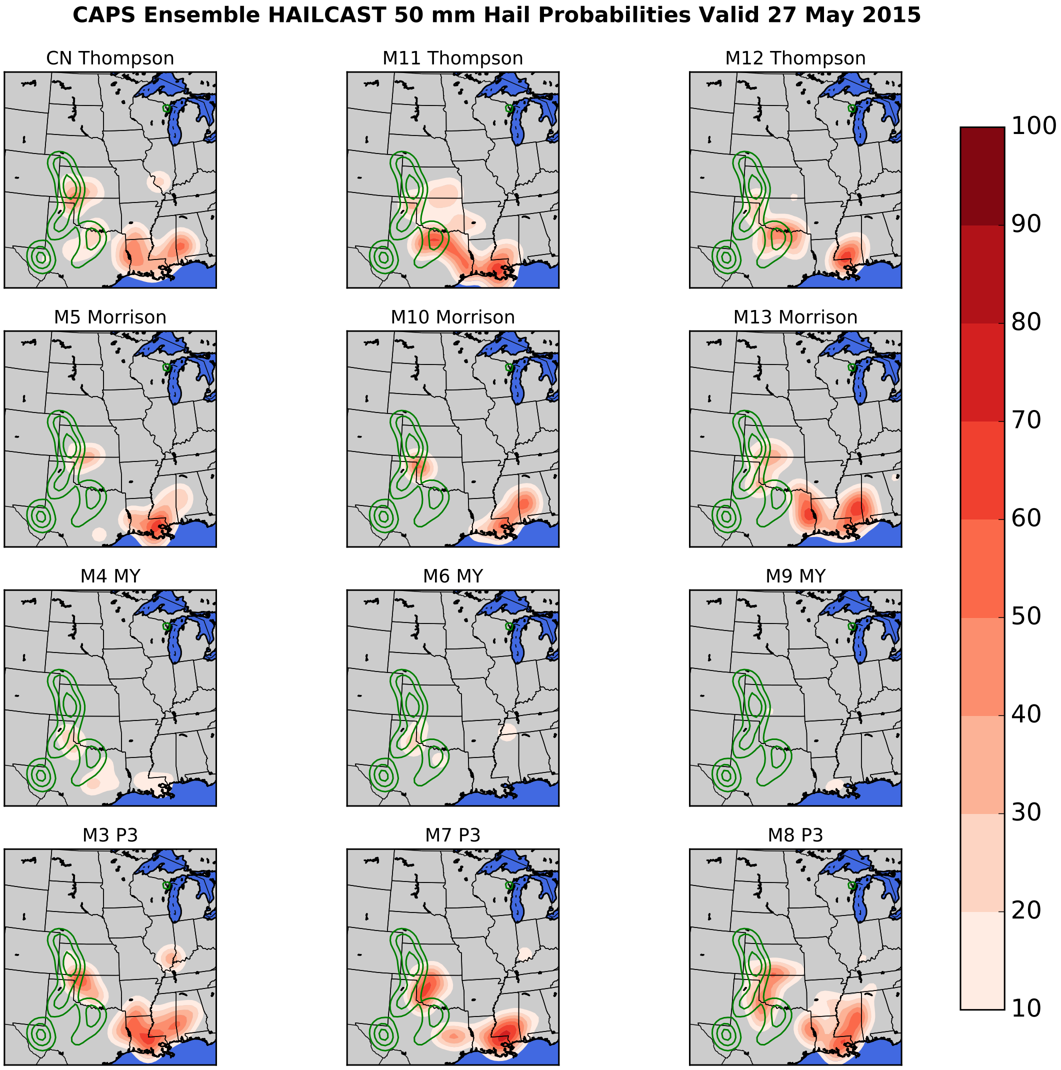


Figure 4.38: Surrogate severe probabilities of hail at least 50 mm in diameter for the period from 12 UTC 27 May 2015 to 12 UTC 28 May 2015. The green contours indicate the practically perfect probabilities of MESH over 50 mm.

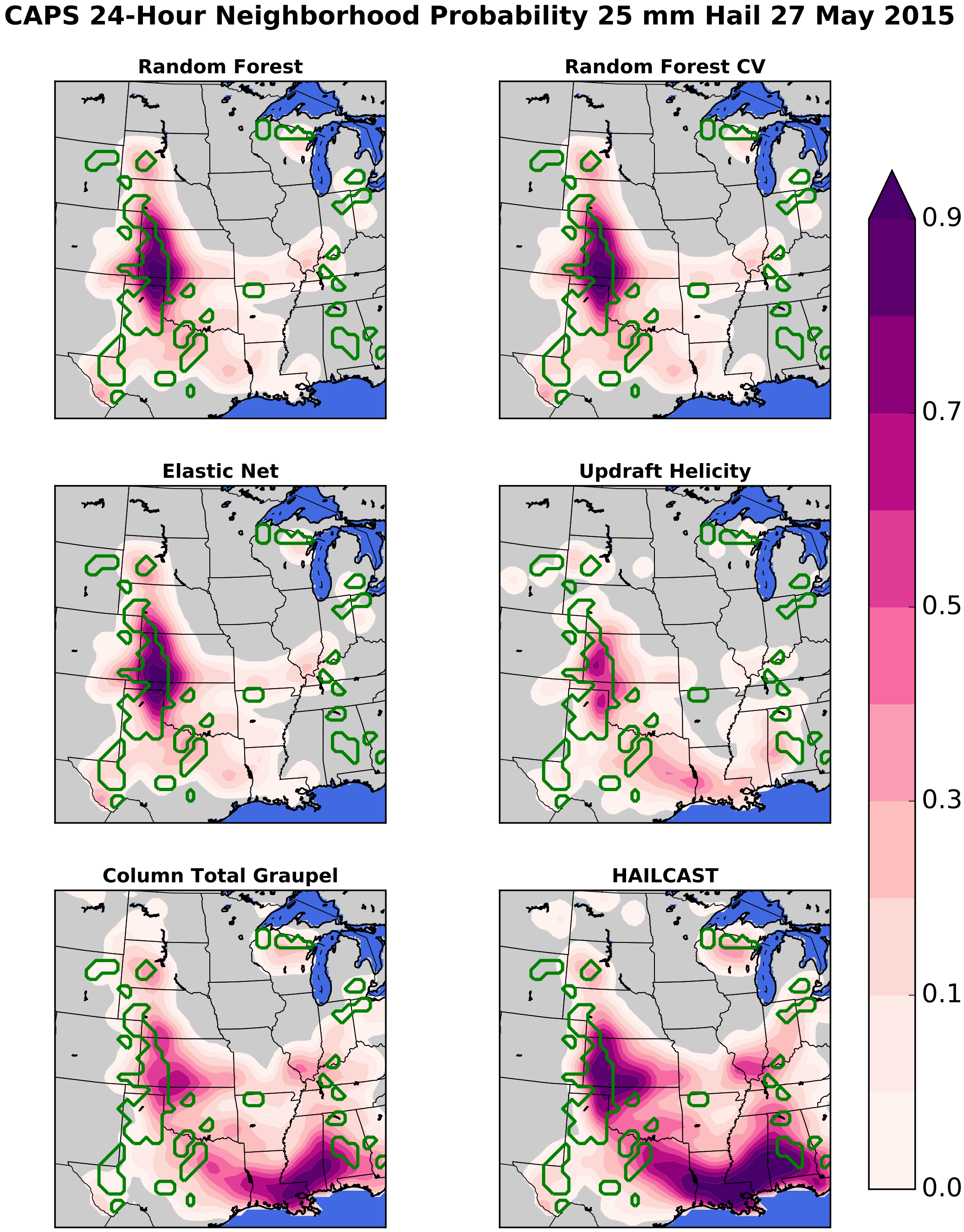


Figure 4.39: CAPS ensemble storm surrogate probabilities of 25 mm hail from each CAPS Ensemble method for 27 May 2015. Observed 25 mm hail is contoured in green.

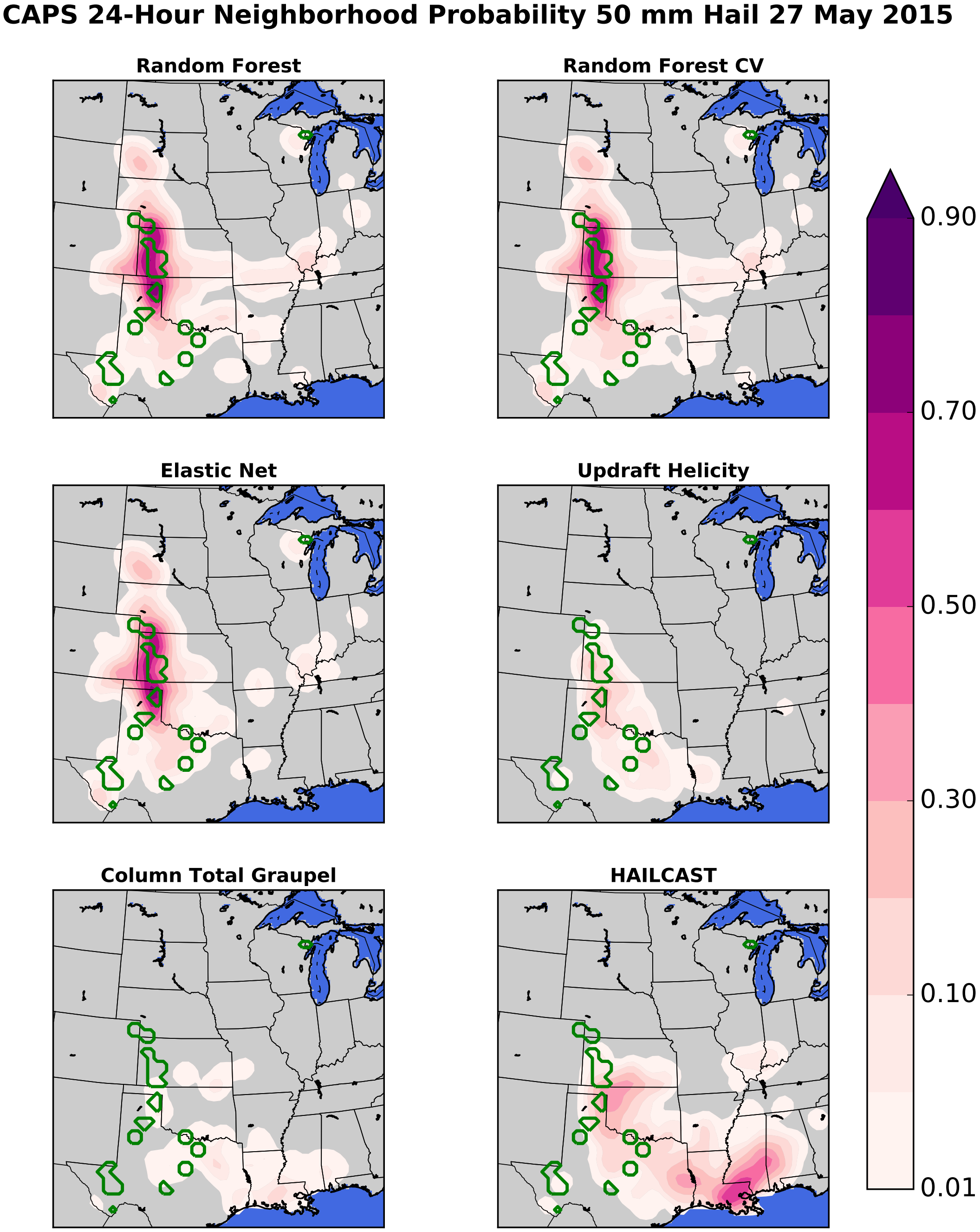


Figure 4.40: CAPS ensemble storm surrogate probabilities of 50 mm hail from each CAPS Ensemble method for 27 May 2015. Observed 50 mm hail is contoured in green.

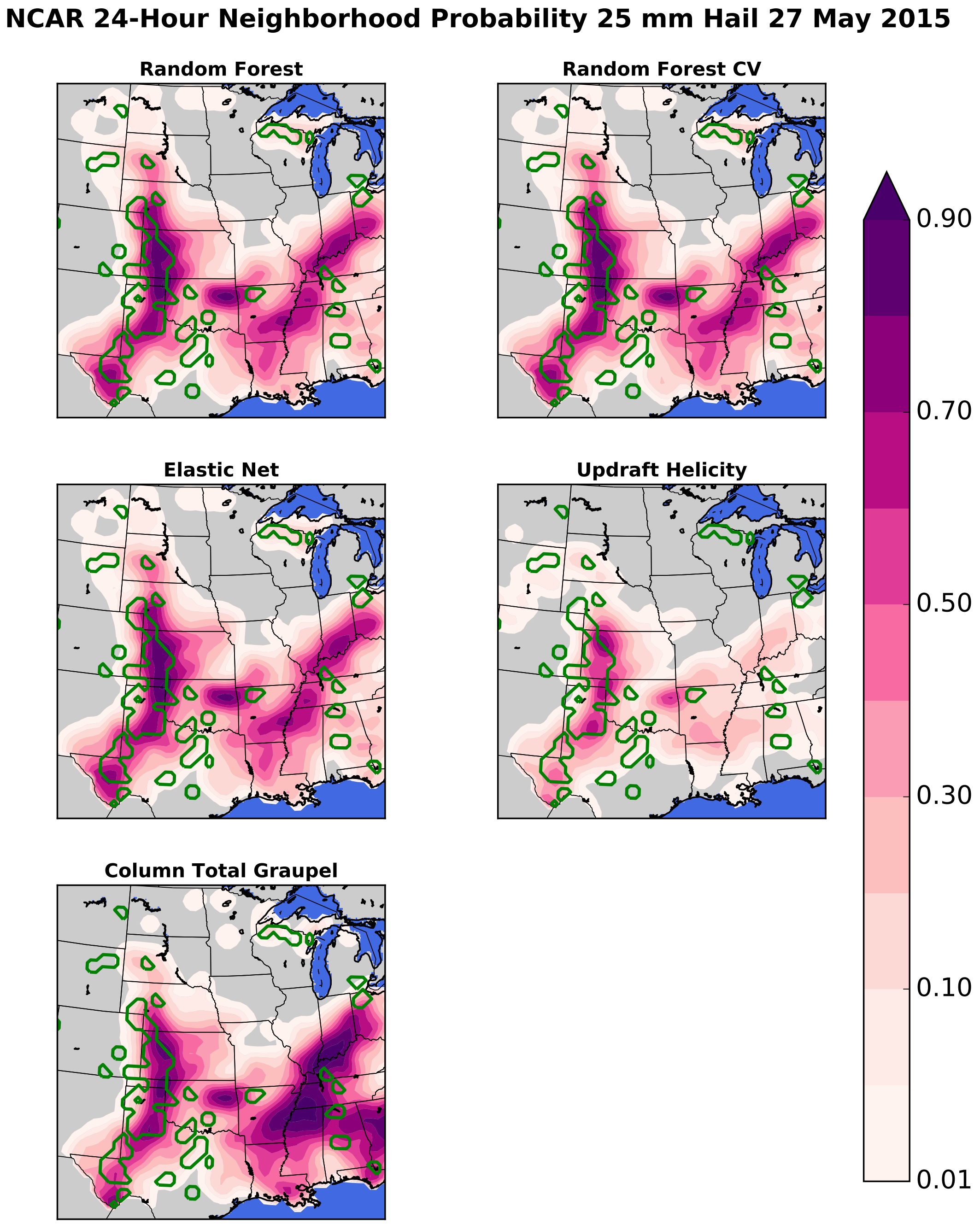


Figure 4.41: NCAR ensemble surrogate probabilities of 25 mm hail from each CAPS Ensemble method for 27 May 2015. Observed 25 mm hail is contoured in green.

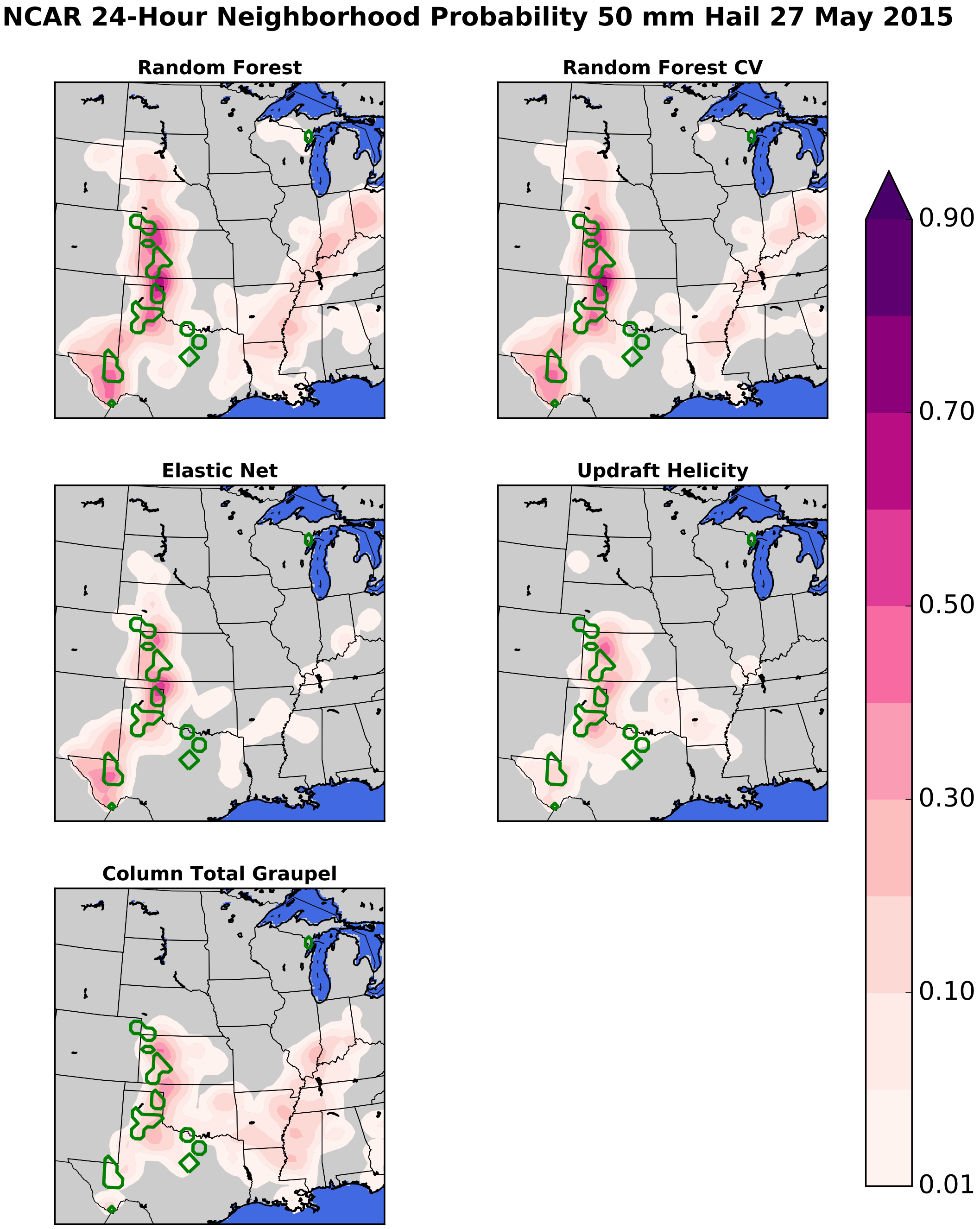


Figure 4.42: NCAR ensemble surrogate probabilities of 50 mm hail from each CAPS Ensemble method for 27 May 2015. Observed 50 mm hail is contoured in green.

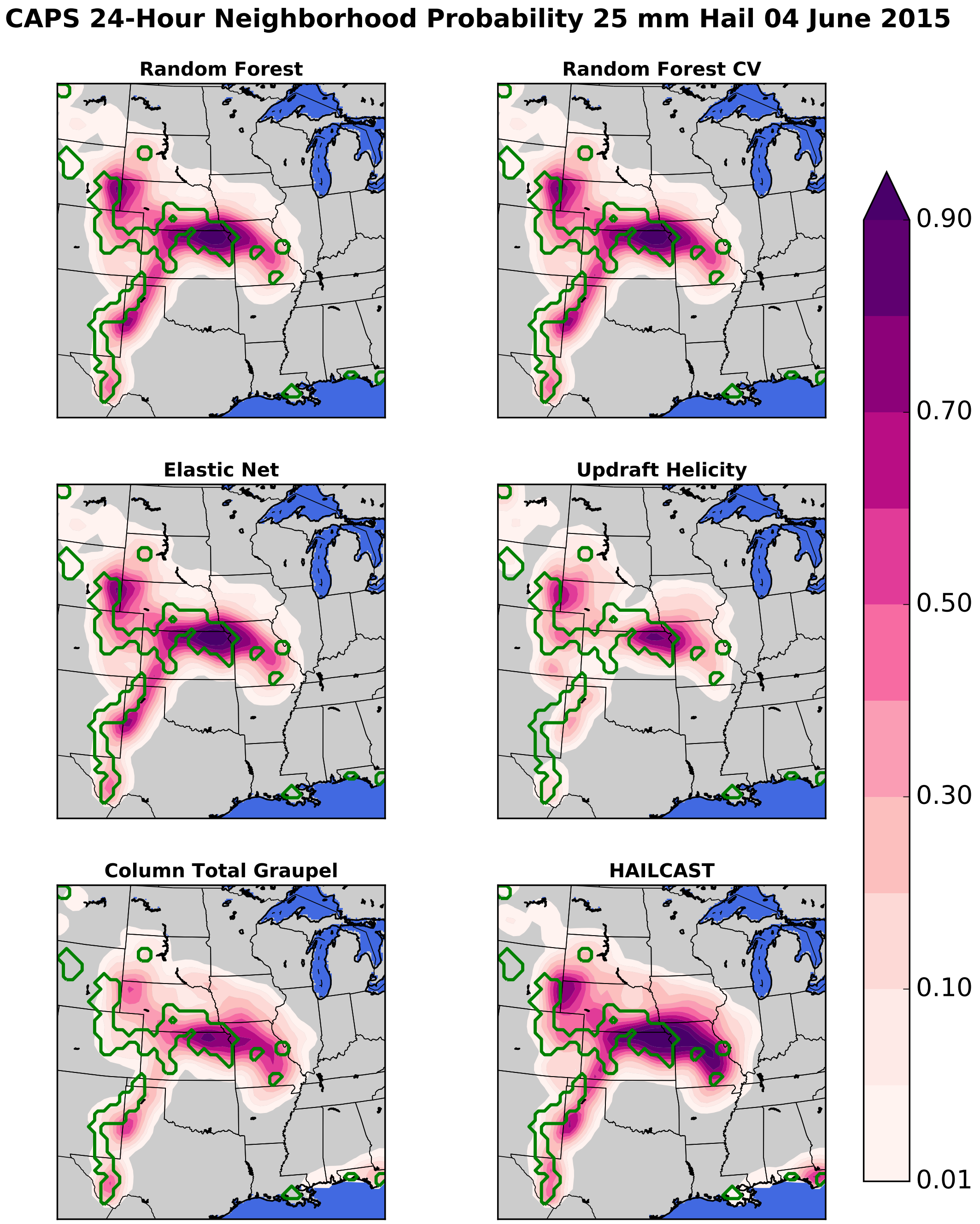


Figure 4.43: CAPS ensemble storm surrogate probabilities of 25 mm hail from each CAPS Ensemble method for 4 June 2015. Observed 25 mm hail is contoured in green.

