UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

FLOW DEPENDENT EVALUATION AND TRAINING OF RANDOM FOREST BASED PROBABILISTIC FORECASTS OF SEVERE WEATHER HAZARDS

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

By

ANDREW SHEARER
Norman, Oklahoma
2023

FLOW DEPENDENT EVALUATION AND TRAINING OF RANDOM FOREST BASED PROBABILISTIC FORECASTS OF SEVERE WEATHER HAZARDS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Aaron Johnson (Chair)

Dr. Xuguang Wang (Co-Char)

Dr. Amy McGovern

Acknowledgments

I would like to give acknowledgements to everyone on my committee. Dr. Aaron Johnson, you have been a great mentor and source of guidance on both research and how to balance that along with personal life. Dr. Xuguang Wang, you pushed me to strive only for the absolute best for which I am thankful. Dr. Amy McGovern, your insight into the way different Machine Learning products work was invaluable and without, a lot of the findings and claims I make in this thesis would not be possible without your work. I appreciate their understanding, insightful, and productive views on my work and many other aspect of graduate life.

I would also like to credit the NOAA JTTI grant NA20OAR4590358 for funding this work. Without this grant, I would not have had been able to conduct any of this research or even go to the University of Oklahoma. This opportunity to do research on severe weather has been my dream since I was young, so to fulfill this dream is such a rewarding experience.

Finally, I want to thank all my friends and family in supporting me. To my friends in particular, Bradley Lampkin, Saurabh Patil, and others who have been there with me through the trying times. I also want to thank Elizabeth Spicer, for being an amazing partner and a source of help and inspiration in both the good times and the bad. Finally, and with great emotion, I want to thank my family and particularly my dad. While your battle of cancer sadly ended, mom, Michael, and myself are eternally grateful for your nearly ten-year fight to allow for many more good memories.

Table of Contents

Acknowledgments **iv**

List Of Tables **vii**

List Of Figures **viii**

Abstract **xi**

1 Introduction **1**

2 Data and Methods **5**

2.1 Data Selection and Preprocessing 5

2.1.1 Ensemble Forecast Model 5

2.1.2 Preprocessing of Predictors 7

2.1.3 Predictands 9

2.2 Random Forest Model Creation 9

2.2.1 Random Forest Description and Selected Hyperparameters . . 10

2.2.2 Predictor Selection 11

2.3 Flow-Dependent (FD) Model Training 12

2.3.1 Domain Average Classification 13

2.3.2 Spatial Pattern Classification 15

2.3.3 Control Experiment (Randomly Trained Model) 17

2.4 Diagnostic and Verification Techniques 18

2.4.1 Brier Score, AUC, Reliability 19

2.4.2 Feature Importance and Permutation Feature Importance . . . 21

2.4.3 Tree Interpreter 22

3 Results **24**

3.1 Evaluation of Non-Flow Dependent Model 24

3.1.1 BSS Comparison of Full Model to FPS Model 25

3.1.2 Exploring Composite Difference? 26

3.1.3 Efficacy of Using Composite Differences to Identify Meteorological Patterns 27

3.1.4 Patterns Determined from Composite Differences 29

3.1.4.1 *maxuh*03 29

3.1.4.2 *maxuh*25 31
3.1.4.3 *u*500 34
3.1.4.4 *mucape* 37
3.1.4.5 Case Study of High and Low Importance 40
3.1.4.6 Summary of Composite Differences 48
3.2 Impacts of Flow Dependent Training 49
3.2.1 Domain Average Classification 49
3.2.1.1 BSS Comparison to NFD Model 49
3.2.1.2 AUC, Reliability Curves, and Model Diagnosis 51
3.2.1.3 Subjective Differences between NFD Model and FD Model 54
3.2.2 Spatial Pattern Classification 63
3.2.2.1 BSS Comparison to NFD 63
3.2.2.2 Reliability and Resolution 64
3.2.2.3 Comparison of Both Methods 65
3.3 Impacts of Sample Size on Model Performance 67
3.3.1 Combined BSS Comparison 67
3.3.2 Reasons for Differences 69

4 Conclusion **71**
4.1 Summary of Work and Findings 71
4.2 Future Work 73

References **75**

List Of Tables

2.1 All possible predictors used for the creation of the model. The predictors used in the creation of the RF model are represented in bold text. . . 8

2.2 Hyperparameters used in the creation of the RF for this study. 10

3.1 The Reliability, Resolutions, and AUC for each PS Model. 54

3.2 The Reliability, Resolutions, and AUC for each PC Model. 65

List Of Figures

2.1 A physical representation of the NWP forecast domain and the regions used for training and verification shaded in green. 7

2.2 All 2021 HWT Domains and their locations throughout the sample cases. The plotted contours are the location in which averages/any computations will be re-centered and plotted for a consistent place of reference while still representing of the spatial scale of the HWT Domain. 12

2.3 CAPE/shear domain average parameter space values for all 29 of the 2021 cases. Additionally, the two thresholds those being the lower being extreme low shear, and upper being extreme high shear for PS classification with marginal cases in the middle. 15

2.4 The EOFs of (a) PC-1 of shear, (b) PC-1 of SBCAPE, (c) PC-2 of shear, and (d) PC-2 of SBCAPE. 17

2.5 A sample (a) ROC curve plot and (b) reliability plot. The ROC can be used to determines how good a model is a detecting severe events and discriminating severe events from null events. The reliability plot shows how well calibrated the model's output probability is compared to observed relative frequency in each forecasted probability. 21

2.6 An example of a single predictor (in this case, 0-1 km helicity) contribution to this case forecast. Red values indicate positive (increasing probability) contribution while blue values indicate negative (decreasing probability) contribution. 23

3.1 The resulting BSS for each iteration of the FPS. The red horizontal line was the model BSS if all predictors were used to forecast the cases. Notice that iterations 4-6 decreased or maintained the same skill before increasing skill above the full model. 26

3.2 PFI values for all predictors used within the NFD Model. It's clear that $maxuh03$, $maxuh25$, $u500$, lat, and $mucape$ are the most important predictors based on PFI. 28

3.3 (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $maxuh03$. 30

3.4 (a) $u500$ wind composite and (b) $v500$ wind composite for the predictor of $maxuh03$. 31

3.5 (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $maxuh25$. 33

3.6 (a) $u500$ wind composite and (b) $v500$ wind composite for the predictor of $maxuh25$. 34
3.7 (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $u500$. 36
3.8 (a) $u500$ wind composite and (b) $v500$ wind composite for the predictor of $u500$. 37
3.9 (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $mucape$. 39
3.10 (a) $u500$ wind composite and (b) $v500$ wind composite for the predictor of $mucape$. 40
3.11 The (a) high importance and (b) low importance cases output probability from the NFD model. 41
3.12 The (a) high importance $maxuh25$ maximum value from the 10 member ensemble and (b) low importance maximum value from the ten-member ensemble cases output probability from the NFD model. 43
3.13 The low importance case (a) 500 mb heights, (b) SBCAPE environment. 45
3.14 The low importance case (a) u500, (b) v500 environment. 46
3.15 The high importance case (a) 500 mb heights, (b) SBCAPE environment. 47
3.16 The high importance case (a) u500 and (b) v500 environment. 48
3.17 The BSS for the (a) PS shear and (b) PS CAPE model. A single SD is plotted for clarity. 51
3.18 On the left, the (a) Low Shear Reliability and (c) ROC curves. On the right, the (b) High Shear Reliability and (d) ROC curves. 53
3.19 (a) The SPC 1630 UTC Outlook overlayed with storm reports and (b) the 1630 UTC SPC Discussion for the 14 May 2021. 56
3.20 FD model output for 13 May 2021. Probability covers the storm report region well, but note a locally higher probability region on the southern portion of the domain. 57
3.21 Two different iterations of the NFD model baseline and the probability generated for each iteration. 58
3.22 Two more different iterations of the NFD model baseline and the probability generated for each iteration. 59
3.23 (a) The FD model FI, (b) the average FI of the NFD model. 60
3.24 The difference between the FD model FI and the average FI of the NFD model. More emphasis was place on low-level shear parameters and less on instability parameters, which reflects what you would expect in a Low CAPE setup. 61
3.25 (a) The 0-1km Helicity Contribution to the FD model, (b) the average contribution of 0-1km Helicity to the NFD model. 62

3.26 The difference between the two. This is the largest difference in contribution between the NFD and FD model, and helps increase probabilities over the storm report region and decrease it to the NW and N of the reports region. 63
3.27 (a) BSS plot for shear PC model and (b) BSS plot for the CAPE PC model. 64
3.28 A comparison of both the PS and PC Models using the Combined BSS as mentioned earlier. No FD model has a clear advantage over the other, but all FD models are statistically significantly better than the baseline. For each model, one SD is plotted for clarity. 67
3.29 A comparison of both the PS and PC Models using the Combined BSS as mentioned earlier on the larger sample size. PS Model remain statistically significantly better than the baseline, but all PC Models (except for PC1 Shear) do not remain statistically better than the baseline. 69

Abstract

There has been an increasing interest over the past ten years in the use of Machine Learning (ML) algorithms such as Random Forests (RF) in the context of severe weather prediction. RF-based methods have even been shown to outperform human-generated operational convective outlook guidance in some cases. However, there remain obstacles to fully integrating the ML algorithms into the operational forecasting process of severe wind, hail and tornado events. For example, the perceived black-box nature of complex RF models can inhibit forecaster confidence in the ML guidance for high impact or atypical events. Since the error characteristics of predictors based on numerical weather prediction, or NWP, and the relationships between these predictors and severe weather risk can vary in different flow patterns, there is a need to better understand the impacts of large-scale flow patterns on RF model performance. In addition to improving confidence in the RF-based forecast products, such understanding can also be incorporated into the model building process to further improve their performance.

This thesis discusses the development and evaluation of a flow-dependent approach to training RF models to produce severe weather convective outlook guidance. This work leverages 53 real-time cases from the 2019 and 2021 real time convection-allowing FV3-based ensemble forecasts produced by the University of Oklahoma (OU) Multi-scale data Assimilation and Predictability (MAP) Lab during the 2021 Hazardous Weather Testbed (HWT) forecasting experiments as model predictors. This study will focus mainly on the 29 cases from 2021. As a first step, the composite difference in large-scale flow between cases with relatively high and low importance of key predictors using Permutation Feature Importance were calculated. These composite differences were used to evaluate if discernible large scale flow patterns could be when the non-flow dependent model would perform the best. Two different methods of classifying cases based on the large-scale flow patterns are then evaluated for the purpose of training separate RF models on cases of similar flow patterns. The appropriate RF model for the pattern is then used to generating convective outlook guidance for a forecast case not included in the RF model training. First, the CAPE/shear parameter space over the region of interest is used as a classification metric. Second, EOF patterns that are qualitatively similar to the previously described composite flow patterns related to predictor importance and used as the classification metric. Finally, both methods will check how sensitive the performance is to changes in sample size by adding the 2019 cases. Both flow-dependent training methods will be compared to the non-flow dependent models and compared to each other. Results will emphasize both objective impacts on forecast skill and physical explanations of the difference in performance among the RF training approaches. Results show that both methods of flow-dependent training initially show improvement in forecasting severe weather compared to the non-flow dependent model. However, the parameter space classification remain to show significant skill with an increase in sample size while not all EOF patterns skill remained significant.

Chapter 1

Introduction

Severe local storms are the most frequent meteorologically-related disaster in the United States (NOAA 2017). From 2011 to 2021, severe thunderstorms and tornadoes have cost the United States (US) $26 billion on average each year (Munich 2023). Furthermore, the United States annually averages about 113 deaths and roughly 1,265 injuries from severe thunderstorms and tornadoes (SPC 2023). The accuracy of severe weather forecasts is important in the preservation of both life and property. Despite inherent predictability limits of individual convective storms (Melhauser and Zhang 2012; Hohenegger and Schar 2007), numerical model guidance from ensembles of convection-allowing models (CAMs) play an important role in operational severe weather prediction. CAM ensembles (CAEs) are useful because they are able to provide probabilistic forecast guidance based on realistically simulated convective systems (Mass et al. 2002; Weisman et al. 2008; Clark and Loken 2022). However, the performance of CAE-based severe weather forecast guidance is limited by an inability to directly resolve the severe hazards (Clark and Loken 2022), the presence of systematic model/physics errors and ensemble underdispersion (Vié et al. 2011; Romine et al. 2014; Schwartz et al. 2014), and an incomplete understanding of why some storms produce severe weather and others do not (Roebber 2009; Anderson-Frey et al. 2016).

Many studies have used Machine Learning (ML) techniques together with CAEs to produce skillful forecasts of tornadoes (Marzban and Stumpf 1996; Alvarez 2014; Sobash et al. 2016; Gallo et al. 2018; McGovern et al. 2019), hail (Marzban and Witt 2001; Brimelow et al. 2006; Adams-Selin and Ziegler 2016; Gagne et al. 2017; McGovern et al. 2019; Burke et al. 2020), and wind (Marzban and Stumpf 1998; Lagerquist et al. 2017) to improve probabilistic severe weather forecasts based on CAEs. One ML technique that has gained a lot of popularity for severe weather forecasting is the Random Forest (RF) (Ahijevych et al. 2016; Hill et al. 2020; Loken et al. 2020). It has been demonstrated that the probabilistic forecasts from RF lead to either comparable (Hill et al. 2020) or better skill (Gagne et al. 2017) to Storm Prediction Center (SPC) forecasts for different types of severe weather. One reason why the RF is chosen frequently in meteorology is that it is relatively easy to interpret compared to other ML methods (Herman and Schumacher 2018; Loken et al. 2020). Another advantage of RF is its ability to identify spatiotemporal relationships and relationships across different predictors that we may not observe (Herman and Schumacher 2018) but we can infer from the trained RF model (Burke et al. 2020; Clark and Loken 2022). The RF approach also allows the resolved model variables to be directly related to unresolvable processes related to the occurrence of severe weather (Clark and Loken 2022). The RF approach can account for the systematic bias of models and mitigate ensemble underdispersion (Loken et al. 2020).

Many previous studies have used one of two methods of subsetting the available training data for training of the RF. These two common methods are training based on region or seasonally (Hill et al. 2020; Loken et al. 2020; Burke et al. 2020). The training samples used for these methods came from either the specific region or season that the RF was used to forecast. This subsetting of training samples is done because of different temperature and/or precipitation climates in different regions and seasons (Bukovsky 2011) and different systems being responsible for severe weather in different regions and seasons (Smith et al. 2012). Therefore, it can be expected that training models based on regions/season can account for different model biases or ensemble underdispersion as well as represent different relationships among variables in these different patterns. When compared to SPC forecasts, general results show that the RF models trained for the central (or midwest) US tends to have the highest forecast skill, with the worst forecast skill region being the Western US (Loken et al. 2020). The combined/aggregate model, or the version that takes the RF of different regions/seasons and combines their average forecasting statistics, of the regional model is similar to SPC forecasts (Hill et al. 2020). The two seasons that produce the poorest results are summer and fall (Hill et al. 2020; Loken et al. 2020). The combined version of the model averaging performance over all regions performs well even with these two seasons performing poorly.

While the previous studies have shown competitive skill of severe weather forecasts with regional and seasonal RF model training, compared to SPC forecasts, some aspects of the region/season-dependent training were not considered. The impact of forecast skill of a specific region/season was not compared against a model trained on all samples. Additionally, systematic model errors and the relationships between severe weather and predictors may also depend on the large-scale flow, which may vary across a single region or within a season. For example, during northwest (NW) flow patterns the RF might benefit from placing more emphasis on different predictors than a traditional Southwest (SW) Flow pattern. Herein, one can hypothesize that an RF model can effectively account for differences in systematic model errors and relationships between severe weather and model predictors by training the RF based on cases with a similar large scale flow pattern.

I will introduce a new approach to determine if RF model training based on the large-scale flow pattern can improve upon RF-based severe weather forecasts trained independently of the flow pattern. Specifically, different methods will be used to classify the large scale flow pattern based on domain average values and spatial patterns of convective available potential energy (CAPE) and bulk wind shear. The flow dependent models are trained separately based on whether or not a case meets the corresponding pattern. The goals of this study are to determine if a flow-dependent RF model can outperform a non-flow dependent model and compare different methods of quantifying the flow pattern.

Chapter 2

Data and Methods

In this chapter, I will introduce the dataset used during this work as well as the region within the CONUS that the forecast are verified with in Section 2.1. I will then discuss the creation of the RF model, predictor selection method, and hyperparameters chosen in Section 2.2. Then, I will introduce the different classification of flow for the creation of the flow-dependent model training in Section 2.3. Finally, I will discuss the different diagnosis tools and methods that I selected for this work to give context of how the research questions were answered in Section 2.4.

2.1 Data Selection and Preprocessing

In this section, I will explain the dataset in detail, including the forecasting domain and where the data comes from. Additionally, I will explain the preprocessing of predictors to show what their inputs will be for the RF model. Finally, I will discuss each of the predictands to give an idea of what each of the models will try to predict.

2.1.1 Ensemble Forecast Model

The numerical weather prediction model used for this study was the University of Oklahoma's Multi-scale data Assimilation and Predictability (MAP) FV3 Convection-Allowing Data Assimilation and Ensemble Forecast System run in real time during the 2021 Hazardous Weather Testbed (HWT) (Gasperoni et al. 2022). This model is a 10-member ensemble forecast at 3-km grid spacing over the entire continental US (CONUS) domain. However, for the purposes of this experiment, the domain labeled in green in Figure 2.1 below was the region used for forecast training and verification. The main reasoning for focusing on the green region is to include climatologically more frequent severe weather compared to the Western United States (SPC 2023), to avoid ocean and international regions in which severe reports are unavailable, and to not have adverse effects from the Rocky Mountains. The cases used in this study consist primarily of the cases from the 2021 HWT, though additional forecasts have since been run retrospectively during the 2019 Spring season. The 2021 cases span the period of April 12 through June 4 for a total of 29 cases. Thirty six-hour forecasts were initialized at 0000 UTC and the hourly outputs at one-hour intervals starting at forecast hour 12 were used to produce the predictors for the RF. This study will focus on severe weather probability over the 1200 UTC to 1200 UTC 24-hour forecast period.

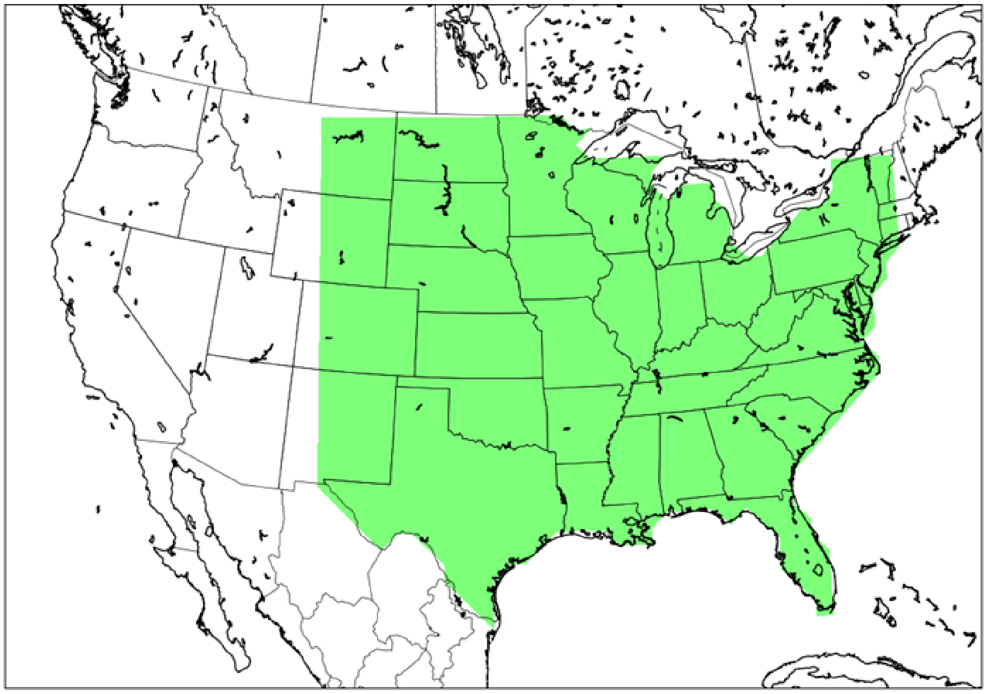


Figure 2.1: A physical representation of the NWP forecast domain and the regions used for training and verification shaded in green.