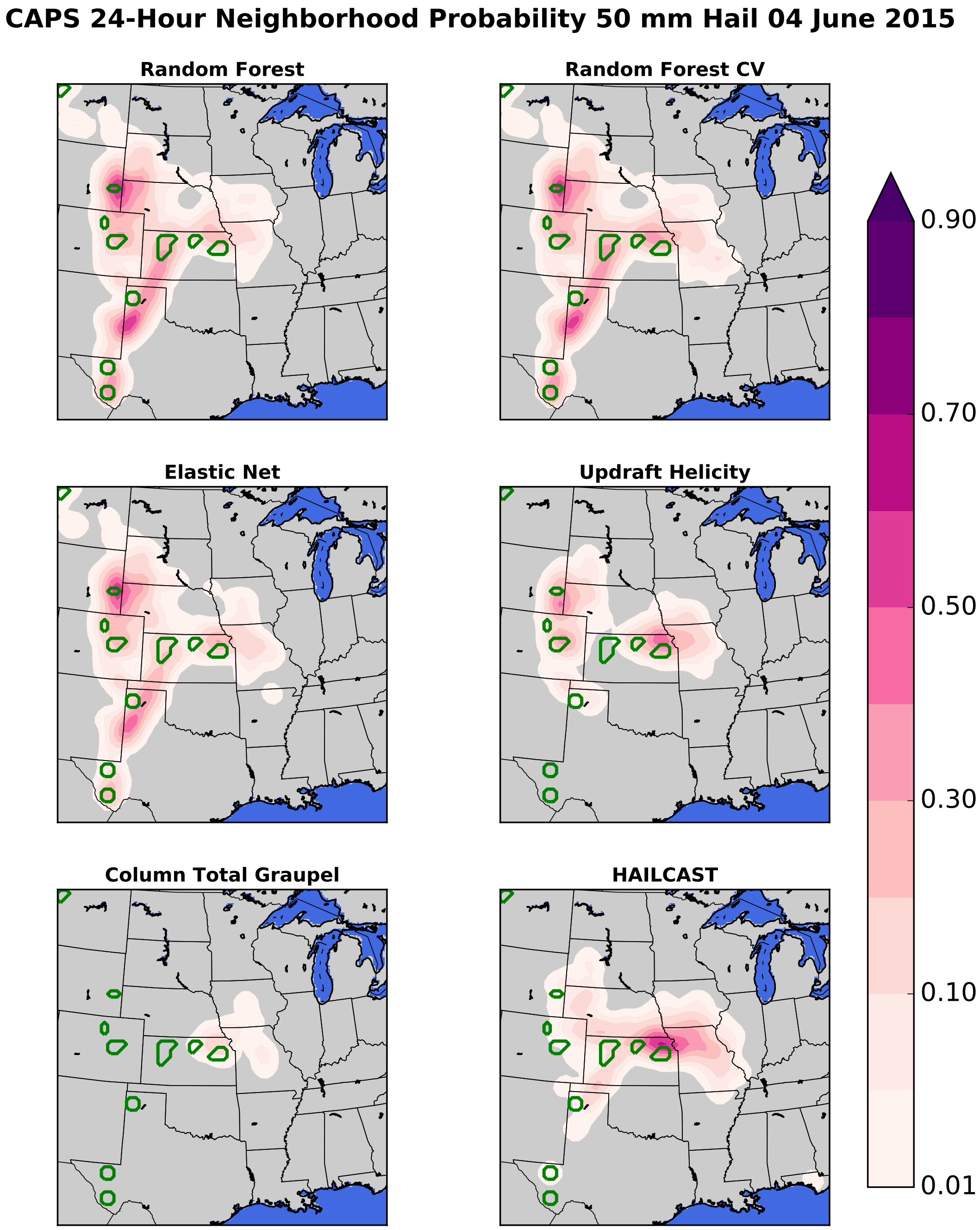


Figure 4.44: CAPS ensemble storm surrogate probabilities of 50 mm hail from each CAPS Ensemble method for 4 June 2015. Observed 50 mm hail is contoured in green.

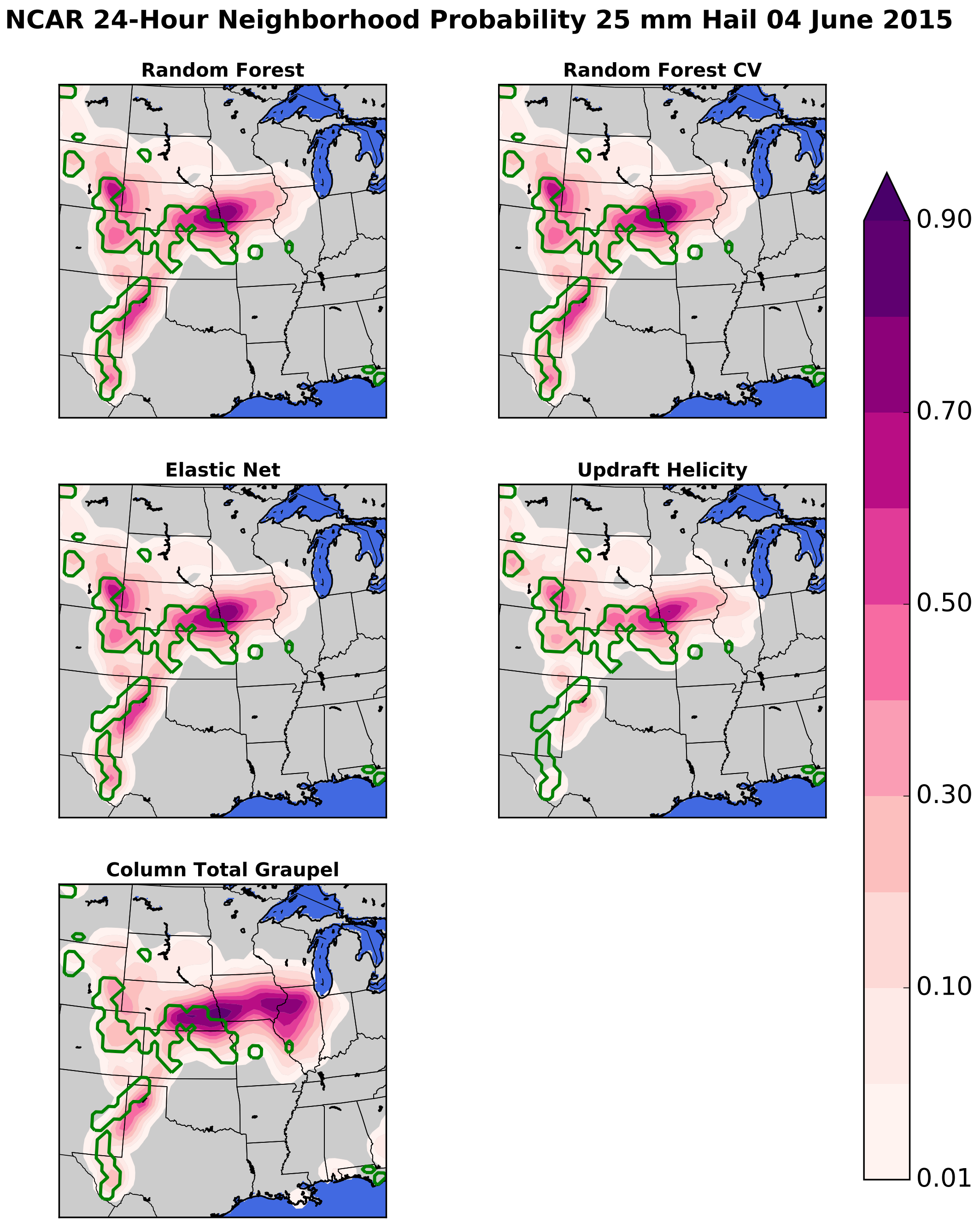


Figure 4.45: NCAR ensemble surrogate probabilities of 25 mm hail from each CAPS Ensemble method for 4 June 2015. Observed 25 mm hail is contoured in green.

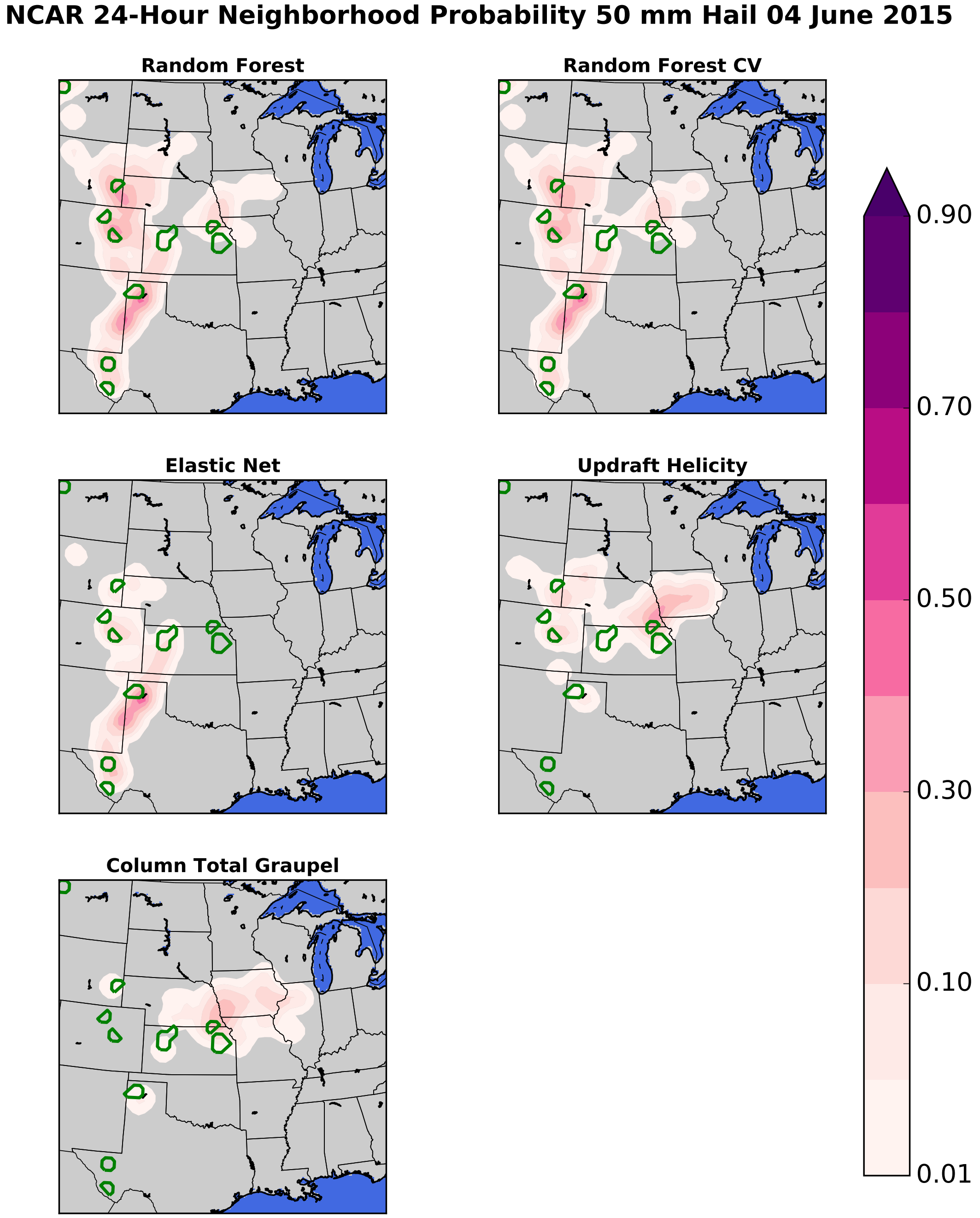


Figure 4.46: NCAR ensemble surrogate probabilities of 50 mm hail from each CAPS Ensemble method for 4 June 2015. Observed 50 mm hail is contoured in green.

4.2.4 Variable Importances

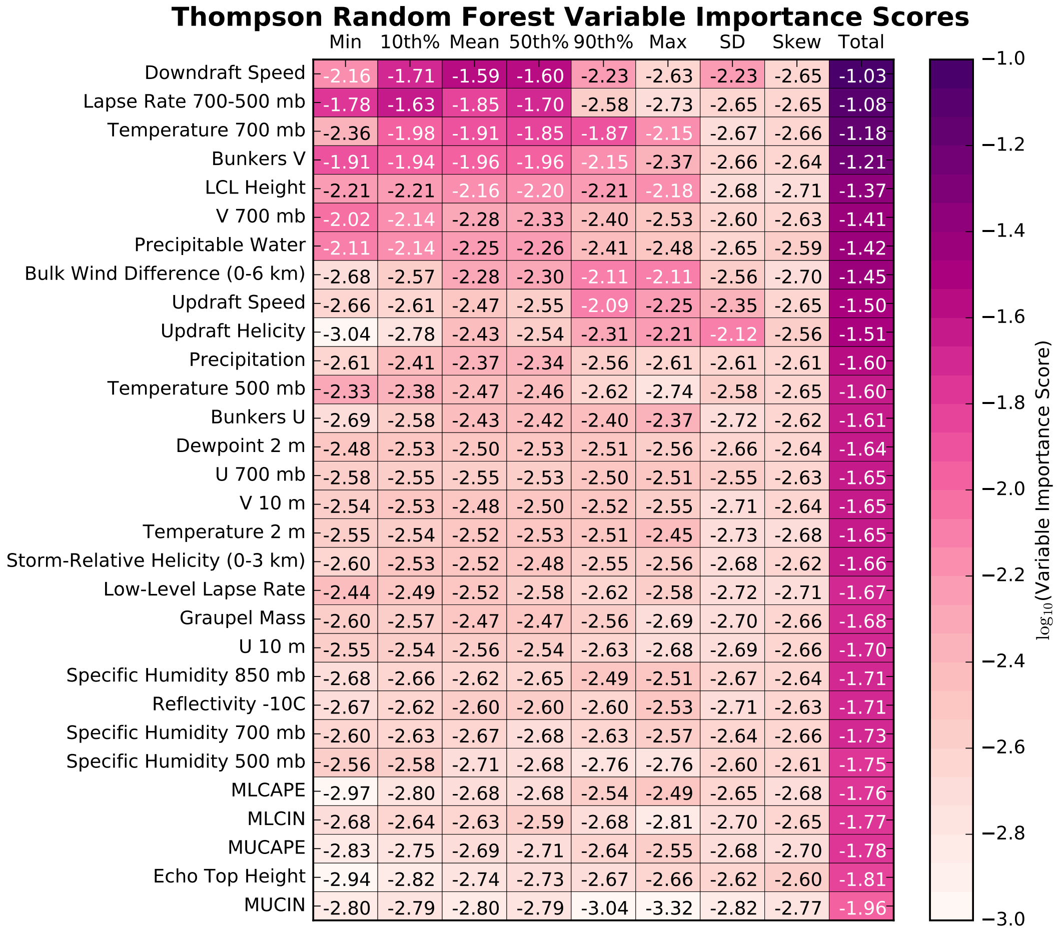
Variable importance scores describe how much each input variable contributes to the performance of a random forest. The variable importance score is calculated by summing the decrease in the impurity metric from parent to child nodes weighted by the number of samples in each node each time a particular variable is used in a decision tree. Then the total decrease in impurity is scaled so that all of the importance scores sum to 1. Finally, the individual tree importance scores are averaged across all trees.

The random forest variable importance procedure was used instead of other feature selection methods, such as wrapper and filter methods, because it accounts for variables that may be useful for segmenting subsets of the training data. Most other feature selection methods rank variables based on their global correlation with the overall output of the training data, which is important for methods like linear regression that are sensitive to overfitting from including too many variables. Random forests, however, are less sensitive to large numbers of input variables, including ones that are highly correlated (Breiman 2001a,b), and can take advantage of variables that may be important only in certain subsets of the feature space to generate increases in predictive accuracy.

Variables with high importance scores tend to be selected more often within each decision tree and impact a larger proportion of the training data cases. If input variables are highly correlated but also have predictive power, then the forest will randomly select among them and effectively split up their total importance (Breiman 2001a). Because of this effect and because multiple statistics were calculated for each input variable, the random forest variable importance scores have been organized into a matrix where each row is an input variable type and each column is a statistic calculated on all the values of that variable within the boundaries of each hailstorm object. The variables are ranked based on the total importance scores summed across all statistics. The logarithms (base 10) of the scores are shown in the matrices to make the differences among the scores more apparent and easily readable.

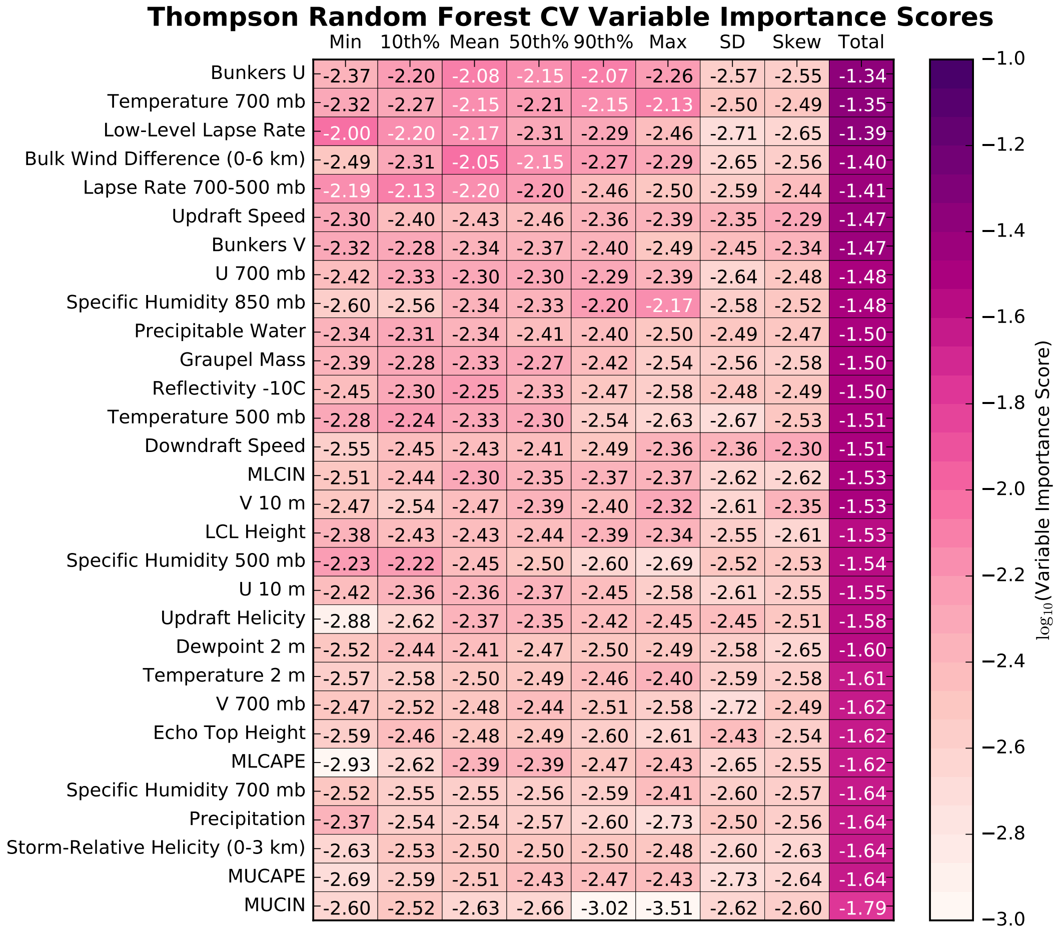
The top variables for predicting hail occurrence are very similar for all microphysics schemes (Fig. 4.47, Fig. 4.49, Fig. 4.51). Downdraft speed is the top variable for all three schemes. Stronger downdrafts allow hail to fall faster and for a longer period through rain-cooled air, which reduces melting and increases the chances of hail reaching the ground. Cooler temperatures at 700 mb and 500 mb also promote hail growth and less melting. Higher 700-500 mb lapse rates result in stronger instability and stronger updrafts in the hail growth zone. Stronger 0-6 km bulk wind differences promote a supercellular storm mode. The Bunkers U and V are tied to mid level winds as well. Finally, strong updrafts and updraft helicity help support the growth of large hail, but updrafts that are too strong can send hail embryos into the anvil before they grow very large.

Some additional variables had more importance for size discrimination (Fig. 4.48, Fig. 4.50, Fig. 4.52). Low-level lapse rates climbed to near the top of the list for each model for size discrimination since it may help distinguish environments supportive of big hail versus any hail at all. While updraft helicity is very helpful for hail occurrence, it does not have a significant impact on determining hail size. Graupel mass is also not the strongest indicator of hail size because it includes hail throughout the atmosphere and not just in the lowest levels.



	Min	10th%	Mean	50th%	90th%	Max	SD	Skew	Total
Downdraft Speed	-2.16	-1.71	-1.59	-1.60	-2.23	-2.63	-2.23	-2.65	-1.03
Lapse Rate 700-500 mb	-1.78	-1.63	-1.85	-1.70	-2.58	-2.73	-2.65	-2.65	-1.08
Temperature 700 mb	-2.36	-1.98	-1.91	-1.85	-1.87	-2.15	-2.67	-2.66	-1.18
Bunkers V	-1.91	-1.94	-1.96	-1.96	-2.15	-2.37	-2.66	-2.64	-1.21
LCL Height	-2.21	-2.21	-2.16	-2.20	-2.21	-2.18	-2.68	-2.71	-1.37
V 700 mb	-2.02	-2.14	-2.28	-2.33	-2.40	-2.53	-2.60	-2.63	-1.41
Precipitable Water	-2.11	-2.14	-2.25	-2.26	-2.41	-2.48	-2.65	-2.59	-1.42
Bulk Wind Difference (0-6 km)	-2.68	-2.57	-2.28	-2.30	-2.11	-2.11	-2.56	-2.70	-1.45
Updraft Speed	-2.66	-2.61	-2.47	-2.55	-2.09	-2.25	-2.35	-2.65	-1.50
Updraft Helicity	-3.04	-2.78	-2.43	-2.54	-2.31	-2.21	-2.12	-2.56	-1.51
Precipitation	-2.61	-2.41	-2.37	-2.34	-2.56	-2.61	-2.61	-2.61	-1.60
Temperature 500 mb	-2.33	-2.38	-2.47	-2.46	-2.62	-2.74	-2.58	-2.65	-1.60
Bunkers U	-2.69	-2.58	-2.43	-2.42	-2.40	-2.37	-2.72	-2.62	-1.61
Dewpoint 2 m	-2.48	-2.53	-2.50	-2.53	-2.51	-2.56	-2.66	-2.64	-1.64
U 700 mb	-2.58	-2.55	-2.55	-2.53	-2.50	-2.51	-2.55	-2.63	-1.65
V 10 m	-2.54	-2.53	-2.48	-2.50	-2.52	-2.55	-2.71	-2.64	-1.65
Temperature 2 m	-2.55	-2.54	-2.52	-2.53	-2.51	-2.45	-2.73	-2.68	-1.65
Storm-Relative Helicity (0-3 km)	-2.60	-2.53	-2.52	-2.48	-2.55	-2.56	-2.68	-2.62	-1.66
Low-Level Lapse Rate	-2.44	-2.49	-2.52	-2.58	-2.62	-2.58	-2.72	-2.71	-1.67
Graupel Mass	-2.60	-2.57	-2.47	-2.47	-2.56	-2.69	-2.70	-2.66	-1.68
U 10 m	-2.55	-2.54	-2.56	-2.54	-2.63	-2.68	-2.69	-2.66	-1.70
Specific Humidity 850 mb	-2.68	-2.66	-2.62	-2.65	-2.49	-2.51	-2.67	-2.64	-1.71
Reflectivity -10C	-2.67	-2.62	-2.60	-2.60	-2.60	-2.53	-2.71	-2.63	-1.71
Specific Humidity 700 mb	-2.60	-2.63	-2.67	-2.68	-2.63	-2.57	-2.64	-2.66	-1.73
Specific Humidity 500 mb	-2.56	-2.58	-2.71	-2.68	-2.76	-2.76	-2.60	-2.61	-1.75
MLCAPE	-2.97	-2.80	-2.68	-2.68	-2.54	-2.49	-2.65	-2.68	-1.76
MLCIN	-2.68	-2.64	-2.63	-2.59	-2.68	-2.81	-2.70	-2.65	-1.77
MUCAPE	-2.83	-2.75	-2.69	-2.71	-2.64	-2.55	-2.68	-2.70	-1.78
Echo Top Height	-2.94	-2.82	-2.74	-2.73	-2.67	-2.66	-2.62	-2.60	-1.81
MUCIN	-2.80	-2.79	-2.80	-2.79	-3.04	-3.32	-2.82	-2.77	-1.96

Figure 4.47: Matrix of the hail occurrence random forest variable importance scores for ensemble members using the Thompson microphysics scheme.