2.1.2 Preprocessing of Predictors

The native 3-km grid was remapped to an 80-km grid following Loken et al. (2020) in order to fit all predictors onto a similar grid to match the probabilities of an SPC outlook. The model variables selected to generate predictors are loosely categorized into 3 distinct groups (Table 2.1), intended to primarily represent synoptic, mesoscale, and storm-scale features. The large-scale predictors output are the ensemble average of the temporal average value over the forecast period to account for the movement of these synoptic features over the forecast period. For example, the $u500$ windspeed would be the ensemble average of the average strength of the u500 winds over the 24-hour forecast. The mesoscale forecast variables ouput are the ensemble average of the maximum/minimum value over the entire forecast period to account for the most favorable mesoscale environment for severe weather during the period. For example, for the output predictor of *sbcape*, the temporal maximum of CAPE is calculated for each ensemble member, then the ensemble average of this most favorable mesoscale environment is used as the predictor. Finally, the storm-scale variables output the maximum ensemble value of the maximum value over the forecast period to see if the models are initializing storms at any time during the valid forecast period and if so, how intense the storms will be based on the model. For example, the maximum updraft helicity in the 2-5 km region would be determined by the max ensemble member's value over the entire 24-hour forecast period. The resulting predictors are similar to previous studies (Hill et al. 2020; Clark and Loken 2022), with a combination of all predictors from previous studies to forecast severe weather.

Table 2.1: All possible predictors used for the creation of the model. The predictors used in the creation of the RF model are represented in bold text.

Large-scale	Meso-scale	Storm-scale
u/v500, 850, 250 (**u250,v250,u500**)	**SBCAPE**, SBCIN	**maxuh25, maxuh03**
Tmp 2m, 850, 700, 500	**MUCAPE, MUCIN**	hrprecip
dew 2m, 850, 700, 500	SB LCL	
Change in z500	**1km helicity**, 3km helicity	
pw		
lat, lon		

2.1.3 Predictands

The predictands, or what the model is trying to predict, for this study were obtained by using the reported severe weather from SPC filter storm reports database and remapping them to the same 80-km grid of an SPC outlook (based on their probabilities). For this study, we treat each grid point as a binary value. A value of one would indicate at least one severe report occurred during the 24-hour period. A value of 0 would indicate that no severe weather was reported within this grid point during the 24-hour period. The goal of mapping the predictands to an 80-km grid is to have the model be able to predict these predictands similar to an SPC convective outlook, which is a probabilistic forecast of any severe weather occurring within 25 miles (40 km) of any given point, which is similar to predicting the probability severe weather occurs within an 80-km grid box.

2.2 Random Forest Model Creation

In this section, I will go in-depth behind the making of the RF model for this study. I will explain how a RF model works, how the predictors are selected for the model, and what the hyperparameters are for the model in this study. Additionally, I will explain how the selection method for this study varies compared to normal applications of that method. Finally, I will present key components of how the model generates a probabilistic forecast.

2.2.1 Random Forest Description and Selected Hyperparameters

A Random Forest is an ensemble of decision trees, where each decision tree is trained on a random subset of the training data. Each decision for a branch of the tree uses a predictor randomly selected from a subset of the predictors seen in Table 2.2. The maximum depth of the tree can be limited as seen in Table 2.2. Eventually, each tree ends with a leaf, which must contain a minimal number of samples as seen in Table 2.2. In this study, a forest size of 1000 trees was used, (2.2) similar to other studies (Hill et al. 2020). The RF model is trained using the leave-one-out technique, wherein the model is trained on all other cases but one and is used to forecast on the remaining case. The leave-one-out training is repeated for all cases within the training set (Bhuiyan et al. 2019). Furthermore, other hyperparameters were determined loosely based on values used in similar studies (Clark and Loken 2022). While a comprehensive optimization testing was not conducted, limited sensitivity tests show that model performance or skill was minimally impacted if values in the table were changed, which is consistent with other studies (Clark and Loken 2022).

Table 2.2: Hyperparameters used in the creation of the RF for this study.

Hyperparameter	**Value**
Number of trees	1000
Maximum depth	10
Predictors used at each branch	$\sqrt{\# of\, predictors}$
Fraction of training samples used for each tree	0.25
Minimum leaf size	30

2.2.2 Predictor Selection

The Forward Predictor Selection (FPS) method is similar to the Ließ et al. (2016) method used to select predictors for the RF. The FPS consists of first training a model with a single predictor, using each predictor in turn. The predictor that provides the highest BSS is retained. Then, a new set of models is trained with the retained predictor(s) plus one additional predictor from the remaining predictors. Then the next predictor added is the one that provides the highest improvement in BSS. This process repeats until a threshold value of improvement in skill is not met (BSS does not improve by 0.001 or more). However, unlike what was done in Ließ et al. (2016), additional iteration(s) are conducted after reaching the stopping condition because of how much poorer the model was compared to a model trained on all predictors. Continuing more iterations led to models that were better (had higher BSS) than the model trained on all predictors. In other words, the stopping condition was required to be met in more than one iteration in a row in order to stop adding predictors to the model. While arguably not necessary to optimize performance of the RF, FPS was used to increase the interpretability of the model by simplifying the number of predictors included without loss of skill. RF may even benefit from the exclusion of redundant or unnecessary predictors which can only add noise to the predictions (Hall et al. 2011; Ahijevych et al. 2016). Although the difference is not likely statistically significant in our study, this potential benefit of FPS may explain the slightly higher skill than the model trained using all possible predictors

2.3 Flow-Dependent (FD) Model Training

The goal of this study is to test the two methods of classifying large scale flow patterns, I explored both the Domain Average Classification and the Spatial Pattern Classification methods. Both methods were calculated based on a regional domain selected by HWT participants as regions of the greatest severe threat during the day of the forecast (Figure 2.2). Each of these methods is explained in respective sections. Additionally, I will discuss a non-flow-dependent control experiment (explained later in this section) was used as a baseline for comparison. The two flow-dependent models will be compared against the baseline to determine the benefits of training the RF on cases with similar flow patterns compared to not using a flow-dependent training.

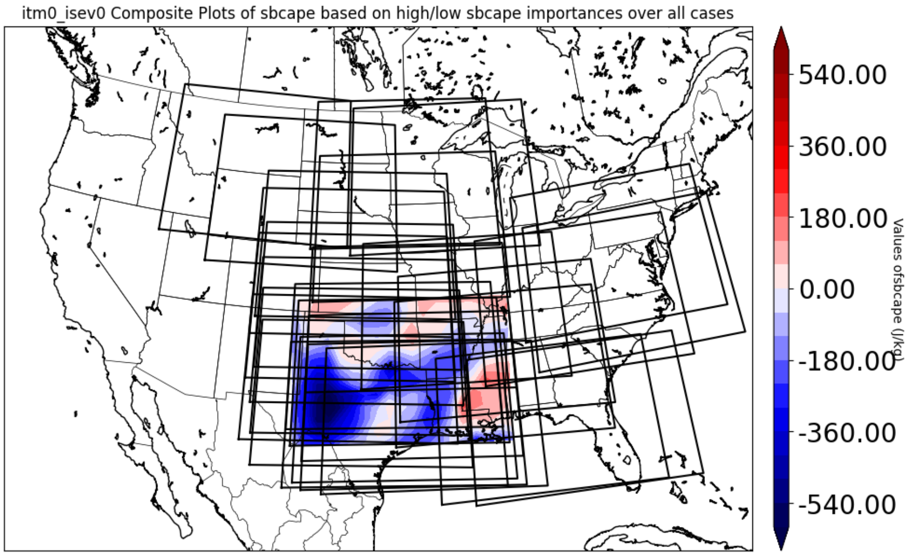

itm0_isev0 Composite Plots of sbcape based on high/low sbcape importances over all cases
540.00
360.00
180.00
0.00
-180.00
-360.00
-540.00
Values ofsbcape (J/kg)

Figure 2.2: All 2021 HWT Domains and their locations throughout the sample cases. The plotted contours are the location in which averages/any computations will be re-centered and plotted for a consistent place of reference while still representing of the spatial scale of the HWT Domain.

2.3.1 Domain Average Classification

The first method of flow-dependent training of the RF is based on two widely used parameters used in severe weather forecasting today. In particular, two parameters used frequently to obtain a first-order estimate of the possibility of severe weather are both CAPE and 0-6 km bulk shear. The CAPE parameter is related to the possible strength of the updraft and resulting strength of the convection (Romps 2016). However, CAPE by itself does not determine the potential for storm organization. Shear affects the organization of thunderstorm updrafts by the separation of the downdraft from the updraft. More separation of the updraft/downdraft leads to a more organized storm and potentially a stronger storm (Weisman and Klemp 1982). While other factors (forcing for ascent, other mesoscale features, etc) are important ingredients to forecast severe weather, the average CAPE (mesoscale predictor) and shear (synoptic predictor) over the HWT domain is used to quantify aspects of the large scale flow that are important for severe potential. The values from the individual forecast cases can be seen in Figure 2.3.

As an initial proof of concept on the impacts of training separate RF models for different categories of large scale flow patterns, two categories of the eight highest and lowest CAPE/shear cases are selected. For each of the extreme cases, or case that do meet the criteria to be classified in a flow pattern, the RF model is trained on the other cases within the same category (cases that also were extreme high or extreme low CAPE/shear). The cases that have only marginally high or low CAPE/shear are not included in the FD training. To get distinct groups, the eight highest and lowest cases are used for the training and forecast evaluation of the RF. The reason behind this is that a case on either side of a single threshold could be more similar to each other than cases in the same category. This can be seen in Figure 2.3 below, where the extreme low and extreme high thresholds are drawn. The cases to the left of the lower bound is extremely low shear in this case and the cases to right of the higher threshold is extreme high shear. The marginal cases are defined as any cases in between the two thresholds. This experiment is intended as a proof of concept of the idea of training only on cases that are unambiguously similar domain average CAPE or shear. For reference throughout the rest of the paper, these models will be called parameter space (or PS) models.

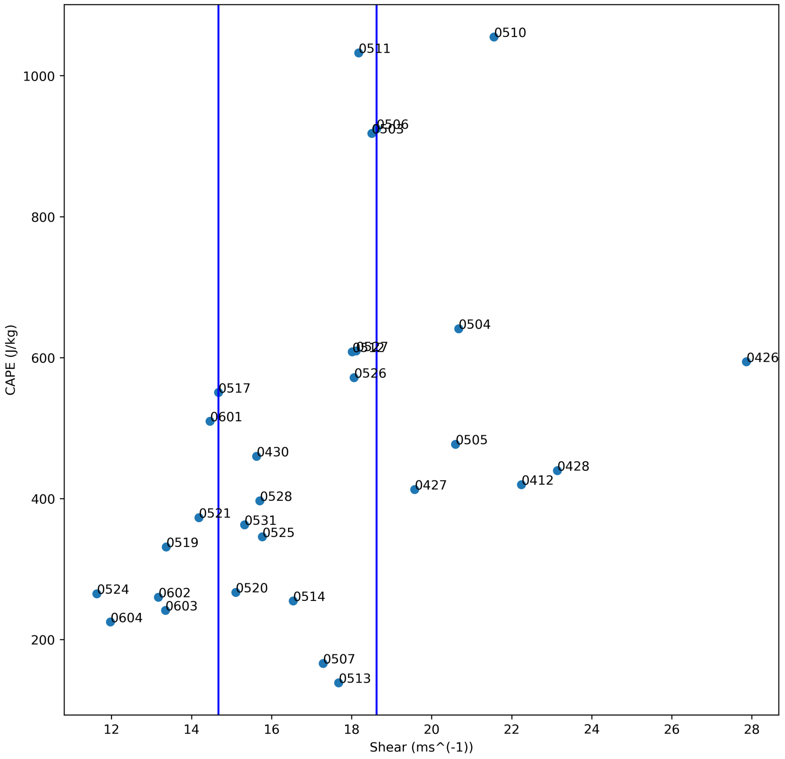
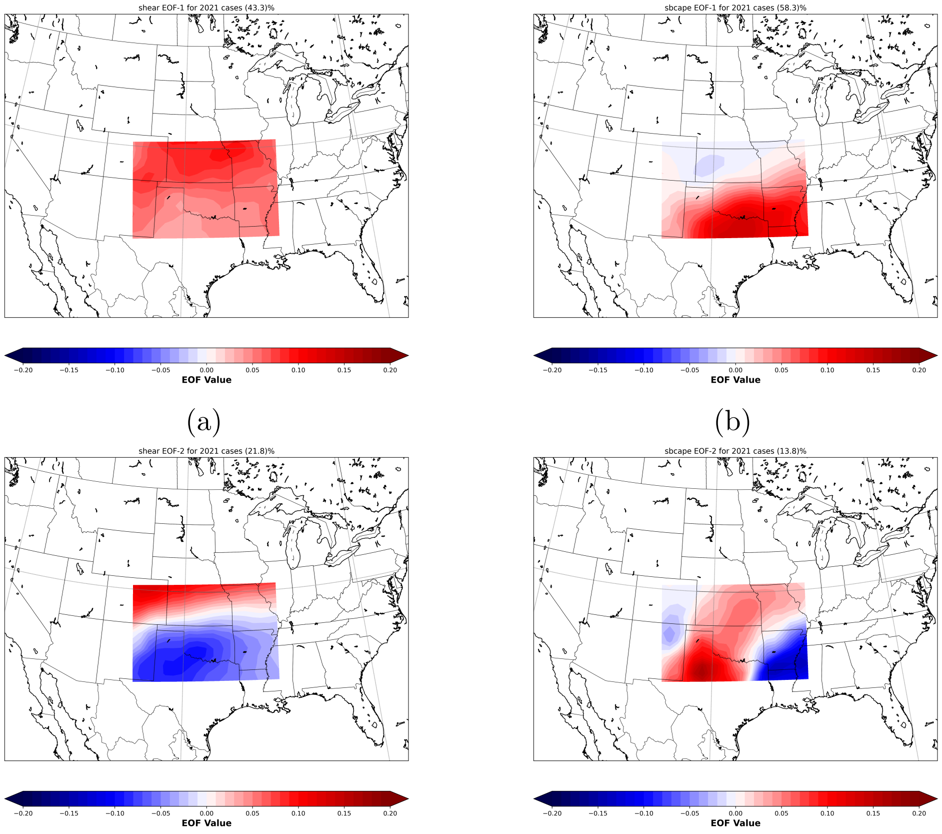


Figure 2.3: CAPE/shear domain average parameter space values for all 29 of the 2021 cases. Additionally, the two thresholds those being the lower being extreme low shear, and upper being extreme high shear for PS classification with marginal cases in the middle.

2.3.2 Spatial Pattern Classification

It has been shown in other meteorological studies that Principal Component Analysis (PCA) can identify patterns within a dataset that explain the majority of the variance within the dataset. Consequently, this study uses PCA to identify spatial patterns (Empirical Orthogonal Functions or EOFs) in both CAPE and shear when quantifying the large scale flow in terms of CAPE and shear in addition to using a domain average value of CAPE/Shear as stated in the previous section. For this study, the two leading modes of variability in both CAPE and shear are used to classify cases and their scalar projection onto these EOFs will be referred to as the principal component value of the case (or PC).

Figure 2.4 shows the leading modes of variability of shear (left column) and CAPE (right column). PC1 of shear shows a higher amount of shear over the entire domain, particularly over the northern portions of the domain, for cases with a positive value of PC1 (Figure 2.4a). PC2 for shear represents shear that is generally lower over the majority of the domain while shear is generally higher over the northern third of the domain, for cases with a positive value of PC2 (Figure 2.4c). Positive values of PC1 of CAPE indicate a higher amount of CAPE over the HWT period except for a bimodal decrease over the northern portion of the domain (Figure 2.4b). Positive values of PC2 of CAPE shows a narrow warm sector, with the highest instability over the southwest corner of the domain (Figure 2.4d). Similarly to excluding the marginal cases for classification based on the domain average value, each PC is used to identify cases that project strongly positively (G1) or strongly negatively (G2) on the corresponding PC. Similar to the classification based on domain average values, the marginal cases are excluded from consideration when evaluating FD model used for cases above the high-threshold or below the low-threshold. However, in order to include the same set of all cases in a skill score for comparing different flow pattern classification methods, the marginal cases are still forecast using a RF model trained on a random set of cases for both the FD and non-FD skill. These models for this study will be reference as the PC models.


shear EOF-1 for 2021 cases (43.3)%
sbcape EOF-1 for 2021 cases (58.3)%
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
EOF Value
shear EOF-2 for 2021 cases (21.8)%
sbcape EOF-2 for 2021 cases (13.8)%

(a) (b) (c) (d)

Figure 2.4: The EOFs of (a) PC-1 of shear, (b) PC-1 of SBCAPE, (c) PC-2 of shear, and (d) PC-2 of SBCAPE.

2.3.3 Control Experiment (Randomly Trained Model)

The purpose of the experiments described in Sections 2.3.1 and 2.3.2 is to determine if the flow dependent RF model forecasts severe weather better than a non-flow dependent RF model. Therefore, a control model is trained on the same number of cases as the model that it is compared against (e.g., eight cases for the high shear model), where the training cases are randomly selected with replacement from the full set of 29 available cases and repeated 27 times. The purpose is to not gain the best possible skill for this type of model, but rather to have a fair model of comparison to see if the training cases that are selected can cause models to forecast severe weather within certain flow patterns better or worse, while controlling the training sample size. This approach also important to consider the impacts of sample size as a way to see if these results are sensitive to small or large sample sizes.

2.4 Diagnostic and Verification Techniques

In this section, I will explain the different diagnosis tools used to determine the model performance. This study uses a variety of plots and scores to determine model skill compared to both a reference forecast and how accurate the forecast is in general. Additionally, I opted to use other tools to determine how important predictors are and how much they contribute to the model performance. Feature Importance was used primarily on the flow-dependent models to quickly compare differences in predictor importance compared to the much slower, albeit more rigorous, permutation feature importance. Since FI is characteristic of the trained model, it is suitable for comparing what predictors different models are using but it is not suitable for evaluating differences in how each predictor contributes to RF performance on different groups of cases. Permutation Feature Importance was ran primarily on the non-flow dependent model to get a rigorous idea of which predictors were importance and in what cases to then further investigate with composite differences. Tree Interpreter purpose is to visualize on specific cases the contribution of predictors and how that looks compared to the normal output of the RF. Therefore, TI can provide physical understanding and qualitative interpretation of case studies that cannot be obtained from FI or PFI. Finally, I will explore the contribution of individual predictors to the forecast on a gridpoint basis to see what meteorological factors are impacting the forecast.

2.4.1 Brier Score, AUC, Reliability

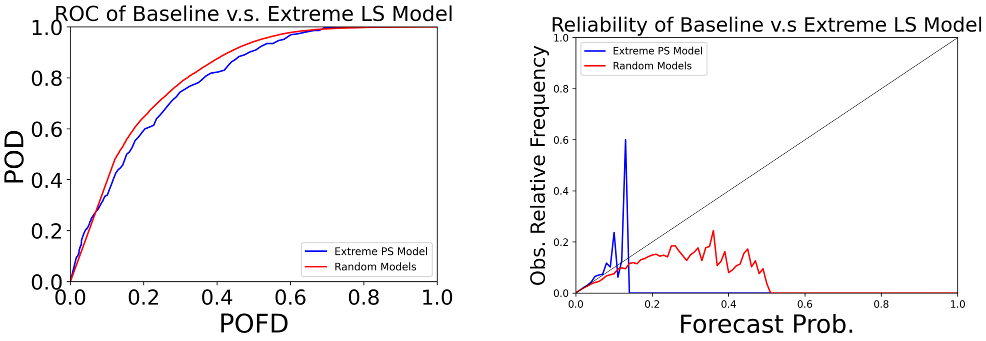
Numerous studies use Brier Score as a way to diagnose the Root Mean Square Error (RMSE) of a model's probabilistic forecast (Berner et al. 2015; Hill and Schumacher 2021; Ahijevych et al. 2016; Hill et al. 2020; Loken et al. 2020). A breakdown of the Brier Score can be seen in the Equation 2.1 where k is the forecast bin, p_k is the forecast probability of the bin, $\bar{o}_k$ is the observed frequency within the bin, and N is the total number of samples and n_k is the number of forecasts in bin k. Finally, $\bar{o}$ is the overall frequency through all k forecast bins. The reliability term is interpreted as the error in the calibration of the forecast probabilities. The resolution shows how the different values of forecast probability resolve different conditional probabilities of the event occurring. Unlike the reliability component, a larger resolution component decreases the overall Brier Scores, indicating a more accurate forecast. Finally, the uncertainty term is just a function of the mean climatology over the verification period. This term is a function of the event occurrence on the cases being forecast and does not depend on the forecast itself.