Thompson Random Forest CV Variable Importance Scores

	Min	10th%	Mean	50th%	90th%	Max	SD	Skew	Total
Bunkers U	-2.37	-2.20	-2.08	-2.15	-2.07	-2.26	-2.57	-2.55	-1.34
Temperature 700 mb	-2.32	-2.27	-2.15	-2.21	-2.15	-2.13	-2.50	-2.49	-1.35
Low-Level Lapse Rate	-2.00	-2.20	-2.17	-2.31	-2.29	-2.46	-2.71	-2.65	-1.39
Bulk Wind Difference (0-6 km)	-2.49	-2.31	-2.05	-2.15	-2.27	-2.29	-2.65	-2.56	-1.40
Lapse Rate 700-500 mb	-2.19	-2.13	-2.20	-2.20	-2.46	-2.50	-2.59	-2.44	-1.41
Updraft Speed	-2.30	-2.40	-2.43	-2.46	-2.36	-2.39	-2.35	-2.29	-1.47
Bunkers V	-2.32	-2.28	-2.34	-2.37	-2.40	-2.49	-2.45	-2.34	-1.47
U 700 mb	-2.42	-2.33	-2.30	-2.30	-2.29	-2.39	-2.64	-2.48	-1.48
Specific Humidity 850 mb	-2.60	-2.56	-2.34	-2.33	-2.20	-2.17	-2.58	-2.52	-1.48
Precipitable Water	-2.34	-2.31	-2.34	-2.41	-2.40	-2.50	-2.49	-2.47	-1.50
Graupel Mass	-2.39	-2.28	-2.33	-2.27	-2.42	-2.54	-2.56	-2.58	-1.50
Reflectivity -10C	-2.45	-2.30	-2.25	-2.33	-2.47	-2.58	-2.48	-2.49	-1.50
Temperature 500 mb	-2.28	-2.24	-2.33	-2.30	-2.54	-2.63	-2.67	-2.53	-1.51
Downdraft Speed	-2.55	-2.45	-2.43	-2.41	-2.49	-2.36	-2.36	-2.30	-1.51
MLCIN	-2.51	-2.44	-2.30	-2.35	-2.37	-2.37	-2.62	-2.62	-1.53
V 10 m	-2.47	-2.54	-2.47	-2.39	-2.40	-2.32	-2.61	-2.35	-1.53
LCL Height	-2.38	-2.43	-2.43	-2.44	-2.39	-2.34	-2.55	-2.61	-1.53
Specific Humidity 500 mb	-2.23	-2.22	-2.45	-2.50	-2.60	-2.69	-2.52	-2.53	-1.54
U 10 m	-2.42	-2.36	-2.36	-2.37	-2.45	-2.58	-2.61	-2.55	-1.55
Updraft Helicity	-2.88	-2.62	-2.37	-2.35	-2.42	-2.45	-2.45	-2.51	-1.58
Dewpoint 2 m	-2.52	-2.44	-2.41	-2.47	-2.50	-2.49	-2.58	-2.65	-1.60
Temperature 2 m	-2.57	-2.58	-2.50	-2.49	-2.46	-2.40	-2.59	-2.58	-1.61
V 700 mb	-2.47	-2.52	-2.48	-2.44	-2.51	-2.58	-2.72	-2.49	-1.62
Echo Top Height	-2.59	-2.46	-2.48	-2.49	-2.60	-2.61	-2.43	-2.54	-1.62
MLCAPE	-2.93	-2.62	-2.39	-2.39	-2.47	-2.43	-2.65	-2.55	-1.62
Specific Humidity 700 mb	-2.52	-2.55	-2.55	-2.56	-2.59	-2.41	-2.60	-2.57	-1.64
Precipitation	-2.37	-2.54	-2.54	-2.57	-2.60	-2.73	-2.50	-2.56	-1.64
Storm-Relative Helicity (0-3 km)	-2.63	-2.53	-2.50	-2.50	-2.50	-2.48	-2.60	-2.63	-1.64
MUCAPE	-2.69	-2.59	-2.51	-2.43	-2.47	-2.43	-2.73	-2.64	-1.64
MUCIN	-2.60	-2.52	-2.63	-2.66	-3.02	-3.51	-2.62	-2.60	-1.79

Figure 4.48: Matrix of size distribution random forest variable importance scores for ensemble members using the Thompson microphysics scheme.

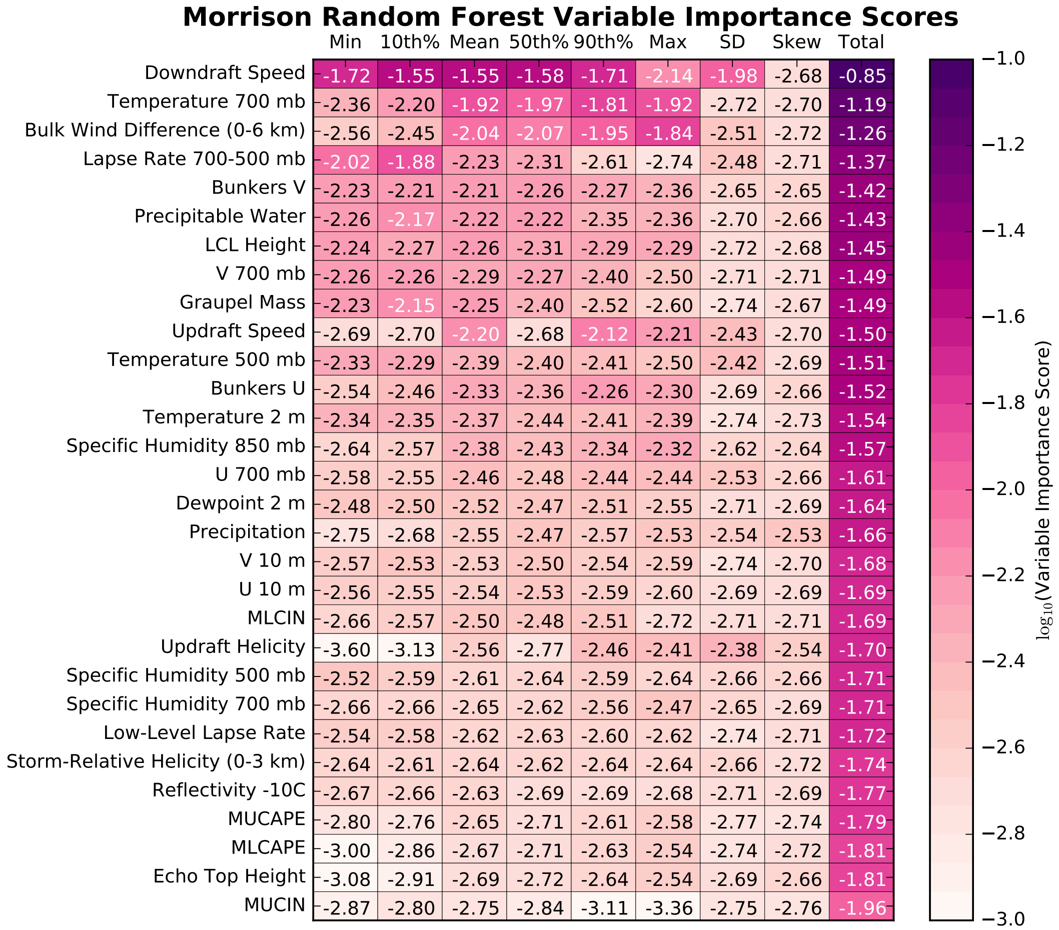
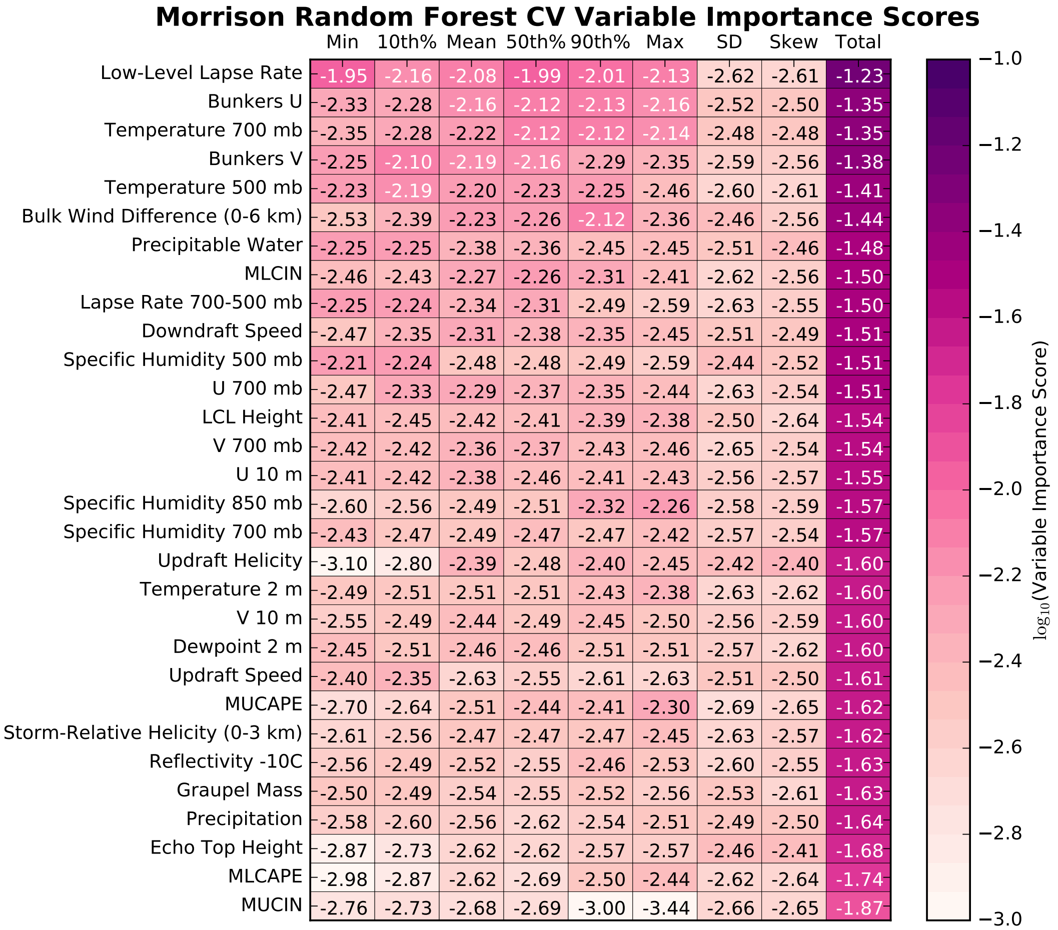
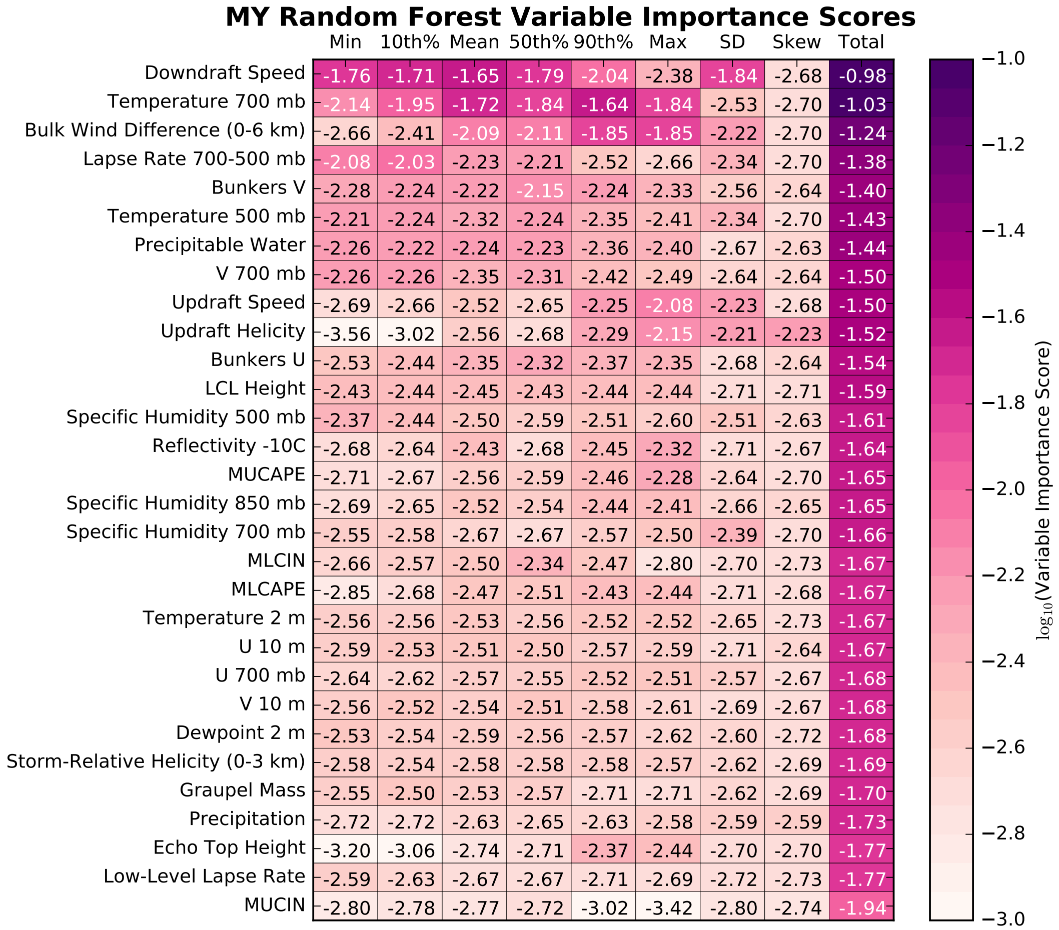


Figure 4.49: Same as Fig. 4.47 but for the hail occurrence random forest trained on Morrison microphysics members.



	Min	10th%	Mean	50th%	90th%	Max	SD	Skew	Total
Low-Level Lapse Rate	-1.95	-2.16	-2.08	-1.99	-2.01	-2.13	-2.62	-2.61	-1.23
Bunkers U	-2.33	-2.28	-2.16	-2.12	-2.13	-2.16	-2.52	-2.50	-1.35
Temperature 700 mb	-2.35	-2.28	-2.22	-2.12	-2.12	-2.14	-2.48	-2.48	-1.35
Bunkers V	-2.25	-2.10	-2.19	-2.16	-2.29	-2.35	-2.59	-2.56	-1.38
Temperature 500 mb	-2.23	-2.19	-2.20	-2.23	-2.25	-2.46	-2.60	-2.61	-1.41
Bulk Wind Difference (0-6 km)	-2.53	-2.39	-2.23	-2.26	-2.12	-2.36	-2.46	-2.56	-1.44
Precipitable Water	-2.25	-2.25	-2.38	-2.36	-2.45	-2.45	-2.51	-2.46	-1.48
MLCIN	-2.46	-2.43	-2.27	-2.26	-2.31	-2.41	-2.62	-2.56	-1.50
Lapse Rate 700-500 mb	-2.25	-2.24	-2.34	-2.31	-2.49	-2.59	-2.63	-2.55	-1.50
Downdraft Speed	-2.47	-2.35	-2.31	-2.38	-2.35	-2.45	-2.51	-2.49	-1.51
Specific Humidity 500 mb	-2.21	-2.24	-2.48	-2.48	-2.49	-2.59	-2.44	-2.52	-1.51
U 700 mb	-2.47	-2.33	-2.29	-2.37	-2.35	-2.44	-2.63	-2.54	-1.51
LCL Height	-2.41	-2.45	-2.42	-2.41	-2.39	-2.38	-2.50	-2.64	-1.54
V 700 mb	-2.42	-2.42	-2.36	-2.37	-2.43	-2.46	-2.65	-2.54	-1.54
U 10 m	-2.41	-2.42	-2.38	-2.46	-2.41	-2.43	-2.56	-2.57	-1.55
Specific Humidity 850 mb	-2.60	-2.56	-2.49	-2.51	-2.32	-2.26	-2.58	-2.59	-1.57
Specific Humidity 700 mb	-2.43	-2.47	-2.49	-2.47	-2.47	-2.42	-2.57	-2.54	-1.57
Updraft Helicity	-3.10	-2.80	-2.39	-2.48	-2.40	-2.45	-2.42	-2.40	-1.60
Temperature 2 m	-2.49	-2.51	-2.51	-2.51	-2.43	-2.38	-2.63	-2.62	-1.60
V 10 m	-2.55	-2.49	-2.44	-2.49	-2.45	-2.50	-2.56	-2.59	-1.60
Dewpoint 2 m	-2.45	-2.51	-2.46	-2.46	-2.51	-2.51	-2.57	-2.62	-1.60
Updraft Speed	-2.40	-2.35	-2.63	-2.55	-2.61	-2.63	-2.51	-2.50	-1.61
MUCAPE	-2.70	-2.64	-2.51	-2.44	-2.41	-2.30	-2.69	-2.65	-1.62
Storm-Relative Helicity (0-3 km)	-2.61	-2.56	-2.47	-2.47	-2.47	-2.45	-2.63	-2.57	-1.62
Reflectivity -10C	-2.56	-2.49	-2.52	-2.55	-2.46	-2.53	-2.60	-2.55	-1.63
Graupel Mass	-2.50	-2.49	-2.54	-2.55	-2.52	-2.56	-2.53	-2.61	-1.63
Precipitation	-2.58	-2.60	-2.56	-2.62	-2.54	-2.51	-2.49	-2.50	-1.64
Echo Top Height	-2.87	-2.73	-2.62	-2.62	-2.57	-2.57	-2.46	-2.41	-1.68
MLCAPE	-2.98	-2.87	-2.62	-2.69	-2.50	-2.44	-2.62	-2.64	-1.74
MUCIN	-2.76	-2.73	-2.68	-2.69	-3.00	-3.44	-2.66	-2.65	-1.87

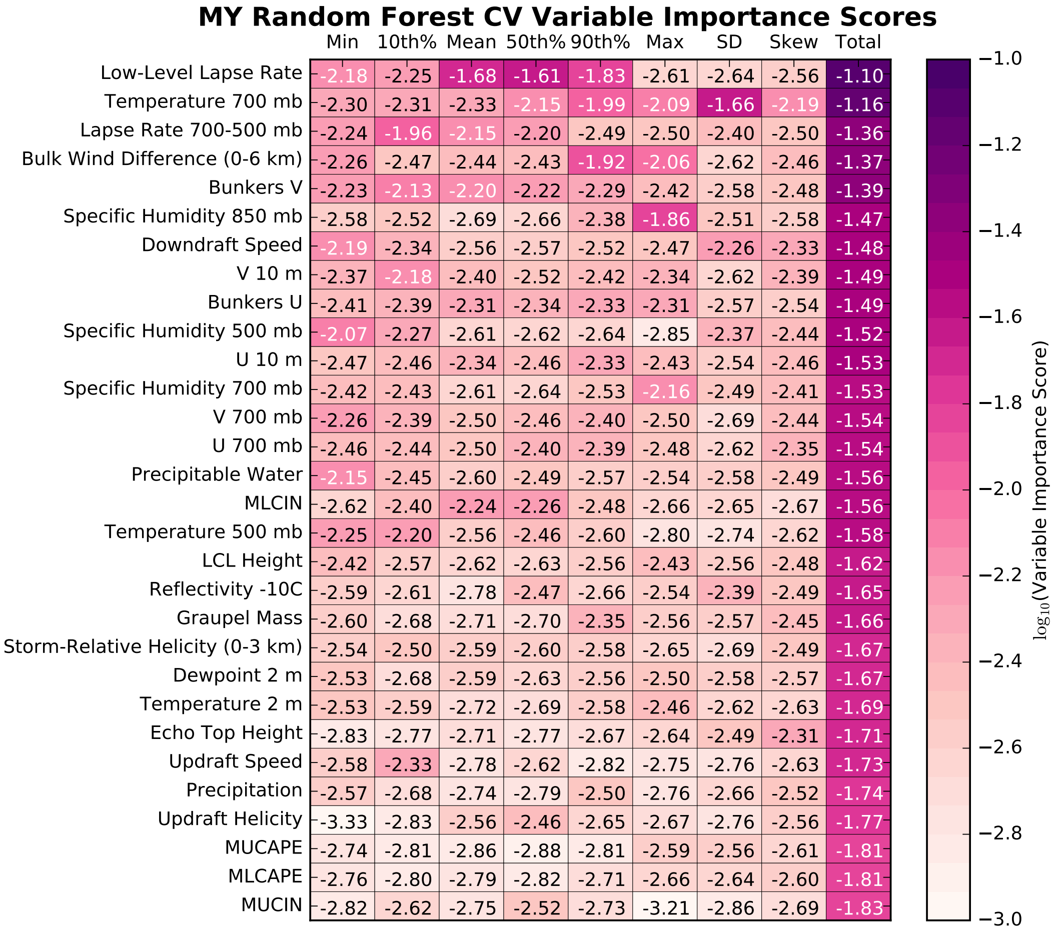
Figure 4.50: Same as Fig. 4.48 but for the size distribution random forest trained on Morrison microphysics members.