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}) \tag{2.1}$$

The Brier Skill Score (BSS) is used to quantify reduction in forecast error compared to a reference forecast (Berner et al. 2015). The equation for the Brier Skill Score can be seen in Equation 2.2. While low BS indicates better forecasts, higher BSS indicates better forecast compared to the reference forecast. This will be one of the main forecasting skill metrics used to compare models against each other in this study. One important note is that to make comparison of different flow dependent model training methods, an equal comparison is needed in addition to the BSS over one group. To achieve a fairer comparison of the different FD methods, both BS and BS_{ref} are calculated over all cases, including the marginal cases which the forecasts use the randomly-trained model for both BS and BS_{ref}.

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{2.2}$$

Two more diagnostic tools that are used in this study are the reliability curve and ROC plots. The ROC plots (Figure 2.5a) show on the x-axis the Probability of False Detection (POFD) and on the y-axis the Probability of Detection (POD). This can demonstrate the model's ability to discriminate an event as a severe event or as a null event as well as indicate the model's ability to detect severe events. The area under the ROC curve (AUC) generates a value between zero and one with 0 meaning the model cannot detect any severe events and 1 meaning the model detects every severe event in the training data. The reliability curve plots shows on the x-axis the model's forecast probability and on the y-axis the observed relative frequency of severe events within that forecast probability. The reliability plot visualizes if the model is well calibrated by if the forecast probability is equal to the observed relative frequency. If the forecast probability does equal the observed relative frequency, then the model is well calibrated as seen by the diagonal line in in Figure 2.5b. If the model forecast probability is lower than the observed relative frequency, the model has an underforecasting bias, while if the model forecast probability is higher than the observed relative frequency, the model has an overforecasting bias.



(a)

(b)

Figure 2.5: A sample (a) ROC curve plot and (b) reliability plot. The ROC can be used to determines how good a model is a detecting severe events and discriminating severe events from null events. The reliability plot shows how well calibrated the model's output probability is compared to observed relative frequency in each forecasted probability.

2.4.2 Feature Importance and Permutation Feature Importance

Feature Importance (FI) is a way to see the entropy gained at each branch by using the data the model is trained on (McGovern et al. 2019). The FI can be interpreted as how effective a predictor is from separating severe occurrences from non-severe occurrences, or how much a predictor decreases impurity (McGovern et al. 2019). In this study, FI will be used on both the FD and baseline models to see how important each predictor is, or how good the predictors are at separating the data. Specifically, the goal of using FI is to demonstrate how the predictor importance changes from the random model to the FD model to understand if the different cases used to train the models lead to different relationships among predictors being represented by the RF.

Permutation Feature Importance (PFI) is different from FI as it tests how the performance of the model changes when the statistical connection between predictors and predictands are broken (McGovern et al. 2019). The PFI tells us that if the model performance does not worsen significantly, then that predictor is considered generally not important. If model performance does begin to worsen, then that predictor is considered important (McGovern et al. 2019). The PFI is used in this study to determine what features are important in the non-flow dependent model to increase skill, then determine on a case-by-case basis which predictors have relatively high or low importance on different cases. This will allow us to visualize differences in the flow patterns on cases with relatively high and low importance of specific predictors to potentially give ideas of how to classify flow patterns.

2.4.3 Tree Interpreter

While there are many good methods to evaluate how a model understands predictors, a method to visualize how predictors contributed to the forecast is preferred. A Python package called Tree Interpreter (Saabas 2023) used in other studies (Loken et al. 2022) has shown to be very useful in understanding how certain predictors contribute to a forecast probability in a given case. Specifically, the Tree Interpreter package reports the changes in probability within a tree resulting from each branch based on a given predictor, then it is the average contribution from all trees over multiple validation samples (Loken et al. 2022). As shown in Equation 2.3, the forecast probability is the sum of the contributions from all predictors plus a bias term. The bias term is the base rate of the forecasted variable before any predictors begin to change the probability. You then add the contributions from your features (different predictors) as seen in Equation 2.3. An example plot of a single feature contribution can be seen below in Figure 2.6. Red values indicate positive contributions to the forecast probability and blue indicates negative contributions to the grid point forecast.

$$predicton = bias + \text{feature}_1\,contribution + ... + \text{feature}_n\,\text{contribution} \qquad (2.3)$$

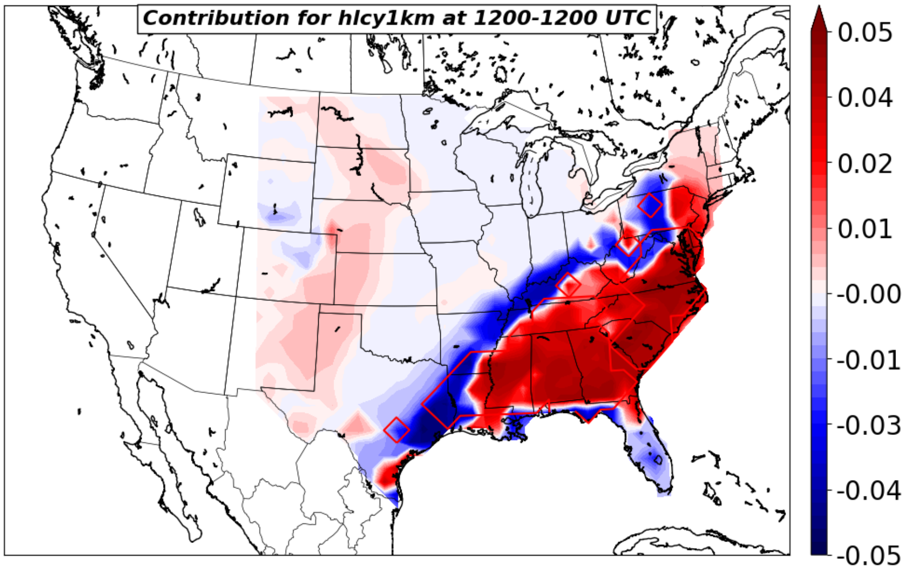


Figure 2.6: An example of a single predictor (in this case, 0-1 km helicity) contribution to this case forecast. Red values indicate positive (increasing probability) contribution while blue values indicate negative (decreasing probability) contribution.

Chapter 3

Results

In this chapter, I will go through the creation and evaluation of the non-flow dependent model. I will also see how well as how predictors were selected for this model while seeing if certain flow patterns lead to good or poor performance of the model in Section 3.1. Then, I will explore the diagnosis of the first flow dependent classification method (Section 3.2), the PS model. I will then discuss the second flow dependent model classification, the PC model, as compared to the PS model (Section 3.2. Finally, I will demonstrate the sensitivity of both methods to sample size by adding more cases to train the models on in Section 3.3.

3.1 Evaluation of Non-Flow Dependent Model

In this section, I will explore the creation of the non-flow dependent model. Section 3.1.1 the performance of this model when predictors were added through the FPS method. Furthermore, I will explore the cases in which certain predictors are important and compare them to cases which those predictors are not important. This will give motivation to the flow-dependent model training discussed in Sections 3.2.

3.1.1 BSS Comparison of Full Model to FPS Model

Before evaluating the differences in performance between flow dependent (FD) and non-flow dependent (NFD) models, an NFD model was built using all available cases within the 2021 training cases. The Forward Predictor Selection (FPS) is used on this NFD model and will be the basis of predictors for all models in this study. Figure 3.1 below shows that although at iteration three the BSS did not improve, adding more variables allowed for further improvement to the NFD model. When multiple iterations did not increase BSS, the resulting model has slightly higher skill than a model using all predictors. The resulting model will be used as a baseline for the majority of the resulting experiments, with only changing the case sizes to allow for equal comparisons of models.

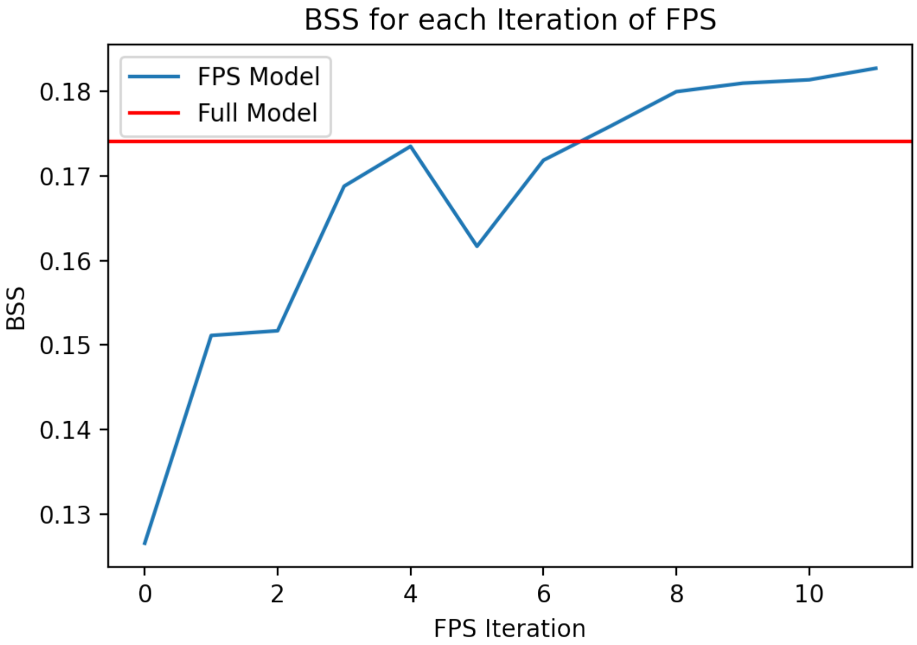


Figure 3.1: The resulting BSS for each iteration of the FPS. The red horizontal line was the model BSS if all predictors were used to forecast the cases. Notice that iterations 4-6 decreased or maintained the same skill before increasing skill above the full model.

3.1.2 Exploring Composite Difference?

Using the PFI, we can get an idea of spatial patterns that occur when a variable is or is not important for that model. A composite difference in this case is the spatial pattern difference between the cases in which a predictor has relatively high importance and the spatial pattern of cases when a predictor has relatively low importance as seen in Equation 3.1. In this equation, we focus on cases of high importance of 500 mb heights and take a difference between cases of low 500mb heights. I can then plot the environmental fields that result from these differences. A hypothesis I created is that if flow pattern is not a factor in the predictor importance, then the composite difference between high and low importance would show no real discernible flow pattern over the domain. Conversely, coherent patterns distinguishing cases with high and low importance of specific predictors would be consistent with a different relationship between predictors and predictands in different large scale flow pattern, further motivating the experiments on FD training of RF.

$$\text{Composite Difference} = mean(hgt500_{high\,importance}) - mean(hgt500_{low\,importance}) \tag{3.1}$$

3.1.3 Efficacy of Using Composite Differences to Identify Meteorological Patterns

To determine the relatively high/low importance, PFI is conducted on the NFD model. The resulting PFI as seen in Figure 3.2 shows that the top five most important variables are, in order, $maxuh03$, $maxuh25$, $u500$, lat, and $mucape$. It is clear that lat that is more important, albeit slightly, than $mucape$ base on their PFI values. However, to focus on the meteorogical differences, the variables that will be used to evaluate composite differences are $maxuh03$, $maxuh25$, $u500$, and $mucape$. Additionally, PFIs for each case were then recalculated to get Z-scores compared to the average PFI, to further emphasize cases which had high or low importance while also normalizing the amount of importance for each variable on a given case. The composite field calculated for these predictors are the change in 500 mb heights, SBCAPE, $u500$ winds, and $v500$ winds. The reasoning behind these fields is that 500 mb heights will represent the large-scale, or synoptic, flow pattern when combined with $u500/v500$ winds. Additionally, some mesoscale information is needed to see if the mesoscale features are similar in these setups or vastly different, so *sbcape* will represent that information. While useful in forecasting convective storms, no storm-scale variables ($maxuh03$, $maxuh25$, or $hrprecip$) were chosen to calculate composite differences on as those variables can vary greatly from one case to another over small spatial scales, making any signal unclear to represent a discernible pattern. For each case, HWT selected regions of interest where the participants believe the highest chance for severe weather would be located and those regions ended up being the regions used in this study. The composites are re-centered on the HWT domains to isolate single flow patterns driving the regional severe weather risk aside from what was driving the weather risk elsewhere in the country on that case and avoid averaging out the important regional flow pattern due to movement of which region is important from one case to the next.

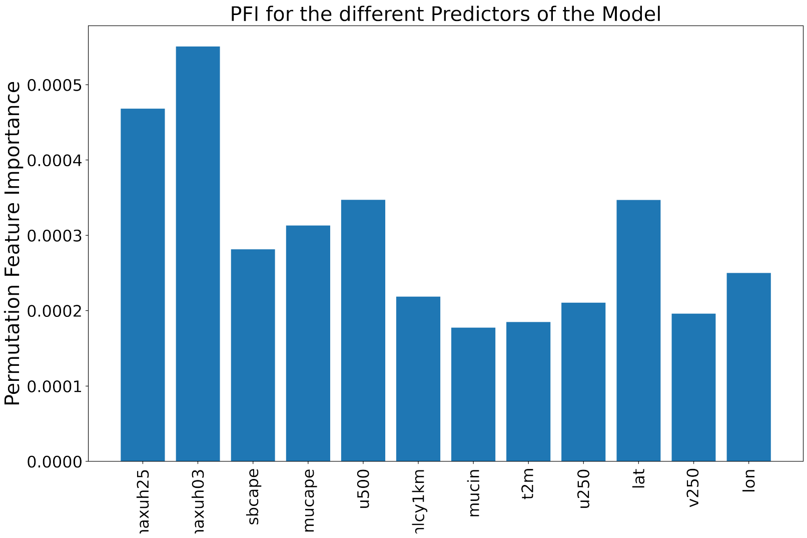
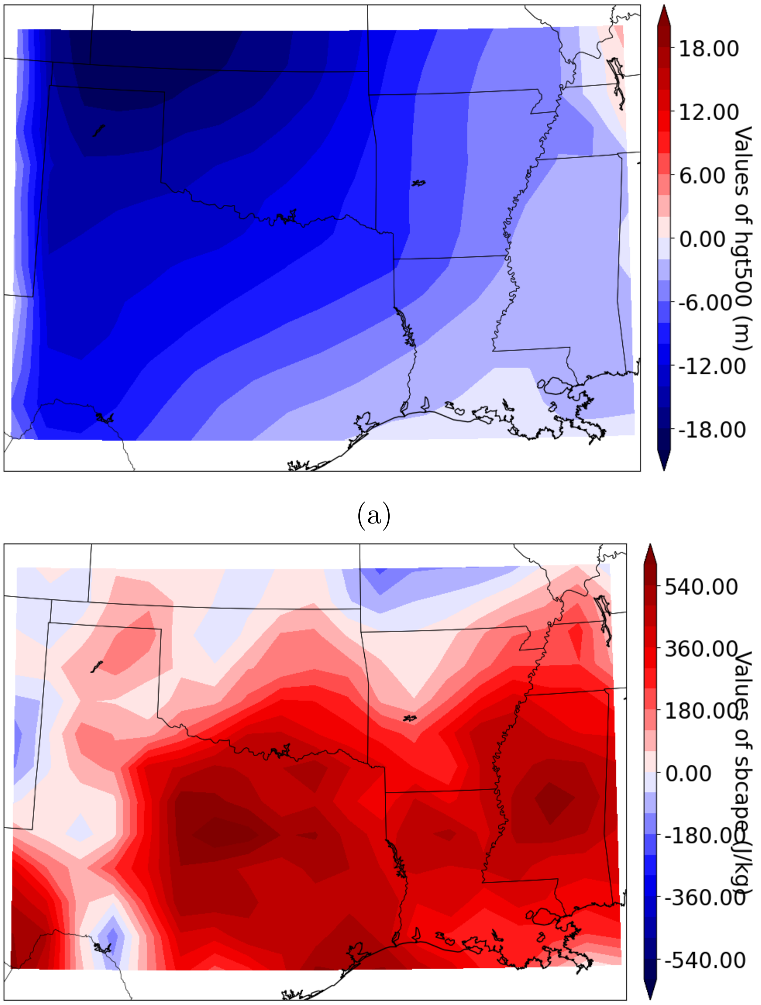


Figure 3.2: PFI values for all predictors used within the NFD Model. It's clear that $maxuh03$, $maxuh25$, $u500$, lat, and $mucape$ are the most important predictors based on PFI.

3.1.4 Patterns Determined from Composite Differences

3.1.4.1 $maxuh03$

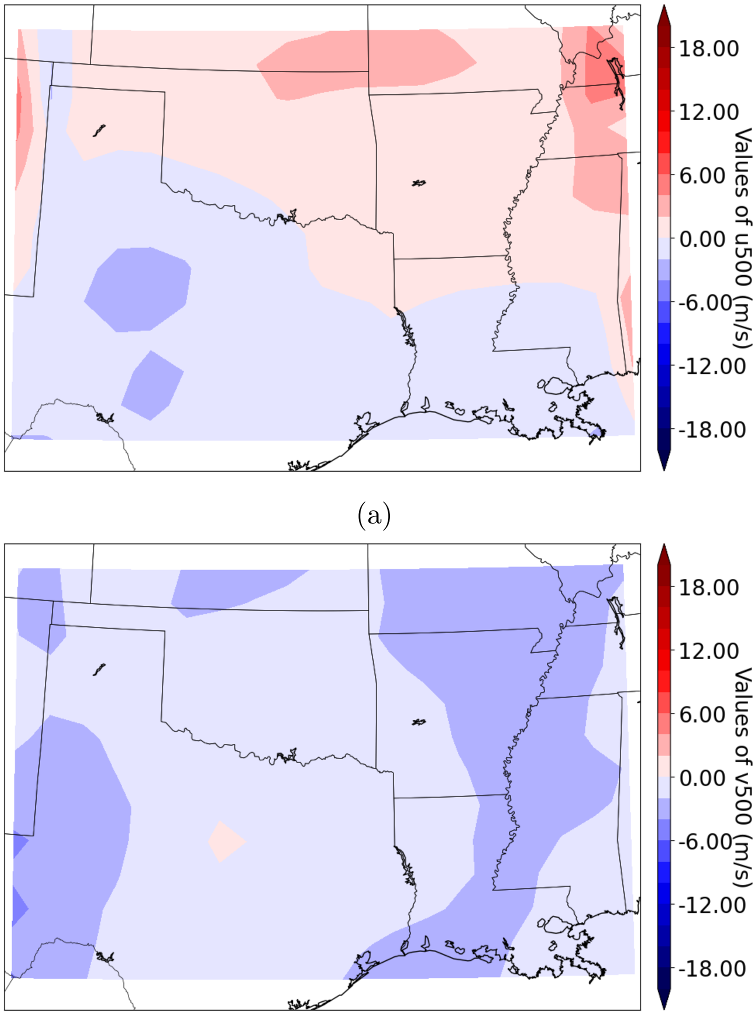
The most important variable from PFI, $maxuh03$, shows a strong signal of more troughing and less ridging (Figure 3.3a) for cases with high importance, particularly over the northwestern quadrant of the domain. Additionally, higher instability on average was noted for high importance cases particularly over the southern portion of the domain in Figure 3.3b. While minimal differences exists for $u500/v500$ winds from high to low importance cases (as seen in Figure 3.4a and Figure 3.4b), a synoptic pattern can still be inferred from the Z and CAPE fields. A possible explanation is that a trough, likely of low amplitude given weak $v500$ winds for high importance cases, moves into the northwestern portion of the forecast domain throughout the period. A strong, unstable airmass setups over the entire HWT domain giving strong amounts of potential energy for thunderstorms. This would be considered a strongly-forced setup when the $maxuh03$ predictor does best in the NFD model.



(a)

(b)

Figure 3.3: (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $maxuh03$.