MY Random Forest Variable Importance Scores

	Min	10th%	Mean	50th%	90th%	Max	SD	Skew	Total
Downdraft Speed	-1.76	-1.71	-1.65	-1.79	-2.04	-2.38	-1.84	-2.68	-0.98
Temperature 700 mb	-2.14	-1.95	-1.72	-1.84	-1.64	-1.84	-2.53	-2.70	-1.03
Bulk Wind Difference (0-6 km)	-2.66	-2.41	-2.09	-2.11	-1.85	-1.85	-2.22	-2.70	-1.24
Lapse Rate 700-500 mb	-2.08	-2.03	-2.23	-2.21	-2.52	-2.66	-2.34	-2.70	-1.38
Bunkers V	-2.28	-2.24	-2.22	-2.15	-2.24	-2.33	-2.56	-2.64	-1.40
Temperature 500 mb	-2.21	-2.24	-2.32	-2.24	-2.35	-2.41	-2.34	-2.70	-1.43
Precipitable Water	-2.26	-2.22	-2.24	-2.23	-2.36	-2.40	-2.67	-2.63	-1.44
V 700 mb	-2.26	-2.26	-2.35	-2.31	-2.42	-2.49	-2.64	-2.64	-1.50
Updraft Speed	-2.69	-2.66	-2.52	-2.65	-2.25	-2.08	-2.23	-2.68	-1.50
Updraft Helicity	-3.56	-3.02	-2.56	-2.68	-2.29	-2.15	-2.21	-2.23	-1.52
Bunkers U	-2.53	-2.44	-2.35	-2.32	-2.37	-2.35	-2.68	-2.64	-1.54
LCL Height	-2.43	-2.44	-2.45	-2.43	-2.44	-2.44	-2.71	-2.71	-1.59
Specific Humidity 500 mb	-2.37	-2.44	-2.50	-2.59	-2.51	-2.60	-2.51	-2.63	-1.61
Reflectivity -10C	-2.68	-2.64	-2.43	-2.68	-2.45	-2.32	-2.71	-2.67	-1.64
MUCAPE	-2.71	-2.67	-2.56	-2.59	-2.46	-2.28	-2.64	-2.70	-1.65
Specific Humidity 850 mb	-2.69	-2.65	-2.52	-2.54	-2.44	-2.41	-2.66	-2.65	-1.65
Specific Humidity 700 mb	-2.55	-2.58	-2.67	-2.67	-2.57	-2.50	-2.39	-2.70	-1.66
MLCIN	-2.66	-2.57	-2.50	-2.34	-2.47	-2.80	-2.70	-2.73	-1.67
MLCAPE	-2.85	-2.68	-2.47	-2.51	-2.43	-2.44	-2.71	-2.68	-1.67
Temperature 2 m	-2.56	-2.56	-2.53	-2.56	-2.52	-2.52	-2.65	-2.73	-1.67
U 10 m	-2.59	-2.53	-2.51	-2.50	-2.57	-2.59	-2.71	-2.64	-1.67
U 700 mb	-2.64	-2.62	-2.57	-2.55	-2.52	-2.51	-2.57	-2.67	-1.68
V 10 m	-2.56	-2.52	-2.54	-2.51	-2.58	-2.61	-2.69	-2.67	-1.68
Dewpoint 2 m	-2.53	-2.54	-2.59	-2.56	-2.57	-2.62	-2.60	-2.72	-1.68
Storm-Relative Helicity (0-3 km)	-2.58	-2.54	-2.58	-2.58	-2.58	-2.57	-2.62	-2.69	-1.69
Graupel Mass	-2.55	-2.50	-2.53	-2.57	-2.71	-2.71	-2.62	-2.69	-1.70
Precipitation	-2.72	-2.72	-2.63	-2.65	-2.63	-2.58	-2.59	-2.59	-1.73
Echo Top Height	-3.20	-3.06	-2.74	-2.71	-2.37	-2.44	-2.70	-2.70	-1.77
Low-Level Lapse Rate	-2.59	-2.63	-2.67	-2.67	-2.71	-2.69	-2.72	-2.73	-1.77
MUCIN	-2.80	-2.78	-2.77	-2.72	-3.02	-3.42	-2.80	-2.74	-1.94

Figure 4.51: Same as Fig. 4.47 but for the hail occurrence random forest trained on Milbrandt and Yau (MY) microphysics members.



	Min	10th%	Mean	50th%	90th%	Max	SD	Skew	Total
Low-Level Lapse Rate	-2.18	-2.25	-1.68	-1.61	-1.83	-2.61	-2.64	-2.56	-1.10
Temperature 700 mb	-2.30	-2.31	-2.33	-2.15	-1.99	-2.09	-1.66	-2.19	-1.16
Lapse Rate 700-500 mb	-2.24	-1.96	-2.15	-2.20	-2.49	-2.50	-2.40	-2.50	-1.36
Bulk Wind Difference (0-6 km)	-2.26	-2.47	-2.44	-2.43	-1.92	-2.06	-2.62	-2.46	-1.37
Bunkers V	-2.23	-2.13	-2.20	-2.22	-2.29	-2.42	-2.58	-2.48	-1.39
Specific Humidity 850 mb	-2.58	-2.52	-2.69	-2.66	-2.38	-1.86	-2.51	-2.58	-1.47
Downdraft Speed	-2.19	-2.34	-2.56	-2.57	-2.52	-2.47	-2.26	-2.33	-1.48
V 10 m	-2.37	-2.18	-2.40	-2.52	-2.42	-2.34	-2.62	-2.39	-1.49
Bunkers U	-2.41	-2.39	-2.31	-2.34	-2.33	-2.31	-2.57	-2.54	-1.49
Specific Humidity 500 mb	-2.07	-2.27	-2.61	-2.62	-2.64	-2.85	-2.37	-2.44	-1.52
U 10 m	-2.47	-2.46	-2.34	-2.46	-2.33	-2.43	-2.54	-2.46	-1.53
Specific Humidity 700 mb	-2.42	-2.43	-2.61	-2.64	-2.53	-2.16	-2.49	-2.41	-1.53
V 700 mb	-2.26	-2.39	-2.50	-2.46	-2.40	-2.50	-2.69	-2.44	-1.54
U 700 mb	-2.46	-2.44	-2.50	-2.40	-2.39	-2.48	-2.62	-2.35	-1.54
Precipitable Water	-2.15	-2.45	-2.60	-2.49	-2.57	-2.54	-2.58	-2.49	-1.56
MLCIN	-2.62	-2.40	-2.24	-2.26	-2.48	-2.66	-2.65	-2.67	-1.56
Temperature 500 mb	-2.25	-2.20	-2.56	-2.46	-2.60	-2.80	-2.74	-2.62	-1.58
LCL Height	-2.42	-2.57	-2.62	-2.63	-2.56	-2.43	-2.56	-2.48	-1.62
Reflectivity -10C	-2.59	-2.61	-2.78	-2.47	-2.66	-2.54	-2.39	-2.49	-1.65
Graupel Mass	-2.60	-2.68	-2.71	-2.70	-2.35	-2.56	-2.57	-2.45	-1.66
Storm-Relative Helicity (0-3 km)	-2.54	-2.50	-2.59	-2.60	-2.58	-2.65	-2.69	-2.49	-1.67
Dewpoint 2 m	-2.53	-2.68	-2.59	-2.63	-2.56	-2.50	-2.58	-2.57	-1.67
Temperature 2 m	-2.53	-2.59	-2.72	-2.69	-2.58	-2.46	-2.62	-2.63	-1.69
Echo Top Height	-2.83	-2.77	-2.71	-2.77	-2.67	-2.64	-2.49	-2.31	-1.71
Updraft Speed	-2.58	-2.33	-2.78	-2.62	-2.82	-2.75	-2.76	-2.63	-1.73
Precipitation	-2.57	-2.68	-2.74	-2.79	-2.50	-2.76	-2.66	-2.52	-1.74
Updraft Helicity	-3.33	-2.83	-2.56	-2.46	-2.65	-2.67	-2.76	-2.56	-1.77
MUCAPE	-2.74	-2.81	-2.86	-2.88	-2.81	-2.59	-2.56	-2.61	-1.81
MLCAPE	-2.76	-2.80	-2.79	-2.82	-2.71	-2.66	-2.64	-2.60	-1.81
MUCIN	-2.82	-2.62	-2.75	-2.52	-2.73	-3.21	-2.86	-2.69	-1.83

Figure 4.52: Same as Fig. 4.48 but for the size distribution random forest trained on Milbrandt and Yau (MY) microphysics members.

4.3 Discussion

Machine learning object-based hail forecasts were evaluated and compared with other diagnostic hail forecasting methods to determine what value may be added to raw CAM ensemble output by these approaches. The different forecasting methods were evaluated using spring and summer 2015 runs of the CAPS and NCAR convection-allowing model ensembles and validated against gridded radar-estimated hail sizes. The machine learning models showed skill in discriminating between forecast storms that produced hail and those that did not but were underdispersive with the hail size forecasts. Additional calibration of the raw machine learning forecasts improved the sharpness of the size forecasts for the NCAR ensemble. For 24-hour hail outlooks, the machine learning methods were better able to identify hail threat areas while minimizing false alarms compared with HAILCAST and the Thompson hail size estimation method. Of the existing storm-surrogates, updraft helicity provided the best indicator of large hail. Based on analysis of variable importance rankings from the machine learning models, the hail forecasts were closely tied to lapse rates near the freezing level, wind shear, and the saturation of the air near the surface. While storm object identification helped constrain where conditions were favorable for hail, the actual storm surrogate values appeared to have little utility in discriminating hail occurrence and hail size. Environmental parameters were more important for that task.

The skill of machine learning hail forecasts generated from CAM ensembles is more sensitive to the choices made during pre-processing than to the choice of machine learning model. Having a training set that is as close in configuration as possible to what is being used operationally is also very important for skillful forecasts. While the 2014 CAPS ensemble dataset provided somewhat skillful forecasts of hail occurrence in 2015, the size forecasts had less ability to discriminate between large and small hail events. The NCAR ensemble configuration, which used the same configuration for all ensemble members and runs, could distinguish a wide range of hail sizes and probabilities. The NCAR ensemble algorithms benefited from having a larger amount of training data due to the longer training period and the ability to aggregate across 10 diverse but similarly configured ensemble members. As the 27 May 2015 example shows, even with only 2 weeks of training data, the machine learning models could still discriminate hail size and location really well. Constraining the hail forecasts to areas where the model produces graupel and to within the US borders results in a less noisy relationship between forecast parameters and observed hail.

The verification statistics, spatial maps, and case studies showed that the hail forecasting methods exhibit more pronounced forecasting biases in particular areas. The machine learning models and updraft helicity work very well for discriminating hail in Plains supercells, particularly storms in the High Plains and just east of the Rockies. Both of these methods take shear and updraft intensity into account. Column Total Graupel and HAILCAST are more influenced by CAPE and less by shear, so they produce hail in any storm that has a strong updraft. Column Total Graupel does not account for melting at all, and HAILCAST uses a bulk melting parameter, so they do not fully account for the melting that occurs with storms in the Southeast and tend to overpredict hail in that area. The machine learning model uses the LCL height to account for relative humidity effects on melting and gives lower probabilities to areas with low LCL heights, but it is too aggressive with lowering probabilities in the Southeast and misses too many storms. Using a lower decision threshold may help capture some of these cases that are currently missed with the 50% decision threshold. Alternatively, regionally calibrated decision thresholds may be useful to capture different storm environments more accurately.

While constraining machine learning hail forecasts to areas where the NWP model produces storms does generally produce better forecasts, there are situations in which this approach will struggle compared with ingredients- or parameter-based methods. If the CAM struggles with the placement, timing, and evolution of storms, then the hail forecasts dependent on those storms will also struggle. These struggles tend to occur in scenarios where large scale forcing is weaker, leading to convection initiation and evolution being governed by poorly observed mesoscale effects. When an area receives multiple days of convection, errors in forecasting the diurnal cycle of convection lead to additional spatial and temporal biases. Hail forecasts based solely on environmental parameters will tend to have better coverage in these situations and will detect hail threat areas that storm-based methods may miss at the expense of more false alarms.

Chapter 5

An Evaluation of Statistical Learning Configurations for Gridded Solar Irradiance Forecasting

Solar-based electricity generation and its share of the power supply has been growing rapidly over the past decade (Shaker et al. 2016). As solar power achieves higher penetration and becomes more critical to the electric infrastructure, the need for accurate forecasts of solar irradiance and solar power increases greatly in order to maintain a steady electricity supply under varying weather conditions (Renné 2014). Current state-of-the-art solar and wind energy forecast systems combine Numerical Weather Prediction (NWP) model output with statistical learning models trained on a historical archive of observed solar irradiance or power output to produce a forecast with minimal bias. This approach is very effective for sites that have been operating for a long period of time, but with new large solar plants coming online more frequently and more people investing in residential rooftop solar panels, accurate solar predictions are needed for larger areas where observing sites either have very short records or are not available at all.

Generating the most accurate predictions at sites without observations requires fusing many static and dynamic data sources together within a statistical learning framework. The amount of solar irradiance at the surface is primarily driven by the position of the sun in the sky as well as the amount and type of aerosols and clouds scattering the sunlight. Obstructions by terrain, buildings, and trees can also impact solar irradiance at lower sun angles. Solar position can be directly calculated given a location and time, and information about terrain and land cover type is available from high resolution gridded datasets. Cloud cover and aerosol information can be extracted from NWP model output, but operational NWP models generally do not represent either very well and may be subject to other systematic biases (Diagne et al. 2013). Statistical learning models can determine cloudiness from other NWP model conditions associated with observed cloudiness and can make corrections based on data sources unavailable to a NWP model, including climatological information and statistics concerning spatial and temporal variability.

Current operational statistical gridded forecasting systems use linear bias correction methods to calibrate raw model output to either observations or analyses and then interpolate those corrections to a fine grid. The National Weather Service Gridded Model Output Statistics (MOS) system (Glahn et al. 2009) performs linear regression corrections at each observation site and then uses the Cressman (1959) successive correction method and an elevation correction to interpolate the site-based MOS forecasts to a grid. The Australian Bureau of Meteorology, which has to account for a sparse observation network across most of the country, performs a bias correction of model output on a coarse grid and then builds a weighted consensus that is statistically downscaled to a fine grid (Engel and Ebert 2012).

The purpose of this chapter is to evaluate different statistical learning models and configurations for gridded solar irradiance forecasting. The primary hypothesis is that ensemble decision tree methods produce more accurate gridded solar irradiance forecasts than linear regression and raw NWP model output. In the pre-processing stage, the set of input variables, NWP model configuration, and division of training data are investigated. Multiple types of machine learning models, as well as different configurations of those models, are all evaluated to see which parameter choices impact performance. Finally, different methods for applying the calibrated machine learning models to unknown sites are compared.

5.1 Methods

5.1.1 Data and Pre-Processing

Observed solar irradiance data come from the Oklahoma Mesonet (McPherson et al. 2007). The Mesonet reports the 5-minute-averaged global horizontal irradiance (GHI) every 5 minutes using Li-Cor LI-200 silicon photodiode-type pyranometers. The instruments are regularly calibrated and are monitored by both humans and automated algorithms for quality assurance. Extraterrestrial solar radiation and solar position angles are calculated using the PVLIB Python library (Holmgren et al. 2015). The solar position calculations are performed using a Python implementation of the National Renewable Energy Laboratory (NREL) Solar Position Algorithm (SPA) (Reda and Andreas 2003). Solar zenith (θ_s), elevation, and azimuth angles are calculated every 5 minutes and are used to estimate the idealized clear-sky irradiance at the top of the atmosphere I_{toa}. The clearness index K_t is calculated from the Mesonet solar irradiance I_s using Eq. 5.1.

$$K_t = \frac{I_s}{I_{toa}\cos\theta_s} \tag{5.1}$$

The 5-minute irradiance and clearness index values are then averaged over the previous hour to determine the hourly-averaged values. The hourly-averaged

K_t is then used as truth for the statistical learning model experiments. Times without sun or with data outages are dropped from the dataset.

The first set of statistical learning model experiments is performed with the NOAA National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) model. The GFS is a global spectral model run operationally by NCEP four times a day out to 16 days. The raw GFS model output is interpolated onto an approximately 4 km grid that uses uniform latitude and longitude values over the contiguous United States. Temporal linear interpolation from the 3-hourly data to hourly values was also performed. Incoming hourly averaged downward short wave radiation at the surface, total cloud cover percentage, and surface temperature are extracted from the 00 UTC runs for the period from 5 June through 30 August 2015. Forecast hours 14 through 24 were used for the analysis. All of the input variables to the GFS machine learning models are listed in the first column of Table 5.1.

A second set of experiments was performed with the Center for Analysis and Prediction of Storms (CAPS) 2016 3DVAR-based Storm-Scale Ensemble Forecast system, referred to as the 2016 CAPS Ensemble from here. The CAPS Ensemble consists of 18 Weather Research and Forecasting (WRF) model members running the Advanced Research WRF (ARW) dynamical core with perturbed initial conditions, lateral boundary conditions, microphysics schemes, and planetary boundary layer (PBL) schemes. The ensemble was run every weekday at 00 UTC from 2 May through 3 June 2016 as part of the NOAA Hazardous Weather Testbed Spring Experiment. Downward shortwave and total irradiance at the surface; relative humidity at 850, 700, and 500 mb; precipitable water; total precipitation; composite reflectivity; and surface height were extracted at

Table 5.1: Input variables for the GFS and 2016 CAPS Ensemble control member machine learning models. Spatial statistics were calculated over a 3 grid point radius for the GFS and a 10 grid point neighborhood radius for the CAPS Ensemble.

GFS	CAPS Ensemble
Valid Hour CST	Forecast Hour
Forecast Hour	GHI at nearest grid cell
Day of Year	GHI Spatial Mean
Sine (Day of Year)	GHI Spatial Max
GHI at nearest grid cell	GHI Spatial Min
GHI Spatial Mean	GHI Spatial SD
GHI Spatial Max	GHI Spatial Skewness
GHI Spatial Min	GHI Spatial Kurtosis
GHI Spatial Correlation	Total Downward Irradiance
Temperature at nearest grid cell	Precipitable Water
Temperature Spatial Mean	Composite Reflectivity
Temperature Spatial Max	Terrain Height
Temperature Spatial Min	850 mb Relative Humidity
Temperature Spatial Correlation	700 mb Relative Humidity
Cloud Cover at nearest grid cell	500 mb Relative Humidity
Cloud Cover Spatial Mean	Precipitation
Cloud Cover Spatial Max	Clear Sky Irradiance
Cloud Cover Spatial Min	Solar Zenith Angle
Cloud Cover Correlation	Solar Azimuth Angle
Solar Zenith Angle	
Solar Azimuth Angle	
Solar Elevation Angle	
Forecast Clearness Index	
Forecast Clear Sky Irradiance	

each Mesonet site. The mean, minimum, maximum, standard deviation, skewness, and kurtosis of the downward shortwave irradiance at grid points within an 30 km box around each site were also extracted for forecast hours 11 through 26. The full list of CAPS ensemble input variables is in the second column of Table 5.1. Hours where the sun was below the horizon were excluded. This analysis focuses on the control member of the ensemble, which uses the Thompson microphysics scheme (Thompson et al. 2008) and Mellor-Yamada-Janjic (MYJ) PBL scheme (Mellor and Yamada 1982).

5.1.2 Statistical Learning Models

Both the statistical learning model type and parameter settings are evaluated to determine their relative impact on forecast performance. Different statistical learning models from the scikit-learn (Pedregosa et al. 2011) Python library and some of their configurations are evaluated. Lasso (Tibshirani 1996), a regularized linear regression model with sparse coefficients, is used as a baseline method. Random forests (Breiman 2001a), an ensemble of randomized decision trees, and Gradient Boosted Regression Trees (Friedman 2001), a stagewise, additive weighted ensemble of decision trees, are also used. The data processing and machine learning modeling procedure is summarized in Fig. 5.1.

In the GFS experiment, the default random forest has 500 trees, a minimum number of samples at a split node of 10, and the square root of the total number of variables sampled at each split node. The “Random Forest Short Trees” model uses a maximum depth of 3 for any branch. The depth of the trees affects how closely each tree fits to individual training cases. A shorter depth should result in smoother predictions throughout the feature space, but the

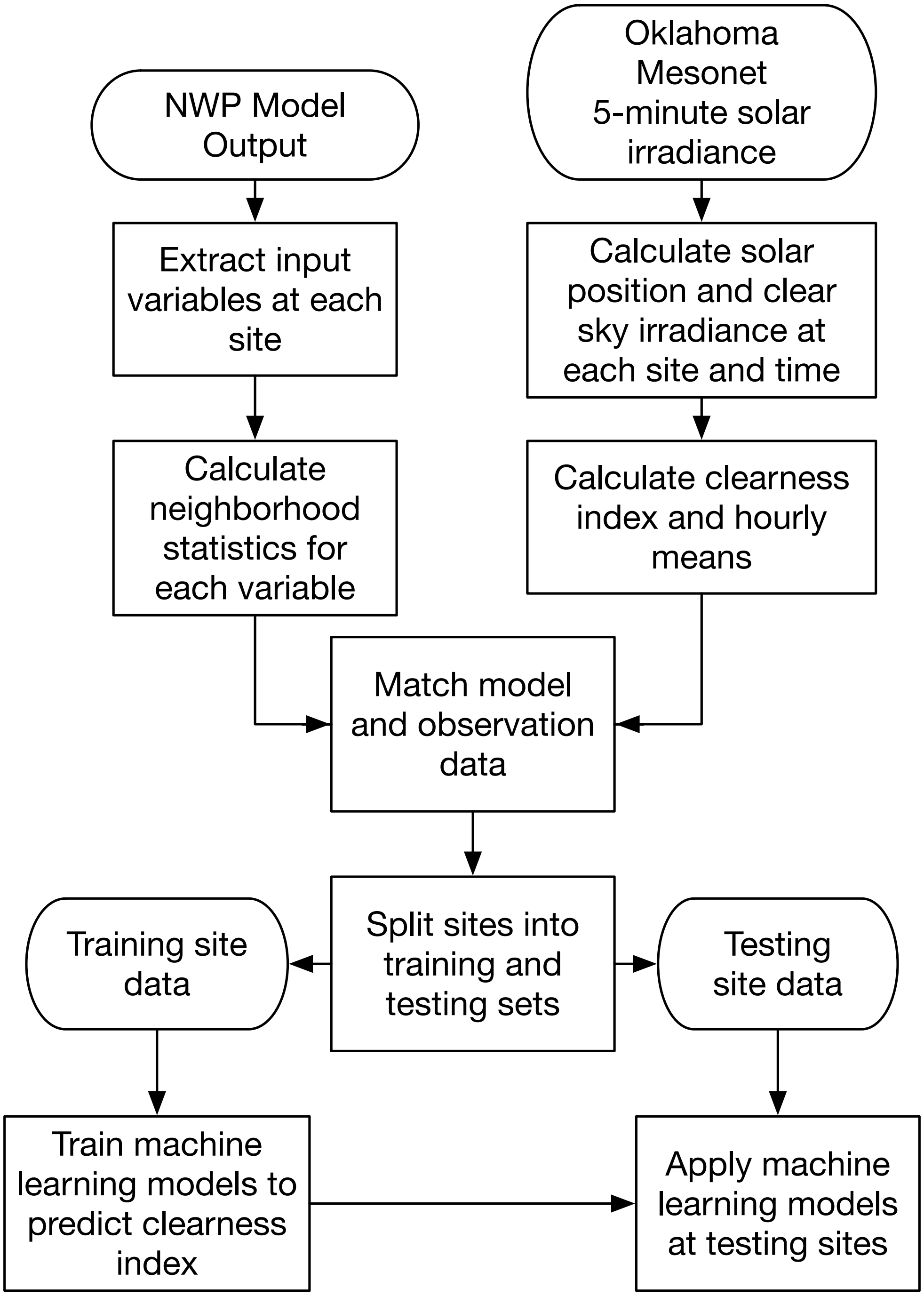

NWP Model Output
Oklahoma Mesonet 5-minute solar irradiance
Extract input variables at each site
Calculate solar position and clear sky irradiance at each site and time
Calculate neighborhood statistics for each variable
Calculate clearness index and hourly means
Match model and observation data
Split sites into training and testing sets
Training site data
Testing site data
Train machine learning models to predict clearness index
Apply machine learning models at testing sites

Figure 5.1: Summary of the procedure for data pre-processing, machine learning model training, and machine learning model application for solar irradiance forecasts.

predictions will also be less sharp. The "Random Forest All Features" model evaluates all input features at each split node instead of a random subset. This change should result in less variability and independence among the individual trees compared with sampling a small number of features.