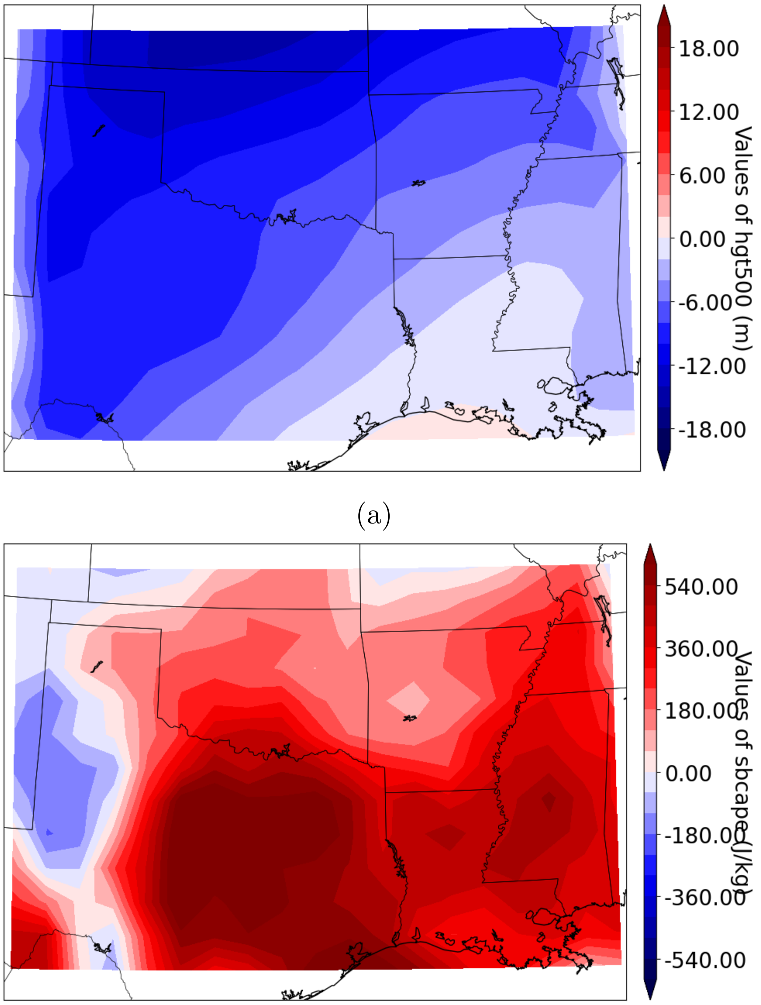
(a)

(b)

Figure 3.4: (a) $u500$ wind composite and (b) $v500$ wind composite for the predictor of $maxuh03$.

3.1.4.2 *maxuh25*

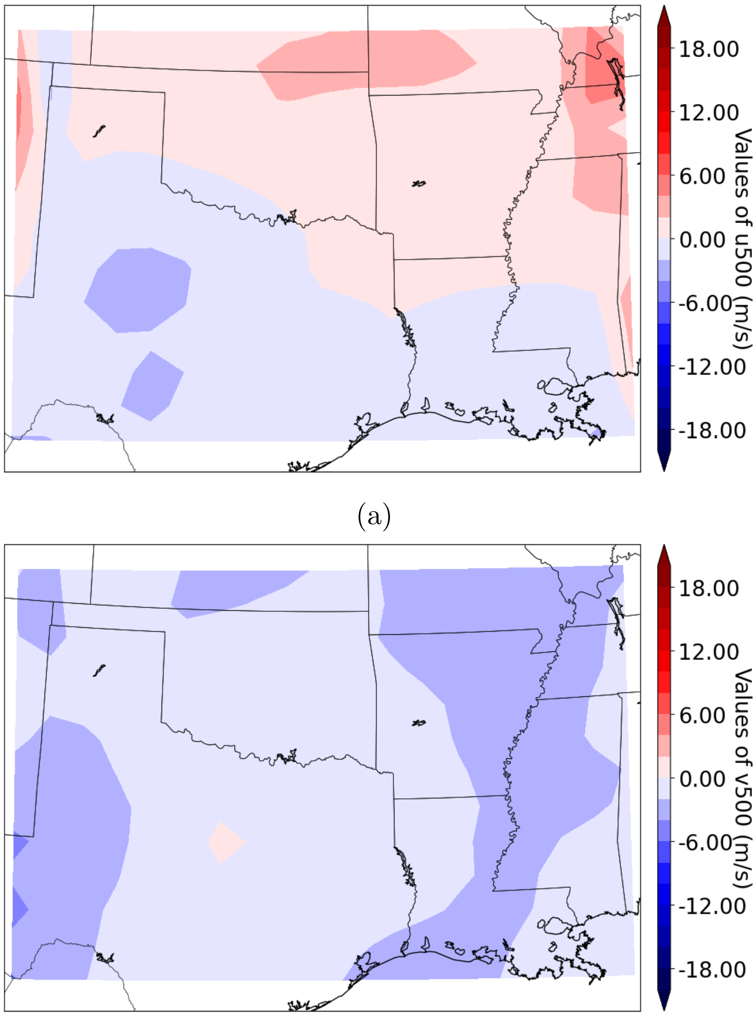
For the second predictor, $maxuh25$, we can notice a lot of similarities between the composite differences of $maxuh03$ and $maxuh25$. Similarly, less ridging, or more troughing, can be observed particularly focused in the northwestern quadrant of the domain in Figure 3.5a. While to a lesser extent, it is still a similar 500 mb height pattern, suggesting that a trough is moving into the western/northwestern portion of the domain. A similar instability pattern exists in $maxuh25$ composite difference for SBCAPE as seen in Figure 3.5b to that of $maxuh03$. The biggest difference between $maxuh03$ composite difference and $maxuh25$ composite difference is the $u500$ winds (Figure 3.6a). The combination of more negative v500 winds (Figure 3.6b) with stronger $u500$ winds for stronger cases indicate anticyclonic shear, similar to that ahead of the main trough east of a ridge axis. Similarly to $maxuh03$, this is likely a low-amplitude trough given the weak $v500$ wind component noted on the composites.



(a)

(b)

Figure 3.5: (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $maxuh25$.



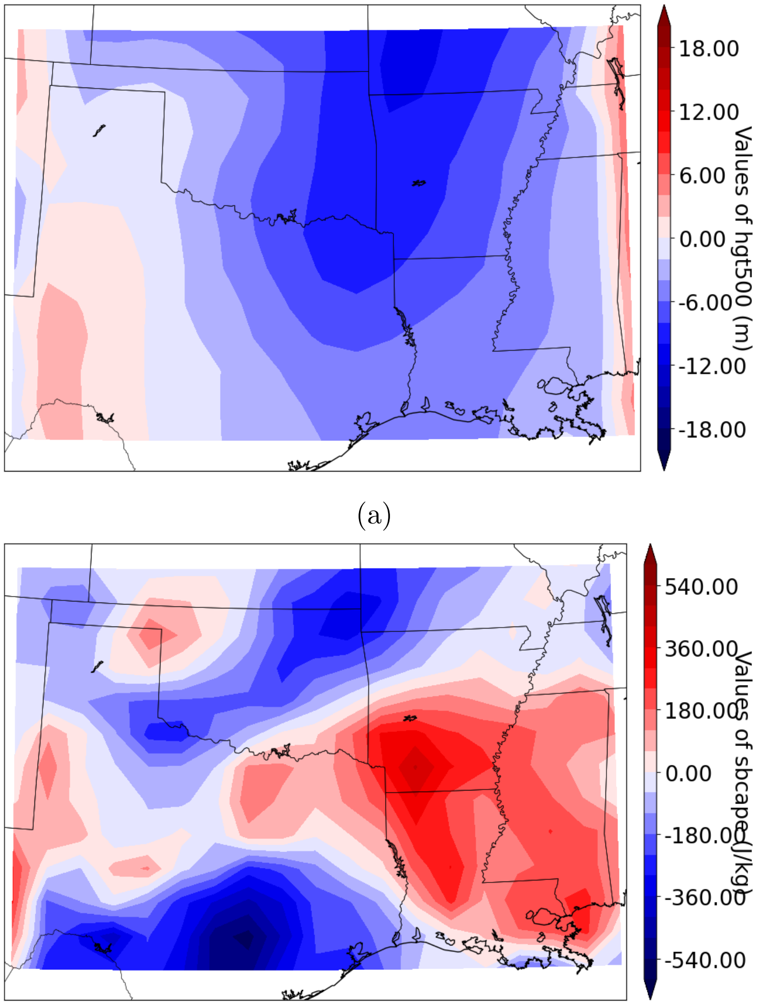
(a)

(b)

Figure 3.6: (a) $u500$ wind composite and (b) $v500$ wind composite for the predictor of $maxuh25$.

3.1.4.3 $u500$

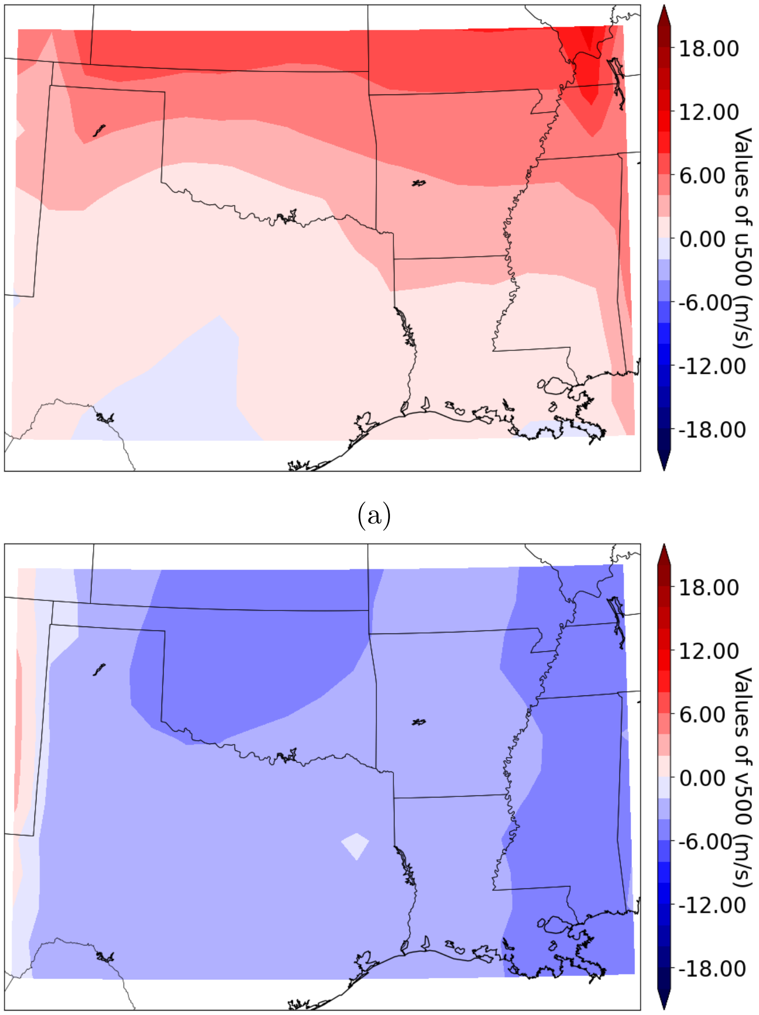
The $u500$ wind composites show a drastically different pattern compared to $maxuh25$ and $maxuh03$. High importance cases do show less trough and more ridging at a similar intensity to that of the other predictors, but this signal is focused in the eastern portion of the domain rather than the western portion of the domain as seen in Figure 3.7a. Additionally, the SBCAPE composite difference reflects this pattern of less ridging (more troughing) over the eastern portion of the domain. This higher instability mainly being focused in the eastern portion as seen in Figure 3.7b of the domain compared to other predictors which focus on the majority of the southern portion of the domain. This indicates that a trough starting in the domain during the forecast period. This trough is likely low amplitude, given the strong $u500$ signal in the high importance cases (as seen in Figure 3.8a) but weak $v500$ signal (as seen in Figure 3.8b) with the warm sector setting up in the eastern portions of the domain.