The default Gradient Boosting model uses 500 trees, a learning rate of 0.1, a least absolute deviance loss function, a max depth of 5, and samples the square root of the number of input variables. The "Gradient Boosting Least Squares" model uses the least squares loss function instead of least absolute deviance and will weigh large errors more heavily. The "Gradient Boosting Big Trees" model uses trees that are grown to split nodes with a minimum number of training samples of 10 instead of a max depth of 5. The "Gradient Boosting "All Features" model samples all features at each node. The "Gradient Boosting Slow Learning Rate" model reduces the learning rate to 0.01. A slower learning rate reduces the contribution of each tree to the ensemble, so more trees are required to reach the same training set error, but the model may be able to optimize predictions more than with a higher learning rate. The Linear Regression model uses a Lasso regression that fits to the 16 top features selected with the highest F-scores and an alpha of 0.5.

The CAPS ensemble experiment tests two ways of aggregating the predictions from the individual trees in the random forest. The default Random Forest for the CAPS ensemble experiment uses 500 trees, minimum samples at the leaf nodes of 1, and samples the square root of the total number of input variables. For regression models, the mean of the tree prediction distribution is used based on the central limit theorem assumption that a large enough set of independent samples will form a Gaussian distribution no matter the identity of the original distribution. In practice, the distribution of individual tree predictions may be multimodal, particularly if there is high uncertainty and the predicted quantity does not have a Gaussian distribution. In some circumstances, such as when there is a larger number of outlier predictions, the median may be a better choice for the consensus than the mean. For bimodal distributions, the mean and median will tend to occur between the two peaks. Because the random forest is composed of a set of randomized, or weaker decision trees, the ensemble spread tends to be very high, so a direct estimation of quantiles may not perform well. The default Gradient Boosting model uses 200 trees, the least absolute deviance loss function, a maximum of 100 leaf nodes, subsamples the 80% of the training data randomly for each tree, and has a learning rate of 0.05. Another Gradient Boosting model uses the Huber loss function (Friedman 2001), a piecewise combination of mean squared error and mean absolute error connected at a split point δ as shown in Eq. 5.2,

$$L(y,F) = \begin{cases} \frac{1}{2}(y-F)^2 & |y-F| \leq \delta \\ \delta(|y-F| - \frac{\delta}{2}) & |y-F| > \delta \end{cases} \tag{5.2}$$

to determine if that produces a more physically realistic forecast distribution by altering the evaluation of the feature splits.

5.1.3 Gridded Forecast Evaluation Procedure

Generating calibrated gridded solar irradiance forecasts requires determining the best estimate of irradiance at a location where irradiance is not observed. In order to simulate this condition and still score the different procedures, half of the 120 Oklahoma Mesonet sites are selected at random as training sites, and the other half are used as testing sites. In addition, testing days are withheld during the period of the experiment to prevent temporal contamination of the training data. Every third day during the training period is used as a testing day to reduce the impact of seasonality bias on the evaluation.

Two procedures are tested for generalized gridded forecasting. In the "Single Site" approach, separate statistical learning models are trained at each training site, predictions are made at each of these sites, and then the predictions are interpolated to the testing site using the Cressman (1959) successive correction interpolation method. For each interpolation point f_i, a distance-weighted average of the predictions at the stations with distances d_j within a radius of influence R is computed such that

$$f_i = \frac{\sum_j^J w_j f_{sj}}{\sum_{j=1}^J w_j}; \quad \begin{array}{ll} w_j = \frac{R^2 - d_j^2}{R^2 + d_j^2} & R < d \\ w_j = 0 & R \geq d \end{array} \tag{5.3}$$

. The test sites were initialized with the mean of the predictions at the training sites, and then four passes were performed with the Cressman filter with a decrease in radius for each pass to capture local effects. The Cressman interpolation method was chosen because the NWS gridded MOS system (Glahn et al. 2009) also uses it for interpolation from training sites to grid points.

In the "Multi Site" approach, the data from all training sites are aggregated together and are used to train one statistical learning model. This model is then applied at the testing sites using the NWP model and clear sky model output at that location. This approach requires training a single statistical model and can thus utilize a larger training set than the Single Site method. Applying a single model at each grid point also eliminates discontinuities that may be found in approaches that use separate statistical models for different regions. This approach is less able to correct for local biases and conditions.

Forecasts from each machine learning model are evaluated based on their accuracy, systemic bias, reliability, discrimination, and sharpness. Forecast accuracy is assessed using the mean absolute error, which is less sensitive to outlier errors. Systemic bias is evaluated through the mean error and determines if the models tend to over or underforecast clearness index. The reliability, or condition bias, of the forecasts is assessed through a reliability diagram in which the forecasts are binned, and the mean observed value is calculated for each bin. Reliable forecasts should have similar average observed values for a given forecast value. Discrimination is assessed by binning observations and calculating the average forecast value for a given observation. If the models show good discrimination, then observations of higher clearness index should have a higher forecast clearness index on average versus observations of lower clearness index. Sharpness is assessed by examining the distribution of forecasts and comparing them with the distribution of observations.

The statistical significance of the verification scores is assessed with bootstrap confidence intervals and permutation tests. A bootstrap replicate size of 10000 is used. Independent bootstrap confidence intervals are used to assess the uncertainty of the verification scores. Ranking the models by score in each bootstrap replicate and counting the frequency of each ranking for each model is used to assess whether certain models consistency outperform others and how much that ranking varies. Permutation tests are finally used to assess whether the difference in scores between two models is statistically significant and what the p-value of that difference is. A global p-value of 0.05 is used to determine statistical significance and is adjusted based on the false discovery rate correction to account for multiple comparisons (Benjamini and Hochberg 1995).

5.2 Results

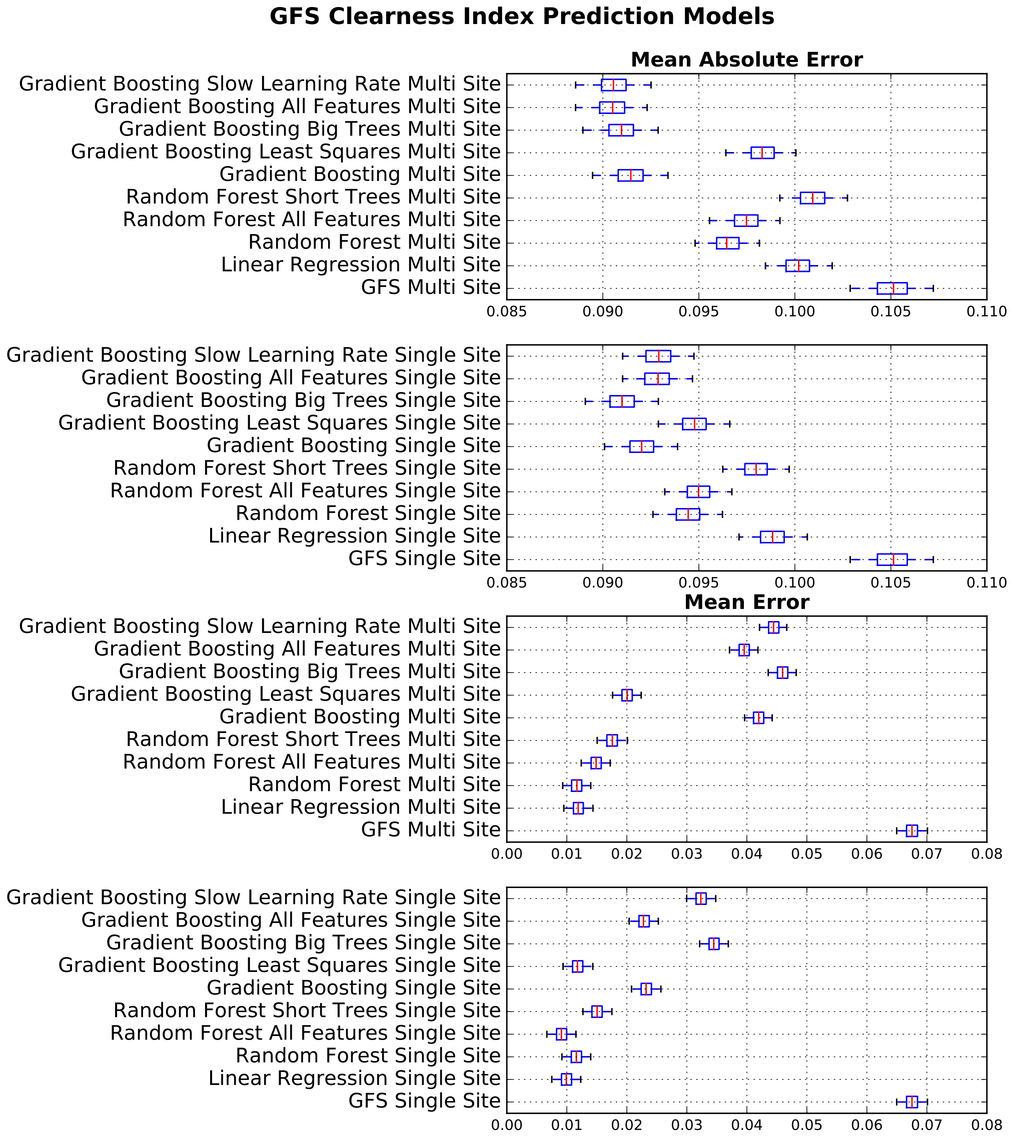
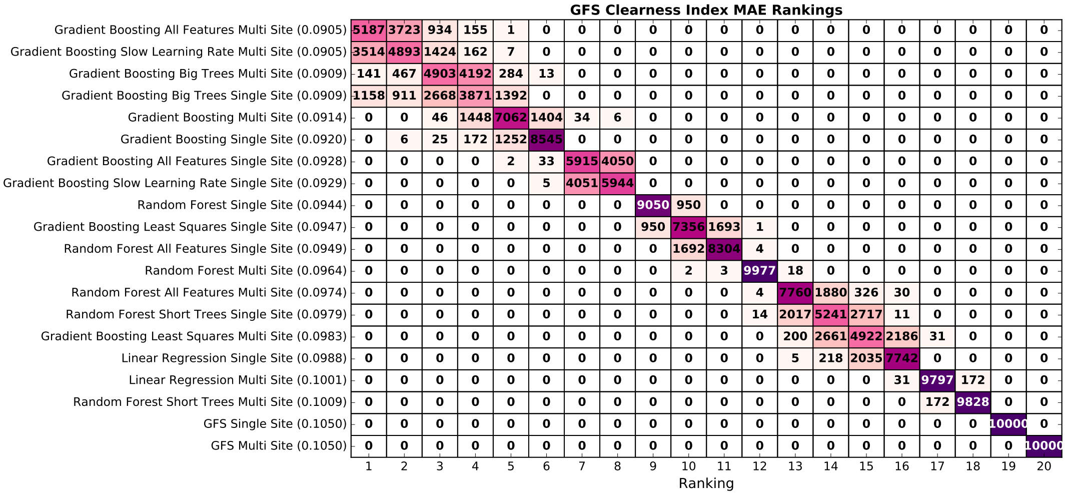


Figure 5.2: Bootstrap confidence intervals for the mean absolute errors and mean errors for each statistical learning model configuration.

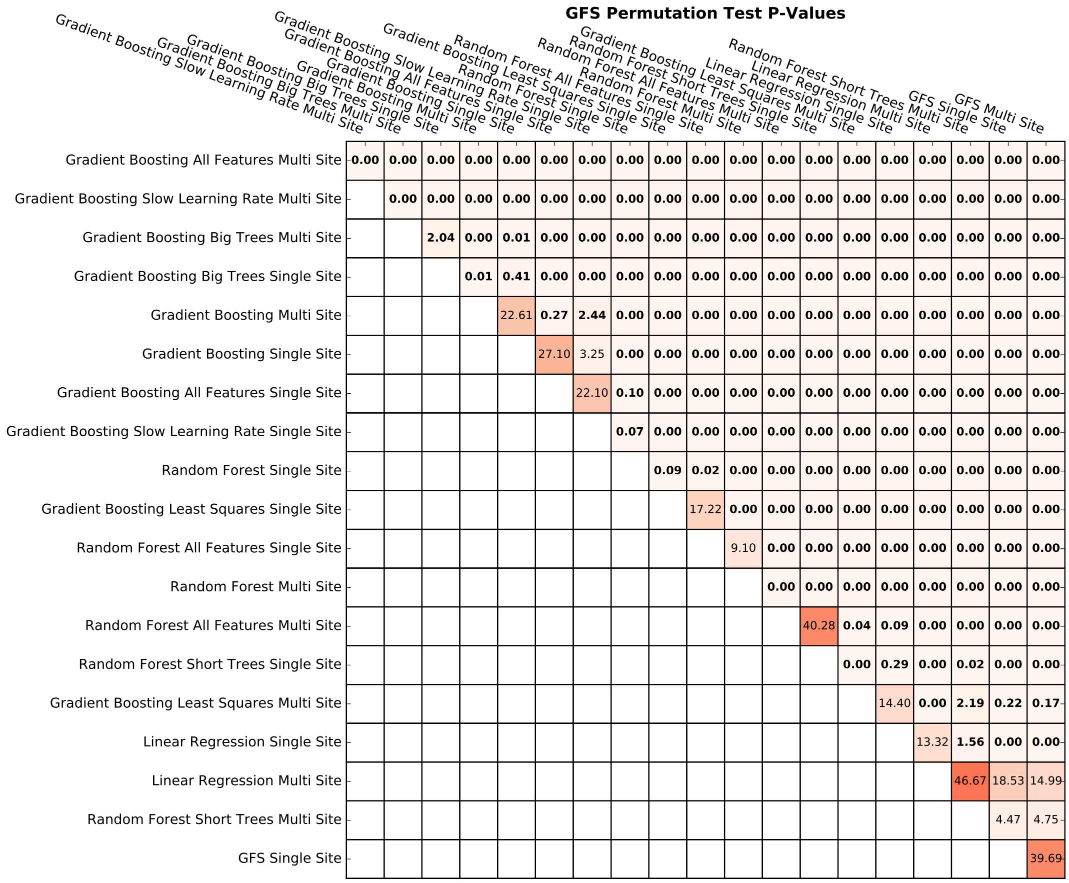
5.2.1 Experiment 1: GFS

Fig. 5.2 shows the bootstrap confidence intervals of the mean absolute error and mean error for each machine learning model. The Single Site models perform slightly better than the Multi Site models with the exception of the top-performing gradient boosting models. The Multi Site Gradient Boosting models with a mean absolute error loss function generally have the lowest mean absolute error but have higher mean error than the other Multi Site models. Of the experimental variations made to the gradient boosting model for this experiment, changing the loss function from mean absolute error to mean squared error had the largest impact on performance. For random forest, changing the depth of the trees had a bigger impact than expanding the number of features evaluated. The linear regression model does exhibit similar errors to some of the configurations of random forest and gradient boosting, suggesting that poor configuration choices can lead to worse performance than simpler models.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Gradient Boosting All Features Multi Site (0.0905)	5187	3723	934	155	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Slow Learning Rate Multi Site (0.0905)	3514	4893	1424	162	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Big Trees Multi Site (0.0909)	141	467	4903	4192	284	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Big Trees Single Site (0.0909)	1158	911	2668	3871	1392	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Multi Site (0.0914)	0	0	46	1448	7062	1404	34	6	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Single Site (0.0920)	0	6	25	172	1252	8545	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting All Features Single Site (0.0928)	0	0	0	0	2	33	5915	4050	0	0	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Slow Learning Rate Single Site (0.0929)	0	0	0	0	0	5	4051	5944	0	0	0	0	0	0	0	0	0	0	0	0
Random Forest Single Site (0.0944)	0	0	0	0	0	0	0	0	9050	950	0	0	0	0	0	0	0	0	0	0
Gradient Boosting Least Squares Single Site (0.0947)	0	0	0	0	0	0	0	0	950	7356	1693	1	0	0	0	0	0	0	0	0
Random Forest All Features Single Site (0.0949)	0	0	0	0	0	0	0	0	0	1692	8304	4	0	0	0	0	0	0	0	0
Random Forest Multi Site (0.0964)	0	0	0	0	0	0	0	0	0	2	3	9977	18	0	0	0	0	0	0	0
Random Forest All Features Multi Site (0.0974)	0	0	0	0	0	0	0	0	0	0	0	4	7760	1880	326	30	0	0	0	0
Random Forest Short Trees Single Site (0.0979)	0	0	0	0	0	0	0	0	0	0	0	14	2017	5241	2717	11	0	0	0	0
Gradient Boosting Least Squares Multi Site (0.0983)	0	0	0	0	0	0	0	0	0	0	0	0	200	2661	4922	2186	31	0	0	0
Linear Regression Single Site (0.0988)	0	0	0	0	0	0	0	0	0	0	0	0	5	218	2035	7742	0	0	0	0
Linear Regression Multi Site (0.1001)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	9797	172	0	0
Random Forest Short Trees Multi Site (0.1009)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	172	9828	0	0
GFS Single Site (0.1050)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000	0
GFS Multi Site (0.1050)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000

Figure 5.3: Bootstrap rankings for the mean absolute errors for each statistical learning model configuration. The mean absolute error for each machine learning model is listed in parentheses.



GFS Permutation Test P-Values

	Gradient Boosting Slow Learning Rate Multi Site	Gradient Boosting Big Trees Multi Site	Gradient Boosting Big Trees Single Site	Gradient Boosting Multi Site	Gradient Boosting Single Site	Gradient Boosting All Features Single Site	Gradient Boosting Slow Learning Rate Single Site	Random Forest Single Site	Gradient Boosting Least Squares Single Site	Random Forest All Features Single Site	Random Forest Multi Site	Random Forest All Features Multi Site	Random Forest Short Trees Single Site	Gradient Boosting Least Squares Multi Site	Linear Regression Single Site	Linear Regression Multi Site	Random Forest Short Trees Multi Site	GFS Single Site	GFS Multi Site
Gradient Boosting All Features Multi Site	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Slow Learning Rate Multi Site		**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Big Trees Multi Site			**2.04**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Big Trees Single Site				**0.01**	**0.41**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Multi Site					22.61	**0.27**	**2.44**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Single Site						27.10	3.25	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting All Features Single Site							22.10	**0.10**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Slow Learning Rate Single Site								**0.07**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest Single Site									**0.09**	**0.02**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Least Squares Single Site										17.22	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest All Features Single Site											9.10	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest Multi Site												**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest All Features Multi Site													40.28	**0.04**	**0.09**	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest Short Trees Single Site														**0.00**	**0.29**	**0.00**	**0.02**	**0.00**	**0.00**
Gradient Boosting Least Squares Multi Site															14.40	**0.00**	**2.19**	**0.22**	**0.17**
Linear Regression Single Site																13.32	**1.56**	**0.00**	**0.00**
Linear Regression Multi Site																	46.67	18.53	14.99
Random Forest Short Trees Multi Site																		4.47	4.75
GFS Single Site																			39.69

Figure 5.4: Permutation test p-values (multiplied by 1000 for the comparison of the differences in mean absolute error between each machine learning model trained on GFS output. P-values in bold are statistically significant based on the false discovery method with a rate of 0.05 (α=0.0325). Darker reds are associated with larger p-values.

The bootstrap rankings of each machine learning model by mean absolute error are shown in Fig. 5.3. Most of the variations on Gradient Boosting Multi Site have overlapping rankings among the top models. Both the Single and Multi Site Gradient Boosting Big Trees models show more variance in their rankings than other models, which may be a product of the increased variance in predictions produced by using larger trees. The Gradient Boosting models using the least absolute deviance loss function do not overlap any of the other models in their rank intervals. The Single Site Random Forests rank better than the Multi Site Random Forests and show no overlap with them. Single Site Linear Regression overlaps in rankings with Multi Site Random Forest and Least Squares Gradient Boosting. The permutation test p-values for the differences in mean absolute error are shown in Fig. 5.4. Using Gradient Boosting models with all input features, a slower learning rate, and larger trees results in statistically significant improvements over the default approach. Single Site and Multi Site Gradient Boosting with default parameters do not have significantly different forecast errors. For random forest, the difference between the Single Site and Multi Site model is statistically significant.

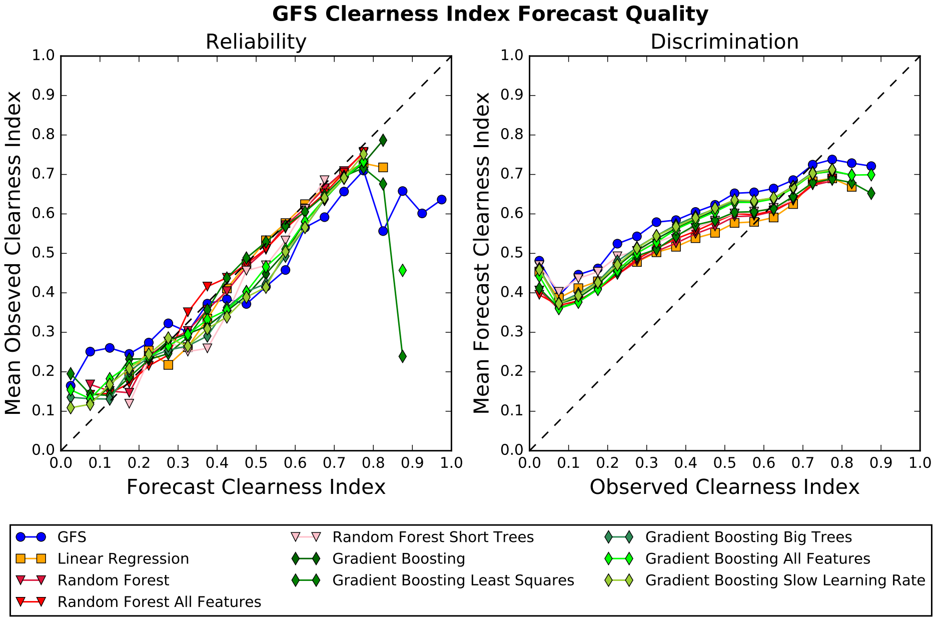


Figure 5.5: Binned mean forecast and observed clearness index values from each model.

The ability of each model to produce reliable clearness index values and discriminate between high and low clearness index cases is shown in Fig. 5.5. All of the models show good reliability although the Gradient Boosting models exhibit a consistent small overforecasting bias. The random forest models are closer to the perfect reliability line. In terms of discrimination, all of the models in general forecast lower values for cloudier hours and higher values for clearer hours, but the models tend to overforecast the clearness index when it is less than 0.5. The machine learning models marginally improve on the GFS but still follow its trends closely.

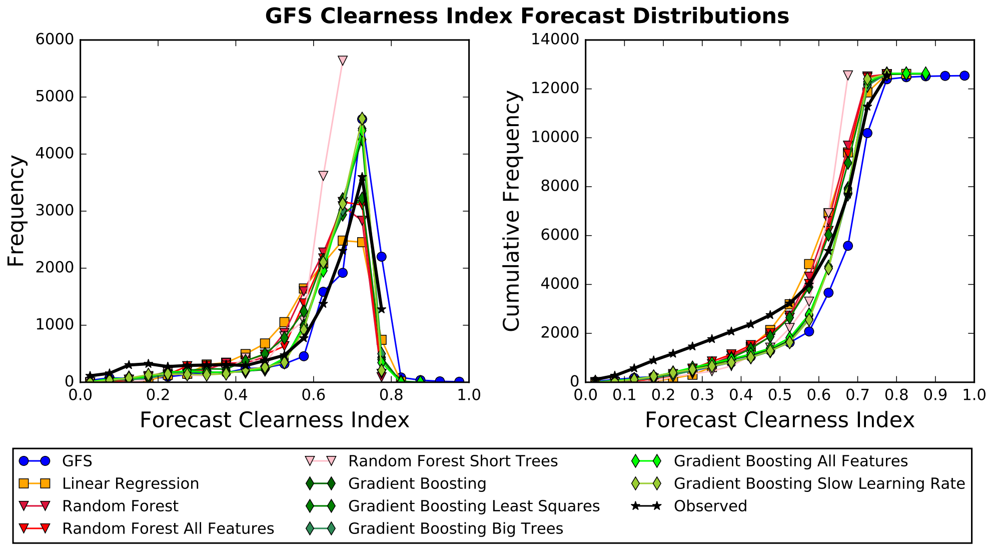


Figure 5.6: Point and cumulative distributions of Multi Site GFS clearness index forecasts and observations.

The biases in the forecast distributions are shown in Fig. 5.6. Compared with the distribution of observations, all models underforecast the occurrence of cloudy days and overforecast clear days, but they do generally capture the correct shape of the distribution. The binned forecast errors in Fig. 5.7 and 5.8 indicate that the largest errors occur under partly cloudy conditions. For all models, mean absolute error peaks between 0.3 and 0.6. For Multi Site models, the peak errors for Gradient Boosting are slightly higher than for the other models (Fig. 5.7). The biggest gains in performance from the raw GFS forecasts occur at small values of clearness index. High errors from the GFS and

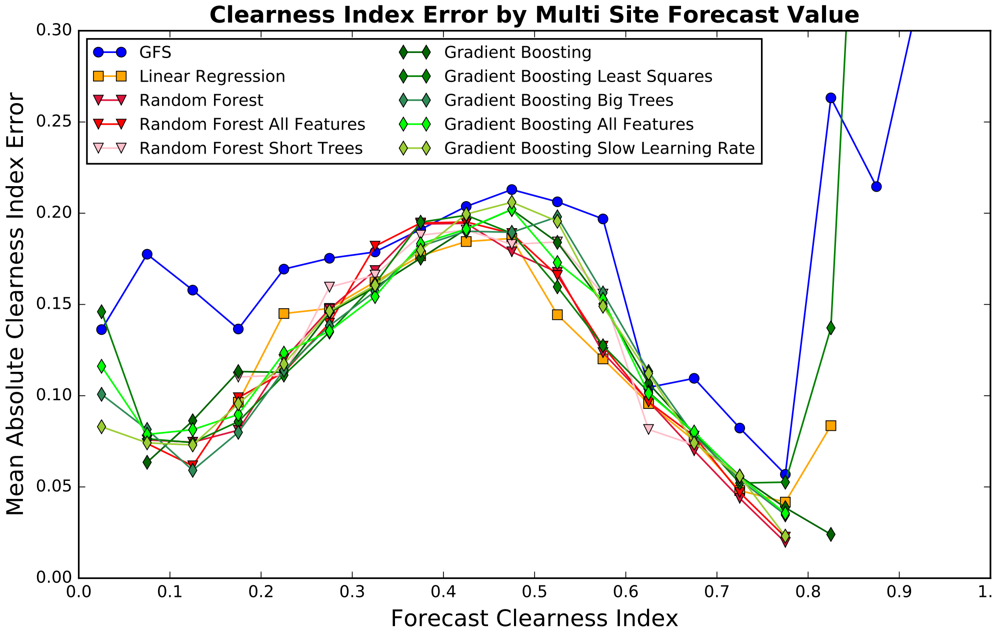


Figure 5.7: Mean absolute error by forecast bin for the Multi Site models.

Gradient Boosting Least Squares models at high forecast clearness index values are due to the models producing irradiances higher than observed near sunrise and sunset where sensitivity to solar angle is greater. The Single Site models show less of a error decrease at small clearness index values, but otherwise all the models show a similar error trajectory (Fig. 5.8).

The mean absolute errors by test site (Fig. 5.9) show some geographic trends in the errors that are fairly consistent across model choice. The highest errors are for the Northeast Oklahoma and the far western Panhandle sites. South central Oklahoma sites generally have the lowest error. Gradient boosting provides the biggest improvements in error in the areas where most of the training sites are located, and the decreases in error are less pronounced elsewhere. The biggest differences in mean error at each site (Fig. 5.10) can be found between

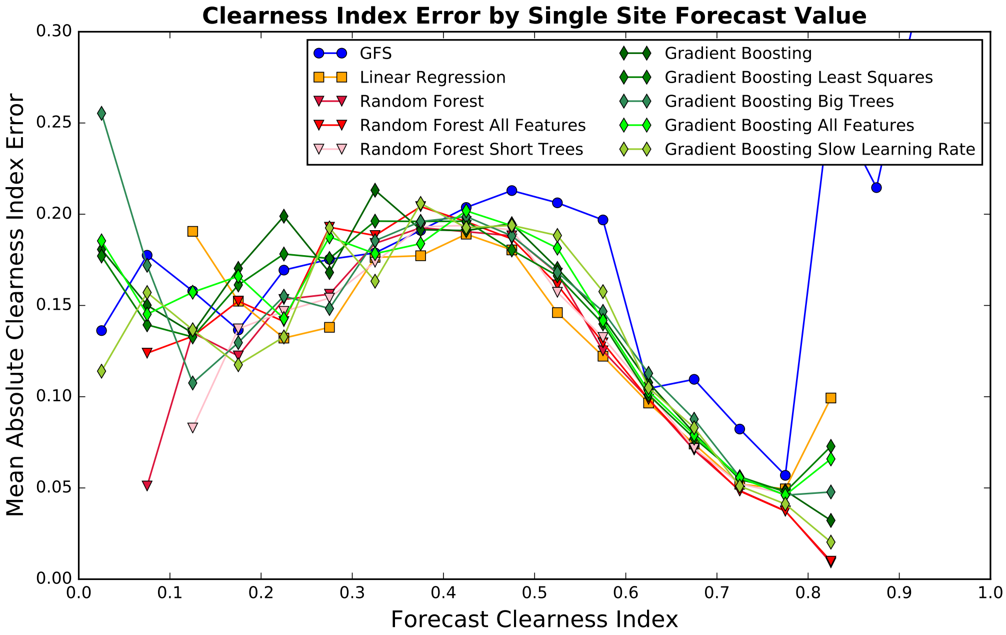


Figure 5.8: Mean absolute error by forecast bin for the Single Site models.

the Multi-Site Gradient Boosting model and everything else. That model features a large positive bias everywhere except the Oklahoma Panhandle, which has a near zero bias while the other models have negative biases. The biggest contributor to the amount of error at a particular site appears to be the amount of cloudiness that occurs at each site (Fig. 5.11). There is a fairly strong correlation between station mean absolute error and the percentage of observations with a clearness index above 0.6. The high outlier points are likely sites in the Oklahoma panhandle that have higher error due to their distance from most of the training sites. Those sites have lower error with the Single Site approach, so forecasts at sites that are poorly represented in the Multi Site training data would benefit more from the Single Site approach. An optimized hybrid of the Multi Site and Single Site approaches could try to take this effect into account

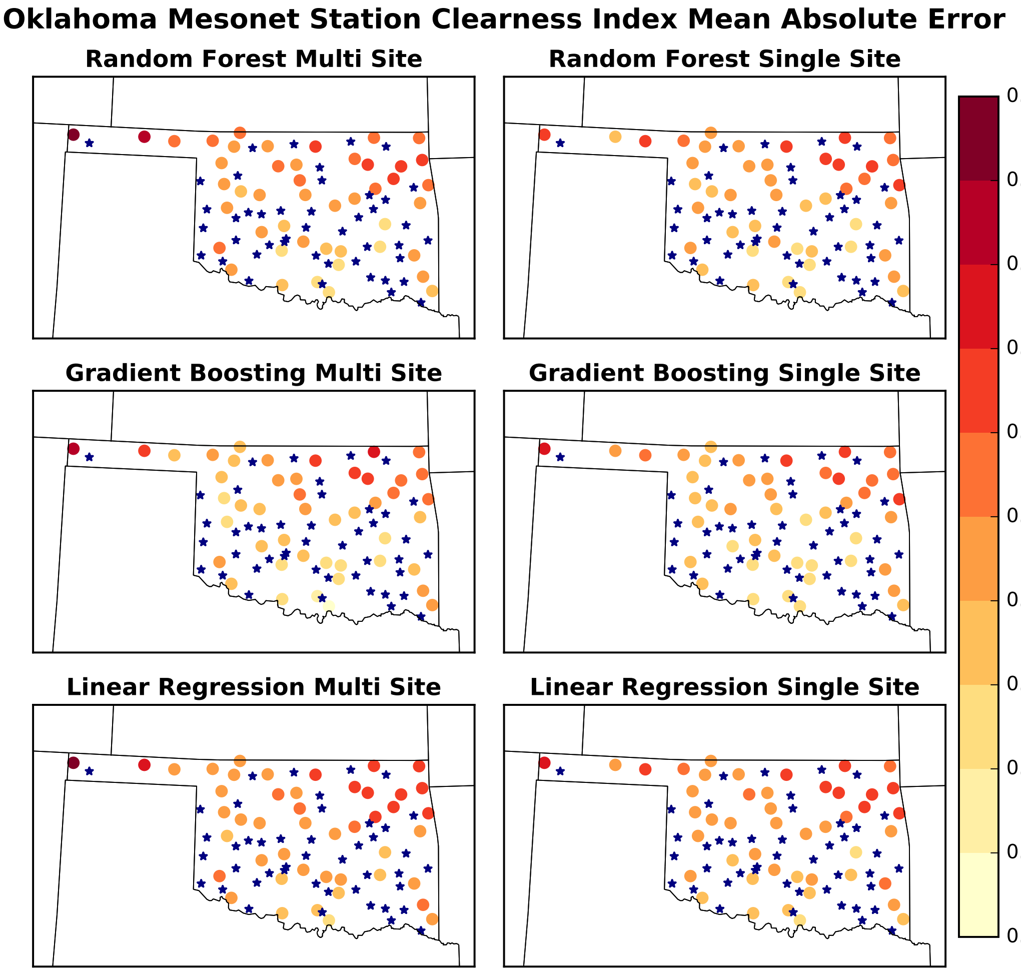

Oklahoma Mesonet Station Clearness Index Mean Absolute Error
Random Forest Multi Site
Random Forest Single Site
Gradient Boosting Multi Site
Gradient Boosting Single Site
Linear Regression Multi Site
Linear Regression Single Site
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05

Figure 5.9: Mean absolute error by station. Stations used for training are indicated with blue stars.

based on information about station spacing and average distance from other sites in the dataset.

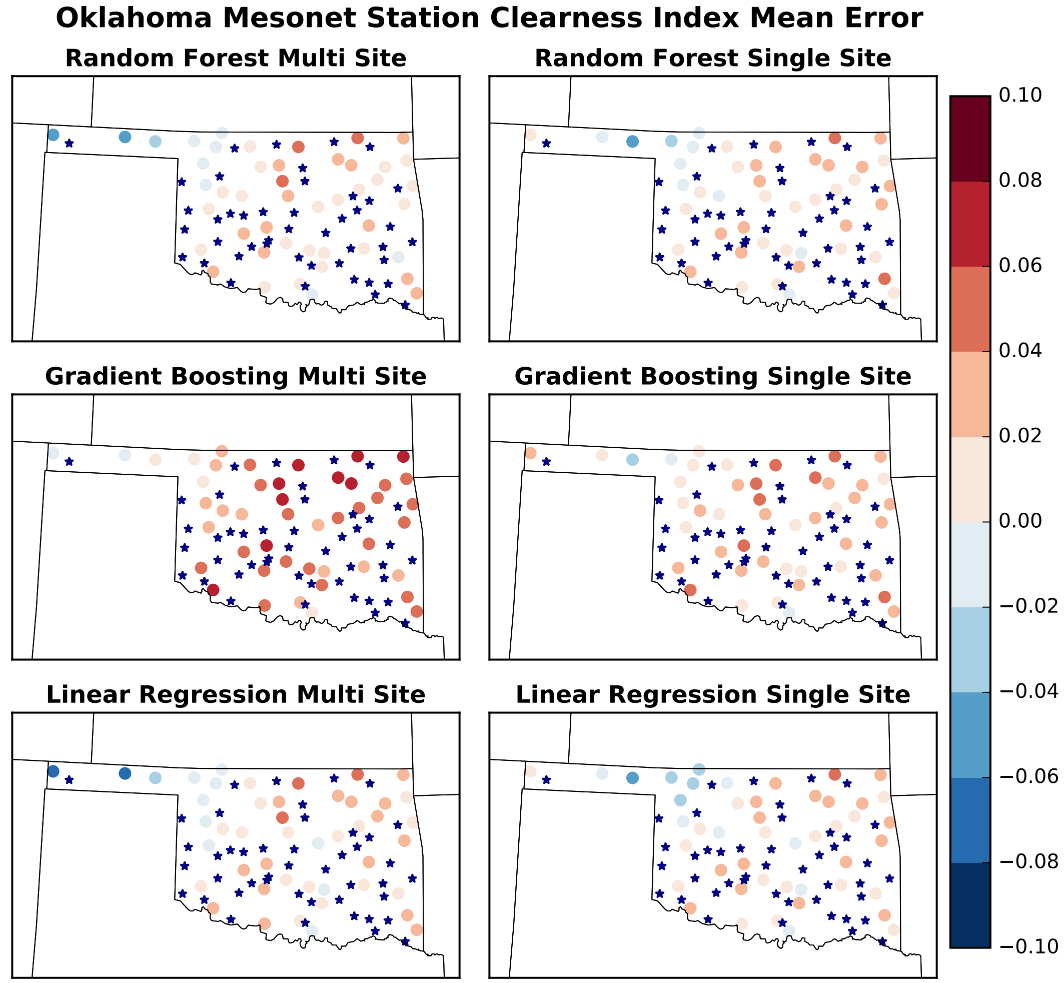

Oklahoma Mesonet Station Clearness Index Mean Error
Random Forest Multi Site
Random Forest Single Site
Gradient Boosting Multi Site
Gradient Boosting Single Site
Linear Regression Multi Site
Linear Regression Single Site
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10

Figure 5.10: Mean error by station. Stations used for training are indicated with blue stars.

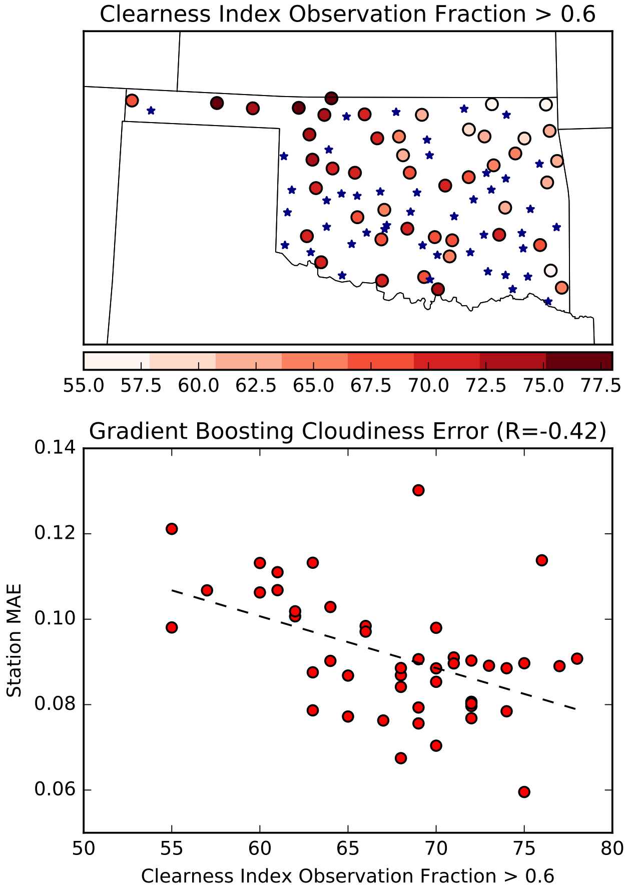


Figure 5.11: Percentage of clearness index observations greater than 0.6 by site, and the relationship between that percentage and mean absolute error.

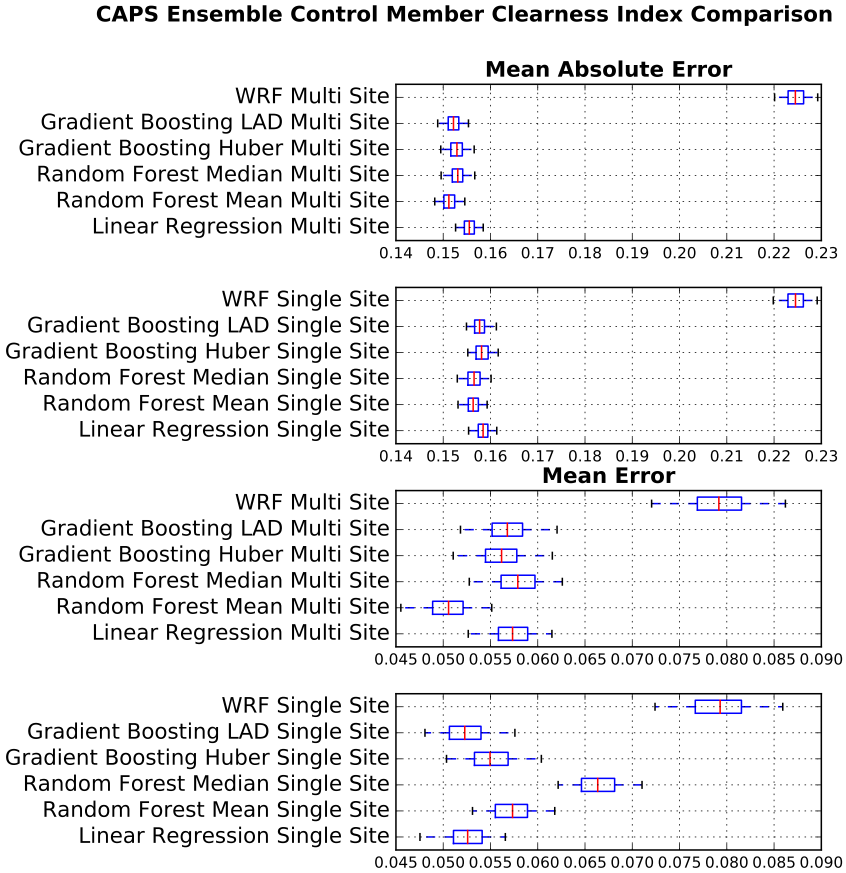


Figure 5.12: Bootstrap confidence intervals of mean absolute error and mean error for models trained on the CAPS control member.

5.2.2 Experiment 2: CAPS Ensemble Control Member

The CAPS Ensemble Control Member experiment tested different aspects of the statistical learning methods as well as the impact of utilizing higher resolution model output and variables describing upper air humidity. The CAPS Ensemble control member error chart (Fig. 5.12) shows that the control member has both a higher error and a strong positive bias. A potential major source of

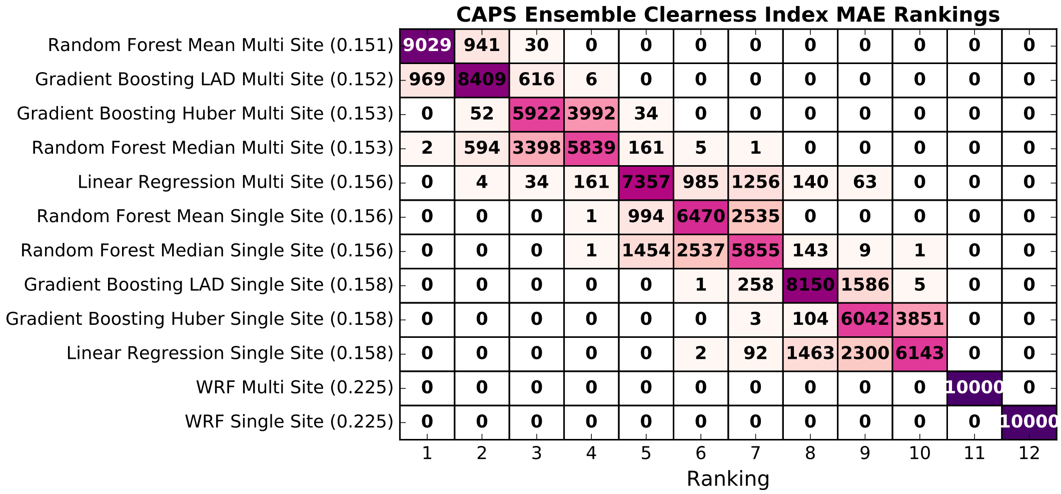
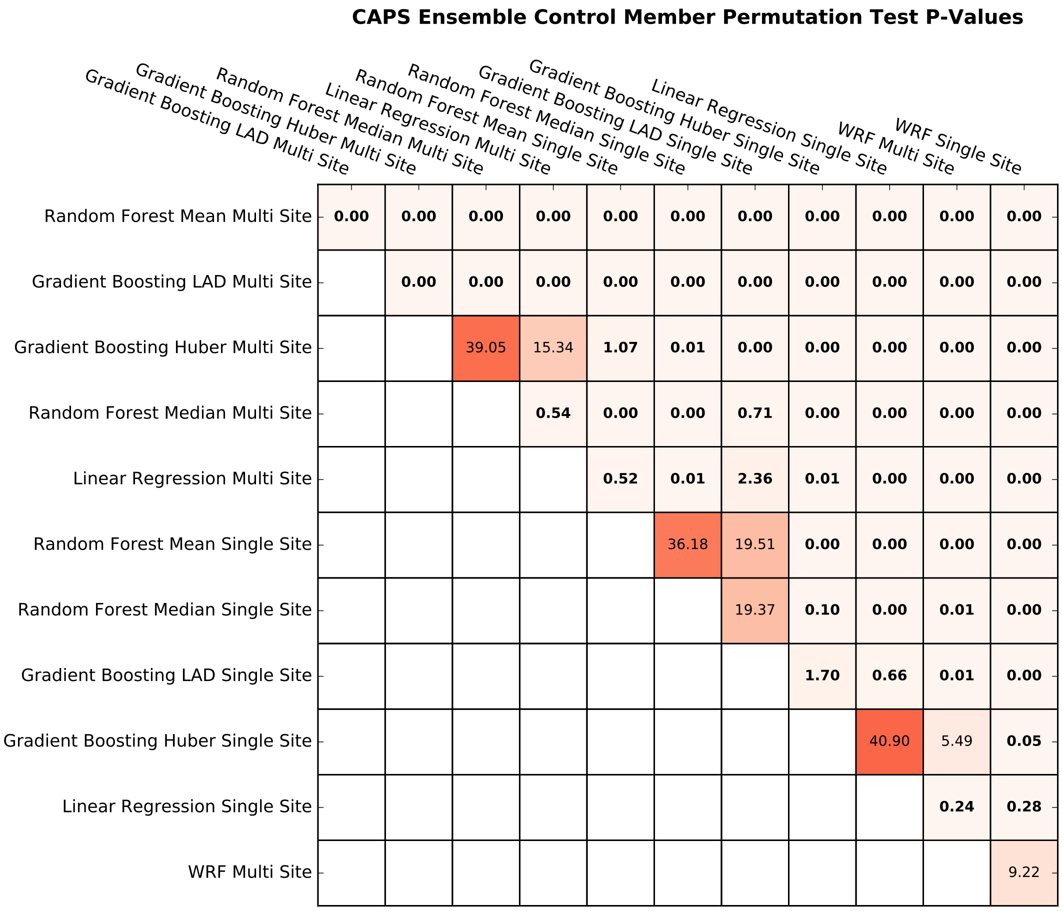


Figure 5.13: Bootstrap rankings of mean absolute error for models trained on the CAPS control member. The mean absolute error for each model is shown in parentheses.

the bias is that the WRF model outputs instantaneous solar irradiance at the top of the hour while the observations are for hourly averaged irradiance. Even with that source of bias, the machine learning models are all able to improve performance significantly. The Multi Site models tend to perform slightly better than the Single Site models, but there are small differences among the type of model or choice of model configuration. While there is some overlap in the bootstrap distributions of mean absolute error for each model, the ranking bootstrap distributions of the models group them into similar cohorts (Fig. 5.13). The Multi Site models outperform all of the Single Site models at statistically significant level (Fig. 5.14). The Random Forest Mean model has the top ranking over 91% of the time, which is statistically significant at the 5 % confidence level (Fig. 5.14), while the next three models tend to trade places in the rankings more often. Gradient Boosting models with the Huber loss function have higher error than those with Least Absolute Deviance.



CAPS Ensemble Control Member Permutation Test P-Values

	Gradient Boosting LAD Multi Site	Gradient Boosting Huber Multi Site	Random Forest Median Multi Site	Linear Regression Multi Site	Random Forest Mean Single Site	Random Forest Median Single Site	Gradient Boosting LAD Single Site	Gradient Boosting Huber Single Site	Linear Regression Single Site	WRF Multi Site	WRF Single Site
Random Forest Mean Multi Site	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting LAD Multi Site		**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Gradient Boosting Huber Multi Site			39.05	15.34	**1.07**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest Median Multi Site				**0.54**	**0.00**	**0.00**	**0.71**	**0.00**	**0.00**	**0.00**	**0.00**
Linear Regression Multi Site					**0.52**	**0.01**	**2.36**	**0.01**	**0.00**	**0.00**	**0.00**
Random Forest Mean Single Site						36.18	19.51	**0.00**	**0.00**	**0.00**	**0.00**
Random Forest Median Single Site							19.37	**0.10**	**0.00**	**0.01**	**0.00**
Gradient Boosting LAD Single Site								**1.70**	**0.66**	**0.01**	**0.00**
Gradient Boosting Huber Single Site									40.90	5.49	**0.05**
Linear Regression Single Site										**0.24**	**0.28**
WRF Multi Site											9.22

Figure 5.14: Permutation test p-values (multiplied by 100) for the comparison of the differences in mean absolute error between each machine learning model trained on CAPS Ensemble control member output. P-values in bold are statistically significant based on the false discovery method with a rate of 0.05 (α=0.03). Darker reds are associated with larger p-values.

The binned and averaged errors displayed in Fig. 5.15 show similar trends to those in Fig. 5.5 with little variation among the statistical learning models. The statistical learning model forecast values contain a more reliable distribution of observed values and improved on the sharpness of the raw model output, even if the average forecast value for a given observation value is still biased.

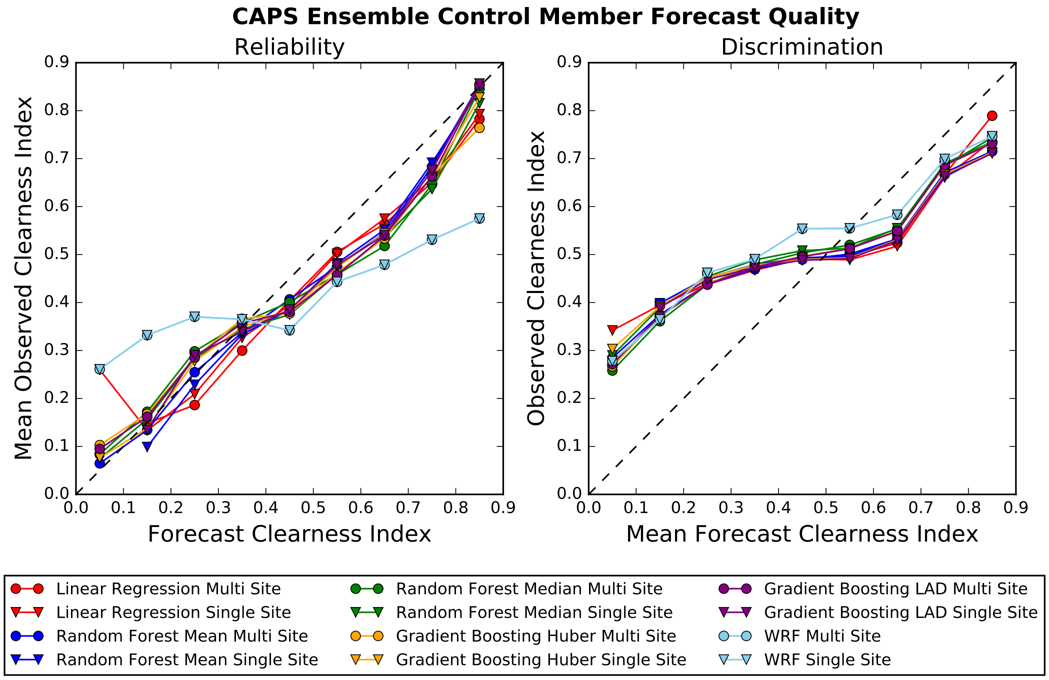


Figure 5.15: Forecast and observed marginal distributions for models trained on the CAPS control member.

The binned mean absolute error chart (Fig. 5.16) shows that the largest improvement over the raw model output occurs at higher and low values of clearness index. The magnitude of the mean and mean absolute errors were successfully reduced by the statistical learning models. The differences in the models were all fairly small except at low levels of clearness index where the linear regression models were either not as sharp or were more biased than the other approaches. The highest errors and smallest reduction in error from the raw ensemble occurred in the partly cloudy regime of clearness index.

The biggest differences in statistical learning model performance appeared when comparing the distributions of the forecast values from each model (Fig. 5.17). The observed clearness index has a bimodal distribution with peaks at 0.75 and 0.2 and a dip in the middle. The raw control member output also roughly captures this pattern but vastly underestimates the number of partly cloudy

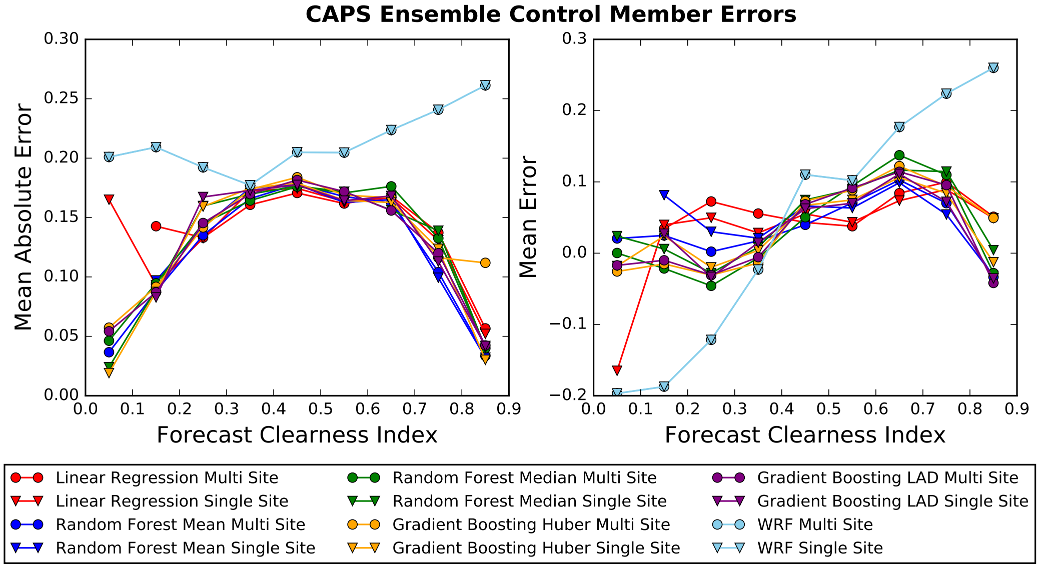


Figure 5.16: Mean absolute error binned by forecast value for models trained on the CAPS control member.

events and overestimates the number of clear sky events. The statistical learning models to varying degrees regress toward the mean of the clearness index distribution depending on their level of complexity and smoothness. The linear regression models display peak forecasts at a clearness index of 0.5 and decrease from there. The random forest using mean aggregation has a similar pattern but starts to capture the main peak at 0.75. The Multi Site gradient boosting models are better able to reduce the proportion of predictions in the middle of the distribution than the Single Site models although they also overestimate clear sky events more. The median random forest model comes the closest to capturing the variability at the low and middle sections of the clearness index spectrum, but it also has a larger positive bias at the high end.

The same evaluation was also performed on two other members of the CAPS ensemble with different microphysics schemes. The member with the Milbrandt and Yau microphysics scheme displayed slightly lower errors (approximately 0.14

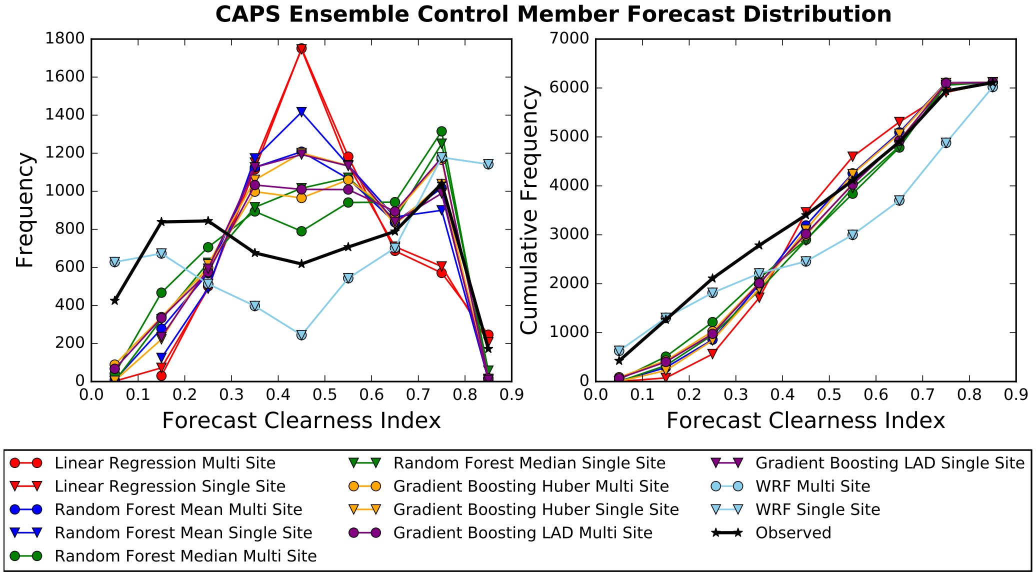


Figure 5.17: Forecast frequency for each clearness index bin for models trained on the CAPS control member.

vs. 0.15) for the statistical learning models and produced the Single Site models that outperform the Multi Site models. Other trends remained the same. The Morrison microphysics member performed slightly worse than the control member but was otherwise fairly similar in terms of performance trends. Decreasing the depth of the trees in the random forest led to forecast distributions closer to that of linear regression.

5.3 Discussion

The results from this experiment show that the statistical learning model configuration can have just as big an impact on performance as the choice of statistical learning model. The parameters with the biggest impact on performance are those that control the sharpness versus smoothness of the model fit. The tree depth parameters have a big impact on the spread of the forecast. Decreasing the tree depth may improve some error statistics, but it comes at the cost of capturing rarer events well. Some of these issues can be addressed by having a larger and more diverse training set, which helped the models in the GFS experiment. Regression models will generally not achieve perfect sharpness and will tend to be underdispersive, but a larger dataset and a model that can scale with the additional data can bring further improvements without a significant investment in model tuning.

Statistical models can generalize well to other sites as long as they are close enough in characteristics to sites in the training data. The Oklahoma Mesonet sites share fairly similar climate and terrain, which allowed the statistical learning model corrections to perform well using both Multi Site and Single Site methods. The type of interpolation used for the Single Site model does make a difference as using nearest neighbor interpolation resulted in a noticeable decrease in performance. The Cressman interpolation approach does appear to be fairly robust even if it is not as optimized or data-driven as the other methods. Aggregating data from multiple sites did not appear to have an overly positive or negative effect on performance for the most part. The conditions across sites in Oklahoma may have been similar enough that aggregating more sites did not add much additional information or cover the parameter space better.

The spatial and temporal window of each input variable into the solar irradiance machine learning model can have a large impact on performance. The observation used for this experiment was hourly averaged clearness index, so any cloud cover occurring over the previous hour has an impact on the clearness index value. The CAPS Ensemble output uses instantaneous solar irradiance while the GFS output is interpolated from 3-hourly output based on instantaneous cloud cover and solar position. The offset in representation leads to biases in the input variables that the machine learning model may not have the information to correct except on a systemic level. Some of the information about potential cloudiness can be inferred by examining the spatial neighborhood of a grid point and its variability in cases where the larger scale cloud cover pattern is similar. Some cloud cover, such as shallow cumulus, may not be resolved by the NWP model or properly parameterized, and so the model will not account for its impact in the irradiance at all. If the model has significant temporal biases in developing convective clouds during the day or in capturing the larger scale synoptic setup, then the irradiance forecasts are also going to be biased. Incorporating information from more times and over a larger spatial area can help account for these issues, but that solution requires more time for data processing and may not be feasible in an operational gridded forecasting system.

Because clearness index does not follow a Gaussian distribution, traditional regression models will not be able to capture the full uncertainty properly. Since clearness index values tend to cluster into clear and cloudy regimes, one alternative method to a single regression model is to create a stacked model to predict the cloudiness regime first and then the clearness index value given the regime. Cloudiness regimes could be identified in the training data with Gaussian Mixture Models applied to the full distribution of clearness index values. Then the machine learning classifier would predict the probability of a given forecast to fall within each mixture distribution. Conditional regression models could then predict the clearness index value within each regime. Finally the regression predictions would be multiplied by the probabilities for each regime to create an optimal deterministic forecast. If the regression models also produced variance estimates, then a full probability density function of clearness index could be created.

There are caveats to generalizing these results to gridded solar forecasting performance in general. The presence of more complex terrain will likely lead to a decreased effectiveness in naive spatial interpolation and will require a more complex method to produce physically consistent forecasts. The CAPS ensemble was only run over a month, resulting in only 8 testing days. The CAPS ensemble was also run during the month of May, which is during a seasonal transition and increased amounts of storms and rain. The GFS experiment was run for the summer, which tends to be clearer and less volatile in Oklahoma. Ideally, an experiment would use multiple years of observations and forecast output from the same NWP model with a static configuration, but that is only available with coarse models like the GEFS Reforecast dataset (Hamill et al. 2013). Large archives of convection-allowing model runs would be beneficial for improving solar irradiance and a host of other high-impact weather predictions, such as the NCAR Ensemble (Schwartz et al. 2015).

Further experiments should be performed over a wider geographic area or one with more complex terrain, but there is a major shortage of well-maintained pyranometers that are both closely spaced and cover a large geographic area. Initial experiments were performed using pyranometer data from Remote Automatic Weather Stations (RAWS) sites in California but were discontinued due to data quality issues. Slater (2016) recently compiled an inventory of all available sources of solar irradiance data in the United States and evaluated the quality of the different sources. Future experiments in gridded solar irradiance forecasting should pull from this wider array of observations and consider using data assimilation techniques to construct the best gridded analysis.

5.4 Conclusions

Different statistical learning techniques and configurations for making gridded solar irradiance forecasts were evaluated to determine their impact on performance and physical realism. All statistical learning models tested were able to improve on raw NWP model output and reduce forecast biases. The models were able to generalize well to sites not included in the training data. Both interpolation of predictions as done in the operational gridded MOS system (Glahn et al. 2009) and direct application of a model trained across multiple sites performed well with little difference in the resulting forecasts. One of the key parameter choices was finding a good balance between the smoothness and sharpness of the forecasts. Biasing toward smoothness resulted in predictions clustering near the average solar irradiance value, which led to underestimation of both clear and cloudy days. All statistical models struggled at capturing the right number of cloudy and partly cloudy events, which is due to errors in the underlying NWP forecasts and the assumptions underlying the statistical learning models. Further explorations of data transformations, sample weighting, and loss functions are needed to capture the crucial extreme events better. Given the bounded nature of solar irradiance and clearness index, more performance improvements should be possible.

Chapter 6

Discussion and Future Work

The primary hypothesis of this dissertation was that properly configured decision tree ensemble machine learning models will produce day-ahead predictions of hail size and solar irradiance that show significantly more skill than raw NWP model output, physics-based diagnostic models, and linear regression. The secondary hypothesis was that properly configured decision tree ensemble machine learning models will produce distributions of forecasts that are physically consistent with distributions of observations. For hail prediction, I developed a probabilistic storm-based machine learning modeling system to predict if a modeled storm would produce hail and what the hail size distribution would be. The machine learning hail modeling system was evaluated on two convection-allowing model ensembles against other hail size and storm surrogate methods. The machine learning hail forecasts either had similar or significantly better performance than the other methods in statistics measuring accuracy, discrimination, and reliability at both the severe and significant severe hail size thresholds. The rankings of the different hail methods varied by metric and ensemble system. The machine learning models consistently produced a smaller proportion of false alarms compared with other methods. The reliability of all the methods varied but could be calibrated further by adjusting the size of the neighborhood and the width of the Gaussian smoother. The linear regression hail size model was a poorer discriminator for significant severe hail but was more reliable than the random forest methods. With these results in mind, I am not able to accept the dissertation hypothesis fully for hail forecasts. The current hail forecasts are competitive with the other methods and exceed them in some areas but do not consistently outperform them in all cases. In the case of the secondary hypothesis, I constrained the predictions of the machine learning models using the storm-based framework, multitask learning, and parametric hail size distributions. The resulting forecasts from the machine learning models maintained the relationship between the shape and scale parameters of the hail size distributions and forecasted higher hail probabilities in areas where large hail is more likely to occur. Because of this, I accept the secondary hypothesis for the machine learning hail forecasts.

For solar irradiance forecasting, I developed a set of gridded machine learning models and evaluated different configurations to determine which setup produces the lowest forecast error. Gradient boosted regression and random forest consistently produced lower errors than linear regression and raw output from the GFS and WRF models. Aggregating data from multiple sites into one machine learning model tended to outperform training separate machine learning models at each site. The machine learning models produced reliable predictions of clearness index and were able to discriminate between sunnier and cloudier days. However, the forecast distribution of clearness index tended to underforecast the frequency of cloudy days and had a slight overforecasting bias for clearer days. Because of the significant improvements in performance and the issues with the forecast distribution matching the observed distribution of clearness index, I accept the primary hypothesis and reject the secondary hypothesis. While some machine learning model configurations produced forecast distributions closer to the observed distributions, they still contained notable biases that could not be addressed with adjustments to the model parameters or model type.

Through my dissertation research, I also contributed to broader impacts by developing Hagelslag, an open source Python library for storm tracking, forecasting, and verification (Gagne II et al. 2016). The machine learning hail forecasts were run in real-time on the CAPS and NCAR ensembles as part of the 2016 Hazardous Weather Testbed Spring Experiment using this software. The storm tracking modules were used to perform an analysis of mid-level versus low-level updraft helicity tracks in the NCAR ensemble. Other academic and government researchers have also expressed interest in using the software as part of their research projects, and further development is planned so that the hail forecasts can be run on operational convection-allowing models.

The dissertation research areas revealed a few overarching insights about machine learning model development for high-impact weather applications. First, the choice of machine learning model had less impact on performance than choices made during pre-processing. The composition of the training data and the amount of stationarity between the training and testing data has a noticeable impact on the structure of the machine learning model and its performance. Between 2014 and 2015, the CAPS ensemble underwent changes in resolution and updates to the model and associated parameterization schemes. The machine learning models were still able to produce predictions with some skill, but the lack of stationarity added some biases to the forecasts and weakened the ability of the model to discriminate accurately among different hail sizes. The NCAR ensemble training setup, on the other hand, kept the model configuration fixed, resulting in less differences between the training and testing data and thus higher performance with less bias. For the solar data, the largest errors tended to come at the sites that were further away from other sites and in more variable terrain. Choosing poor model parameters, particularly for more complex models, led to overfitting.

Second, statistical and machine learning models need to be constrained in many ways to produce physically realistic forecasts. The initial hail forecast system described in Chapter 3 was set up to directly predict the maximum hail size with a regression model, but due to the high uncertainty of the data and the exponential distribution of hail sizes, the models would tend to forecast closer to the mean hail size in the dataset and would show a high variance in hail sizes across storms in similar environments. The second generation of hail forecasts in Chapter 4 adopted a set of choices that constrained the models while making them more robust and realistic. The hail size models predicted parameters of a gamma distribution fitted to a set of MESH values instead of the maximum MESH value within an object. Candidate hailstorms were matched to tracks instead of individual timesteps. Most importantly, the same model predicted the shape and scale parameters together, instead of optimizing them independently, which resulted in more realistic depictions of hail sizes within the objects. Finally, instead of applying a single hail size value to all points within an object, the spatial distribution of column-integrated graupel values was used in order to preserve the spatial structure of the storms. These choices not only led to forecasts that looked more physically realistic, but they also performed better and captured extreme events better. With the solar data, growing deeper trees, using loss functions that weighed large errors linearly instead of quadratically, and using the random forest ensemble median instead of the mean led to sharper forecasts even with smaller datasets.

Even with these adjustments, predicting extreme, rare events well with decision tree ensemble models is still challenging. While the individual trees in the random forest tend to produce sharp, if not highly accurate forecasts, the averaging process brings the consensus result closer to climatology and tends to produce forecast distributions that are Gaussian even when the observed distribution is not. Weighted averages of the tree predictions based on out-of-bag error estimates and fuzzy combinations of tree predictions have only led to marginal improvements in accuracy (Kuncheva 2003; Bonissone et al. 2010; Shahzad et al. 2015), and these methods were not evaluated on sharpness nor reliability. A regularized linear model fit to the individual tree predictions based on validation set performance may add more sharpness and predictive skill, particularly if the optimization function contained a term minimizing the difference in forecast and observation standard deviations. There may also be benefits to applying kernel dressing approaches, such as Bayesian model averaging, to the individual trees to estimate the prediction uncertainty better.

While this dissertation showed that machine learning could be very beneficial for improving prediction of high impact weather, it did not definitively show how machine learning could also improve physical understanding of the phenomena being predicted. While variable importance rankings from random forests do provide some insight into how the the models are structured, the importance scores themselves are subject to many sources of variability, including collinearity with other input variables. Some variables with low importance scores still showed some predictive skill when used in isolation. Using multiple variable importance metrics, especially ones that account for co-linearity, could help provide stronger evidence for physical connections. Combining raw variable importance scores with analysis of the discrimination skill of each variable may also be useful. There is value in discovering which cases are not predicted well by machine learning models and discovering why by analyzing the underlying data.

Given the large stakes associated with high impact weather, humans still have a key role in the forecasting process. In order to stay engaged in the forecast process, the forecaster workflow needs to be designed so that they contribute their time to areas where automation struggles (Pagano et al. 2016). Automation should be integrated in such a way to be complementary to forecaster skills and not be antagonistic. In this framework, forecasters would interact with different sources of guidance as needed to evaluate the quality of their representation of the atmosphere (Doswell III 1992). Forecasters would be able to identify which scenarios presented by the automated system are more realistic and deemphasize those that are not in the final product. More time would be spent on constructing narratives and recommendations for end users instead of on forecast entry, although being more removed from the forecast generation process may make in-depth communication more difficult (Pagano et al. 2016). Visual analytics systems that enable realtime interactive with forecast data could help with this. Forecaster evaluation would focus less on skill against the automated guidance but on how effectively they communicate their forecasts to the public.

6.1 Future Work

Statistical models tend to perform better when the relationship between the inputs and the prediction is less noisy. The input variables and datasets for this dissertation were limited to what was already available. In the future, the data scientist developing the machine learning framework for a NWP modeling system should work closely with the NWP modelers to extract variables that are as closely tied to the physical processes driving a weather phenomenon as possible. For hail prediction, this would involve extracting more information about conditions within the hail growth zone and within the melting layer instead of relying on bulk severe weather statistics, such as CAPE and shear. Solar irradiance models could benefit greatly from having more variables that capture the evolution of clouds over the timeframe when model output is not stored, similar to the hourly-maximum fields used for severe weather forecasting. Cumulative measures of the cloud cover variability over an hour at different heights would be very useful for capturing the partly cloudy events that tend to have the highest forecast errors.

Probabilistic forecasts of rare events would likely benefit from using more flexible probability distribution representations. A single parametric distribution will struggle with capturing the behavior of both the common events in the distribution and the tails. Mixture distributions and splines offer more degrees of freedom to represent heavy-tailed and multi-modal distributions that may occur in some circumstances. One potential approach for predicting mixture distributions is first training a classifier to predict the probability of the event falling within a particular quantile of the overall distribution. Then, separate regression models are trained to predict the parameters of the distribution describing that quantile. The classifier probabilities would then be used to create a blended distribution from the quantile distributions. The weighted combination of distributions hedges against the risk of picking the wrong quantile. This method could be particularly useful for representing the probability of cloud cover for solar irradiance forecasting and the variability during time scales smaller than that output by the NWP models. This additional uncertainty information could be very useful for electric utilities at both hourly and day-ahead time scales.

While the predictive accuracy of machine learning models has been extensively studied, there has been little work done to determine how forecasters interpret the products and to determine what products are most valuable for them. Machine learning predictions, particularly ones that produce probability distributions, can be displayed with varying degrees of complexity. This dissertation focused on producing forecasts from the machine learning models that conformed to existing products, but that left a lot of information hidden. An interactive visualization system, similar to the Probabilistic Hazards Information tool (Karstens et al. 2015), could allow forecasters to query hazard probabilities from a set of storms and compare them with storm climatologies and other weather variables. This system could generate trust in the machine learning methods at the expense of requiring more analysis time and a higher cognitive load. Additional studies of forecaster interaction with machine learning products are needed to assess these issues.

While the machine learning models in this dissertation did incorporate some spatial and temporal data into their predictions, they were not able to interpret more complex spatial structures and patterns. Traditional machine learning models assume each input variable is independent of the others, which is generally not the case for variables that are spatially and temporally related. Spatial statistics summarize the local variability but hide structure and texture that may have predictive usefulness. Human forecasters can identify spatial structures in weather phenomena, such as hook echoes on supercells, that provide predictive information that would be missed if only the mean and standard deviation of radar reflectivity were available. The spatiotemporal relational random forest (McGovern et al. 2013) can utilize the predictive power of some of these spatial structures but requires a human expert to identify and extract these structures in advance. Convolutional neural networks (Dieleman et al. 2015) can learn multiple sets of spatial filters that identify features and textures at small and large scales in gridded data. While pre-processing is still required to isolate the area being studied, the model is able to learn features on its own from the raw pixel values while being constrained by the neural network structure, max pooling to reduce the impact of minor translation errors, and dropout regularization to strengthen signals that are found in the data. Convolutional neural networks and other forms of deep learning have already made impressive gains in predictive performance in many challenging domains but have yet to receive wide use in meteorology. The challenge with implementing deep learning on meteorological data will be posing the problems in ways that add predictive skill while being computationally efficient.

Reference List

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

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic Forecasts of Mesoscale Convective System Initiation Using the Random Forest Data Mining Technique. *Wea. Forecasting*, **31**, 581–599. doi:10.1175/WAF-D-15-0113.1.

Benjamin, S., 2014: From the RAPv3/HRRRv2 deterministic era to the NARRE/HRRRE ensemble era. *2014 Warn-On-Forecast and High Impact Weather Workshop*, Norman, OK, National Severe Storms Laboratory, 1.

Benjamini, Y. and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **B57**, 289–300.

Blend, R. and M. Schubert, 2000: Cyclone Tracking in Different Spatial and Temporal Resolutions. *Mon. Wea. Rev.*, **128**, 377–382.

Bonissone, P., J. M. Cadenas, M. C. Garrido, and R. A. Diaz-Valladares, 2010: A fuzzy random forest. *International Journal of Approximate Reasoning*, **51**, 729–747. doi:10.1016/j.ijar.2010.02.003.

Breiman, L., 2001a: Random Forests. *Mach. Learn.*, **45**, 5 – 32.

Breiman, L., 2001b: Statistical Modeling: The Two Cultures. *Statist. Sci.*, **16**, 199–231.

——, J. Friedman, C. J. Stone, and R. A. Olshen, 1984: *Classification and regression trees*. CRC Press, 358 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Brimelow, J. C., G. W. Reuter, and E. R. Poolman, 2002: Modeling Maximum Hail Size in Alberta Thunderstorms. *Wea. Forecasting*, **17**, 1048–1062.

——, ——, R. Goodson, and T. W. Krauss, 2006: Spatial Forecasts of Maximum Hail Size Using Prognostic Model Soundings and HAILCAST. *Wea. Forecasting*, **21**, 206–219.

Brooks, H. E., 2004: Tornado-Warning Performance in the Past and Future: A Perspective from Signal Detection Theory. *Bull. Amer. Meteor. Soc.*, **85**, 837–843.

Brown, T. M., W. H. Pogorzelski, and I. M. Giammanco, 2015: Evaluating Hail Damage Using Property Insurance Claims Data. *Weather, Climate, and Society*, **7**, 197–210. doi:10.1175/WCAS-D-15-0011.1.

Cangialosi, J. P. and J. L. Franklin, 2015: 2014 National Hurricane Center Forecast Verification Report. NOAA/NWS/NCEP/National Hurricane Center Tech. Rep. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2014.pdf.]

Caruana, R., 1997: Multitask Learning. *Mach. Learn.*, **28** (1), 41–75.

—— and A. Niculescu-Mizil, 2006: An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, New York, NY, ACM, 161–168.

Changnon, S. A., 2009: Increasing major hail losses in the U.S. *Climate Change*, **96**, 161–166.

——, R. A. Pielke Jr., D. Changnon, R. T. Sylves, and R. Pulwarty, 2000: Human Factors Explain the Increased Losses from Weather and Climate Extremes. *Bull. Amer. Meteor. Soc.*, **81**, 437–442.

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

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

——, R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of Object-Based Time-Domain Diagnostics for Tracking Precipitation Systems in Convection-Allowing Models. *Wea. Forecasting*, **29**, 517–542. doi:10.1175/WAF-D-13-00098.1.

——, S. J. Weiss, J. S. Kain, and Coauthors, 2012a: An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

——, ——, ——, and ——, 2012b: An overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

Cressman, G. P., 1959: An Operational Objective Analysis System. *Mon. Wea. Rev.*, **87**, 367–374.

Cumming, G. and S. Finch, 2005: Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, **60**, 170–180. doi:10.1037/0003-066X.60.2.170.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The Method fo Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267.

Diagne, M., M. David, P. Lauret, J. Boland, and N. Schmutz, 2013: Review of solar irradiance forecasting methods and proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, **27**, 65–76. doi:10.1016/j.rser.2013.06.042.

Dieleman, S., K. W. Willett, and J. Dambre, 2015: Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, **450**, 1441–1459. doi:10.1093/mnras/stv632.

Dixon, M. and G. Wiener, 1993: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A Radar-based Methodology. *J. Atmos. Oceanic Technol.*, **10**, 785–797. doi:10.1175/1520-0426(1993)010⟨0785:TTITAA⟩2.0.CO;2.

Doswell III, C. A., 1992: Forecaster Workstation Design: Concepts and Issues. *Wea. Forecasting*, **7**, 398–407. doi:10.1175/1520-0434(1992)007⟨0398:FWDCAI⟩2.0.CO;2.

——, 2004: Weather Forecasting by Humans– Heuristics and Decision Making. *Wea. Forecasting*, **19**, 1115–1126.

Dunn, O. J., 1961: Multiple Comparisons Among Means. *Journal of the American Statistical Association*, **56**, 52–64. doi:10.1080/01621459.1961.10482090.

Efron, B. and R. J. Tibshirani, 1994: *An Introduction to the Bootstrap.* CRC Press, 456 pp.

Engel, C. and E. E. Ebert, 2012: Gridded Operational Consensus Forecasts of 2-m Temperature over Australia. *Wea. Forecasting*, **27**, 301–322.

Falls, L. W., 1974: The Beta Distribution: A Statistical Model for World Cloud Cover. *Journal of Geophysical Research*, **79**, 1261–1264.

Fawbush, E. F. and R. C. Miller, 1953: A method for forecasting hailstone size at the earth's surface. *Bull. Amer. Meteor. Soc.*, **34**, 235–244.

Friedman, J., 2001: Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189–1232.

Gagne II, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Oceanic Technol.*, **26** (7), 1341–1353. doi:10.1175/2008JTECHA1205.1.

Gagne II, D. J., A. McGovern, N. Snook, R. Sobash, J. Labriola, J. K. Williams, S. E. Haupt, and M. Xue, 2016: Hagelslag: Scalable Object-Based Severe Weather Analysis and Forecasting. *Proceedings of the Sixth Symposium on Advances in Modeling and Analysis Using Python*, New Orleans, LA, Amer. Meteor. Soc., 447. [Available online at https://ams.confex.com/ams/96Annual/webprogram/Paper280723.html.]

Gilbert, G. K., 1884: Finley's tornado predictions. *American Meteorological Journal*, **1**, 166–172.

Gilbert, K. K., J. P. Craven, D. R. Novak, T. M. Hamill, J. Sieveking, D. P. Ruth, and S. J. Lord, 2015: An Introduction to the National Blend of Global Models Project. *Proceedings, Special Symposium on Model Postprocessing and Downscaling*, Phoenix, AZ, Amer. Meteor. Soc., 3.1. [Available online at https://ams.confex.com/ams/95Annual/webprogram/Paper267282.html.]

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The Gridding of MOS. *Wea. Forecasting*, **24**, 520–529.

Glahn, H. R. and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Data Set. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565. doi:10.1175/BAMS-D-12-00014.1.

Han, L., S. Fu, L. Zhao, Y. Zheng, H. Wang, and Y. Lin, 2009: 3D convective storm identification, tracking, and forecasting–An enhanced TITAN algorithm. *J. Atmos. Oceanic Technol.*, **26**, 719–732.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning*, 2nd ed. Springer, 745 pp.

Hewson, T., 1998: Objective fronts. *Meteorol. Appl.*, **5**, 37–65.

Ho, T. K., 1998: The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 832–844.

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

Holmgren, W. F., R. W. Andrews, A. T. Lorenzo, and J. S. Stein, 2015: PVLIB Python 2015. *Proceedings of the 42nd Photovoltaic Specialists Conference*, New Orleans, LA, IEEE, 1–5.

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

Hsu, W.-R. and A. H. Murphy, 1986: The Attributes Diagram: A Geometrical Framework for Assessing the Quality of Probability Forecasts. *International Journal of Forecasting*, **2**, 285–293. doi:10.1016/0169-2070(86)90048-8.

Jensen, D. D. and P. R. Cohen, 2000: Multiple Comparisons in Induction Algorithms. *Machine Learning*, **38**, 309–338. doi:10.1023/A:1007631014630.

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

Jimenez, P. A., and Coauthors, 2016: WRF-Solar: An augmented NWP model for solar power prediction. Model description and clear sky assessment. *Bull. Amer. Meteor. Soc.*, In Press pp. doi:10.1175/BAMS-D-14-00279.1.

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

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. W. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The Storm Cell Identification and Tracking Algorithm: An Enhanced WSR-88D Algorithm. *Wea. Forecasting*, **13** (2), 263–276. doi:10.1175/1520-0434(1998)013⟨0263:TSCIAT⟩2.0.CO;2.

Jurado, M., J. M. Caridad, and V. Ruiz, 1995: Statistical distribution of the clearness index with radiation data integrated over five minute intervals. *Solar Energy*, **55**, 469–473.

Kain, J. S., and Coauthors, 2008: Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP. *Weather and Forecasting*, **23** (5), 931–952. doi:10.1175/WAF2007106.1.

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

Karstens, C. D., and Coauthors, 2015: Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed. *Weather and Forecasting*, **30** (6), 1551–1570. doi:10.1175/WAF-D-14-00163.1.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective Prediction of Five-Day Mean Temperatures during Winter. *Journal of Meteorology*, **16**, 672–682.

Kong, F., 2014: 2014 CAPS Spring Forecast Experiment Program Plan. Center for Analysis and Prediction of Storms Tech. Rep.

Kuncheva, L. I., 2003: ”Fuzzy” Versus ”Nonfuzzy” in Combining Classifiers Designed by Boosting. *IEEE Transactions on Fuzzy Systems*, **11**, 729–741.

Lakshmanan, V. and T. Smith, 2009: Data Mining Storm Attributes from Spatial Grids. *J. Atmos. Oceanic Technol.*, **26**, 2353–2365.

—— and ——, 2010: An Objective Method of Evaluating and Devising Storm-Tracking Algorithms. *Wea. Forecasting*, **25**, 701–709. doi:10.1175/2009WAF2222330.1.

——, 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.

——, K. L. Elmore, and M. B. Richman, 2010: Reaching Scientific Consensus Through a Competition. *Bull. Amer. Meteor. Soc.*, **91**, 1423–1427.

Lim, K.-S. and S.-Y. Hong, 2010: Development of an Effective Double-Moment Cloud Microphysics Scheme with Prognostic Cloud Condensation Nuclei (CCN) for Weather and Climate Models. *Mon. Wea. Rev.*, **138**, 1587–1612.

Limbach, S., E. Schömer, and H. Wernli, 2012: Detection, tracking and event localization of jet stream features in 4-D atmospheric data. *Geosci. Model Dev.*, **5**, 457–470.

Lindquist, M. A. and A. Mejia, 2015: Zen and the Art of Multiple Comparisons. *Psychosomatic Medicine*, **77**, 114–125. doi:10.1097/PSY.0000000000000148.

Lorenz, E., J. Hurka, D. Heinemann, and H. G. Beyer, 2009: Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **2**, 2–10.

Lowry, D. A. and H. R. Glahn, 1976: An Operational Model for Forecasting Probability of Precipitation- PEATMOS PoP. *Mon. Wea. Rev.*, **104**, 221–232.

Manzato, A., 2013: Hail in Northeast Italy: A Neural Network Ensemble Forecast Using Sounding-Derived Indices. *Wea. Forecasting*, **28**, 3–28. doi:10.1175/WAF-D-12-00034.1.

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

Mass, C. F. and L. E. Madaus, 2014: Surface Pressure Observations from Smartphones: A Potential Revolution for High-Resolution Weather Prediction? *Bull. Amer. Meteor. Soc.*, **95**, 1343–1349. doi:10.1175/BAMS-D-13-00188.1.

McGovern, A., D. J. Gagne II, J. Basara, T. M. Hamill, and D. Margolin, 2015: Solar energy prediction: an international contest to initiate interdisciplinary research on compelling meteorological problems. *Bull. Amer. Meteor. Soc.*, **96**, 1388–1395. doi:10.1175/BAMS-D-14-00006.1.

McGovern, A., D. J. Gagne II, J. K. Williams, R. A. Brown, and J. B. Basara, 2013: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Machine Learning*, 1–24. doi:10.1007/s10994-013-5343-x.

McPherson, R. A., and Coauthors, 2007: Statewide Monitoring of the Mesoscale Environment: A Technical Update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, **24**, 301–321.

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

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

Moore, J. T. and J. P. Pino, 1990: An interactive method for estimating maximum hailstone size from forecast soundings. *Wea. Forecasting*, **5**, 508–526.

Morrison, H. and A. Gettelman, 2008: A New Two-Moment Bulk Stratiform Cloud Microphysics Scheme in the Community Atmosphere Model, Version 3 (CAM3). Part I: Description and Numerical Tests. *J. Climate*, **21**, 3642–3659. doi:10.1175/2008JCLI2105.1.

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

——, ——, G. H. Bryan, K. Ikeda, S. A. Tessendorf, and G. Thompson, 2015: Parameterization of Cloud Microphysics Based on the Prediction of Bulk Ice Particle Properties. Part II: Case Study Comparisons with Observations and Other Schemes. *J. Atmos. Sci.*, **72**, 312–339. doi:10.1175/JAS-D-14-0066.1.

Morss, R. E. and F. M. Ralph, 2007: Use of Information by National Weather Service Forecasters and Emergency Managers during CALJET and PACJET-2001. *Wea. Forecasting*, **22**, 539–555.

Munkres, J., 1957: Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, **5**, 32–38.

Murphy, A. H., 1973: A New Vector Partition of the Probability Score. *J. Appl. Meteor.*, **12**, 595–600.

——, 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. Forecasting*, **8**, 281–293. doi:10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2.

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

Nateghi, R., S. Guikema, and S. M. Quiring, 2014: Power Outage Estimation for Tropical Cyclones: Improved Accuracy with Simpler Models. *Risk Analysis*, **34** (6), 1069–1078. doi:10.1111/risa.12131.

Nelson, S. P., 1983: The Influence of Storm Flow Structure on Hail Growth. *J. Atmos. Sci.*, **40**, 1965–1983.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and Temperature Forecast Performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504. doi:10.1175/WAF-D-13-00066.1.

——, D. R. Bright, and M. J. Brennan, 2008: Operational Forecaster Uncertainty Needs and Future Roles. *Wea. Forecasting*, **23**, 1069–1084.

Pagano, T. C., F. Pappenberger, A. W. Wood, M.-H. Ramos, A. Persson, and B. Anderson, 2016: Automation and human expertise in operational river forecasting. *Wiley Interdisciplinary Reviews: Water*, In Press pp. doi:10.1002/wat2.1163.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pliske, R. M., B. Crandall, and G. Klein, 2004: Competence in Weather Forecasting. *Psychological Investigations of Competence in Decision Making*, K. Smith, J. Shanteau, and P. Johnson, Eds., Cambridge University Press, 40–70.

Reda, I. and A. Andreas, 2003: Solar Position Algorithm for Solar Radiation Applications. NREL Tech. Rep. TP-560-34302.

Renné, D. S., 2014: Emerging Meteorological Requirements to Support High Penetrations of Variable Renewable Energy Sources: Solar Energy. *Weather Matters for Energy*, A. Troccoli, L. Dubus, and S. E. Haupt, Eds., Springer, 257–273.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24**, 601–608.

Rosencrants, T. D. and W. S. Ashley, 2015: Spatiotemporal analysis of tornado exposure in five US metropolitan areas. *Nat. Hazards*, **78**, 121–140. doi:10.1007/s11069-015-1704-z.

Schwartz, C. S., and Coauthors, 2010: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership. *Wea. Forecasting*, **25**, 263–280.

——, G. S. Romine, R. A. Sobash, K. Fossell, and M. L. Weisman, 2015: NCAR's Experimental Real-Time Convection-Allowing Ensemble Prediction System. *Wea. Forecasting*, **30**, 1645–1654. doi:10.1175/WAF-D-15-0103.1.

Shahzad, R. K., M. Fatima, N. Lavesson, and M. Boldt, 2015: Consensus Decision Making in Random Forests. *International Workshop on Machine Learning, Optimization and Big Data* Springer 347–358.

Shaker, H., H. Zareipour, and D. Wood, 2016: Impacts of large-scale wind and solar power integration on California's net electrical load. *Renewable and Sustainable Energy Reviews*, **58**, 761–774. doi:10.1016/j.rser.2015.12.287.

Skitka, L. J., K. L. Mosier, and M. Burdick, 1999: Does automation bias decision-making? *Int. J. Human-Computer Studies*, **51**, 991–1006. doi:10.1006/ijhc.1999.0252.

Slater, A. G., 2016: Surface Solar Radiation in North America: A Comparison of Observations, Reanalyses, Satellite, and Derived Products. *J. Hydrometeor.*, **17**, 401–420. doi:10.1175/JHM-D-15-0087.1.

Snellman, L. W., 1977: Operational Forecasting Using Automated Guidance. *Bull. Amer. Meteor. Soc.*, **58**, 1036–1044.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. Fossell, and M. L. Weisman, 2016: Severe Weather Prediction Using Storm Surrogates from an Ensemble Forecasting System. *Wea. Forecasting*, **31** (1), 255–271. doi:10.1175/WAF-D-15-0138.1.

——, J. S. Kain, D. R. Bright, A. R. Dean, M. 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.

Stewart, T. R., 2001: Improving Reliability of Judgmental Forecasts. *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Springer US, 81–106.

——, P. J. Roebber, and L. F. Bosart, 1997: The Importance of the Task in Analyzing Expert Judgement. *Organizational Behavior and Human Decision Processes*, **69**, 205–219.

——, W. R. Moninger, K. F. Heideman, and P. Reagan-Cirincione, 1992: Effects of Improved Information on the Components of Skill in Weather Forecasting. *Organizational Behavior and Human Decision Processes*, **53**, 107–134.

Sukoriansky, S., B. Galperin, and V. Perov, 2005: Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. *Bound.-Layer Meteor.*, **117**, 231–257. doi:10.1007/s10546-004-6848-4.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115.

Tibshirani, R., 1996: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Tversky, A. and D. Kahneman, 1974: Judgment under Uncertainty: Heuristics and Biases. *Science*, **185**, 1124–1131.

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.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The Resolution Dependence of Explicitly Modeled Convective Systems. *Mon. Wea. Rev.*, **125**, 527–548. doi:10.1175/1520-0493(1997)125⟨0527:TRDOEM⟩2.0.CO;2.

Wilks, D. S., 2006: On “Field Significance” and the False Discovery Rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189. doi:10.1175/JAM2404.1.

——, 2011: *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Academic Press, 676 pp.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Machine Learning*, **95** (1), 51–70. doi:10.1007/s10994-013-5346-7.

Williams, J. K., P. P. Neilley, J. P. Koval, and J. McDonald, 2016: Adaptable Regression Method for Ensemble Consensus Forecasting. *Proceedings, Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, AZ, AAAI, 3915–3921. [Available online at http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12492.]

Witt, A., M. D. Eilts, G. J. Stumph, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An Enhanced Hail Detection Algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303.

Zamo, M., O. Mestre, P. Arbogast, and O. Pannekoucke, 2014: A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. *Solar Energy*, **105**, 792–803. doi:10.1016/j.solener.2013.12.006.

Zhang, J., K. Howard, C. Langston, and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338.

Zou, H. and T. Hastie, 2005: Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.