(a)

(b)

Figure 3.7: (a) 500 mb height composite and (b) SBCAPE composite for the predictor of $u500$.



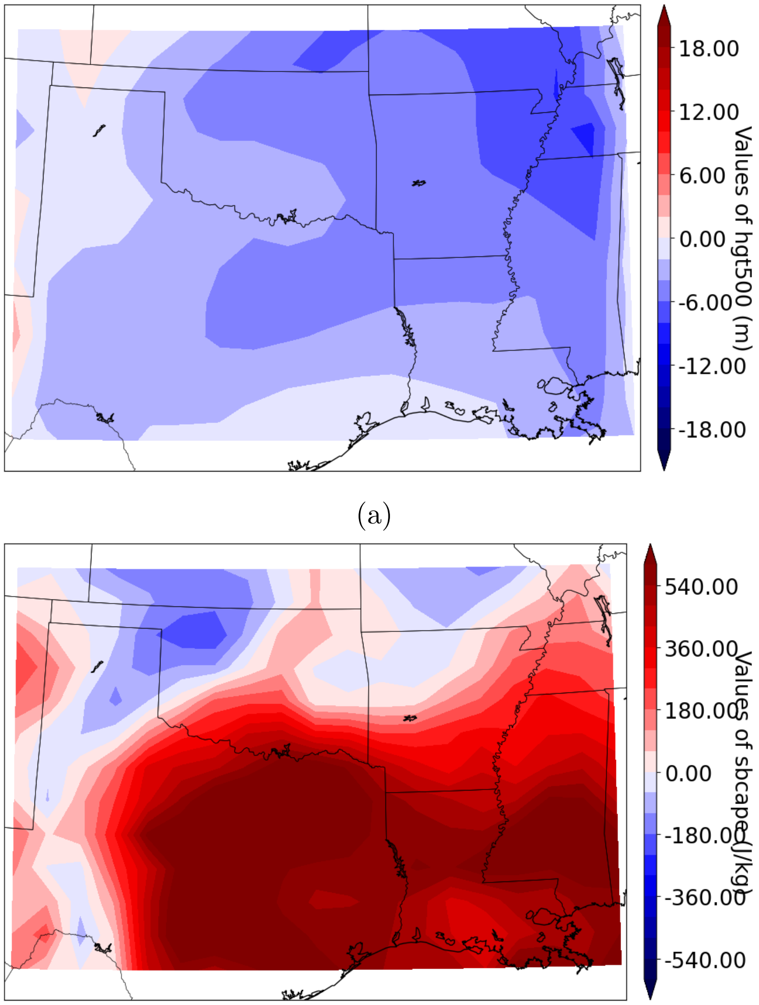
(a)

(b)

Figure 3.8: (a) *u*500 wind composite and (b) *v*500 wind composite for the predictor of *u*500.

3.1.4.4 *mucape*

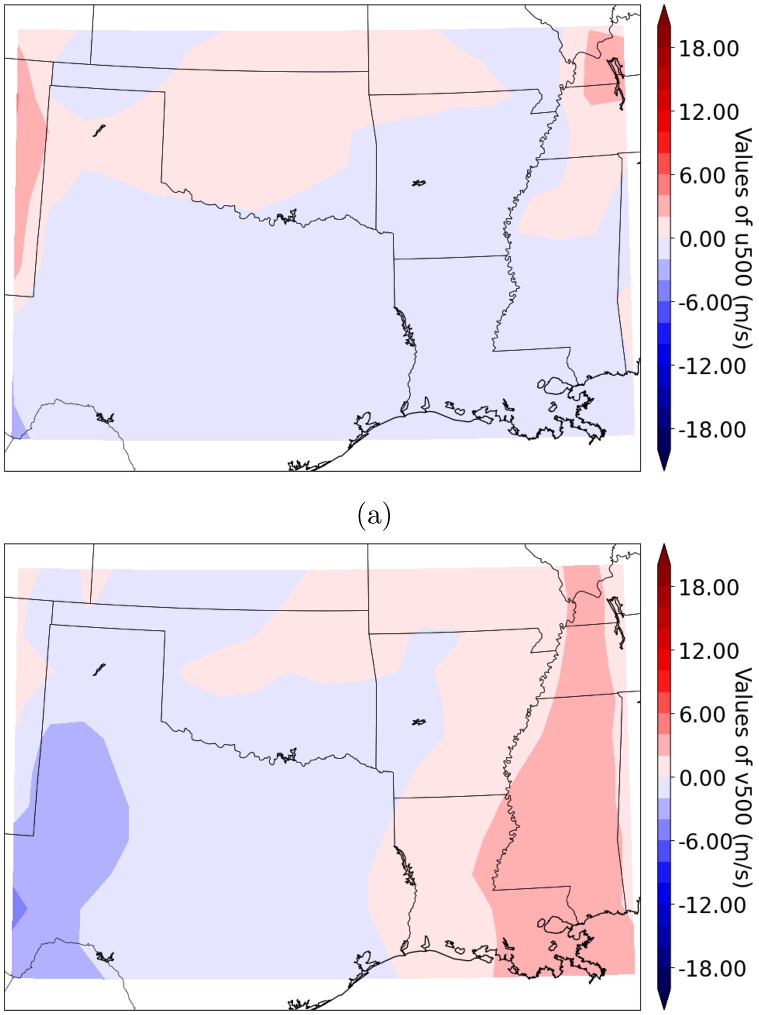
The final predictor composite difference of this study, *mucape*, has the most unique spatial patterns of all features that have been examined. *mucape* has a 500 mb height pattern that shows less ridging/more troughing over the domain, has a very weak signal focused in the eastern portion of the domain as seen in Figure 3.9a. A strong instability pattern is present throughout the domain, except for a small section in the northern part of the domain as seen in Figure 3.9b. Furthermore, there are some hints of a higher amplitude feature on the eastern edge of the domain with strong $u500$ and $v500$ winds as seen in both Figure 3.10a and Figure 3.10b, although just barely in the domain. This pattern is a very uncommon placement for severe weather. One reason for this is that the HWT Domain is relatively small for these larger scale features, so it is hard to determine if the entire trough is inside the domain given the small spatial extent. While the composite difference for *mucape* pattern seems atypical, a discernable pattern is still able to be visualized.



(a)

(b)

Figure 3.9: (a) 500 mb height composite and (b) SBCAPE composite for the predictor of *mucape*.



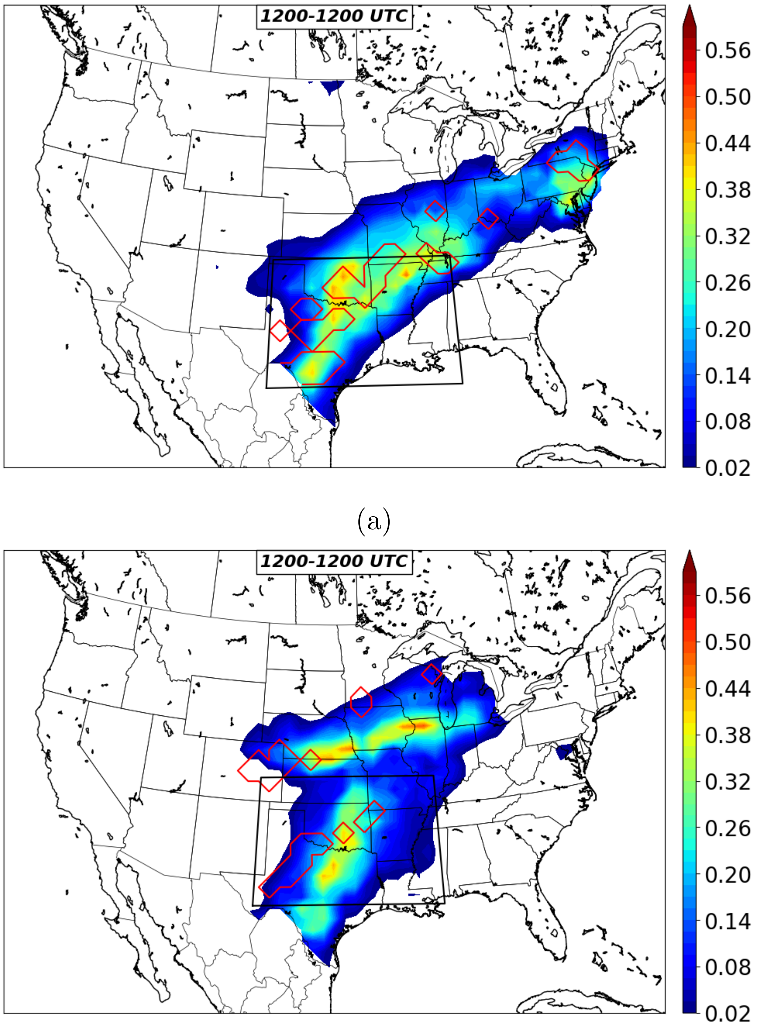
(a)

(b)

Figure 3.10: (a) *u*500 wind composite and (b) *v*500 wind composite for the predictor of *mucape*.

3.1.4.5 Case Study of High and Low Importance

While we can suggest some of these flow patterns based on the composite differences, do those patterns actually occur within high/low importance cases? For this, we will look at the example of the high/low importance case for *maxuh*25 for a high importance (28 April 2021) case and a low importance case (27 April 2021). First, we can examine the model's output probability and see that both cases produce relatively high probabilities (40-50%) as seen in both Figure 3.11a and Figure 3.11b. Similar to the composite differences, this case study will examine the 500 mb heights, SBCAPE, $u500$, and $v500$ winds to see if the cases match up with their composite differences.

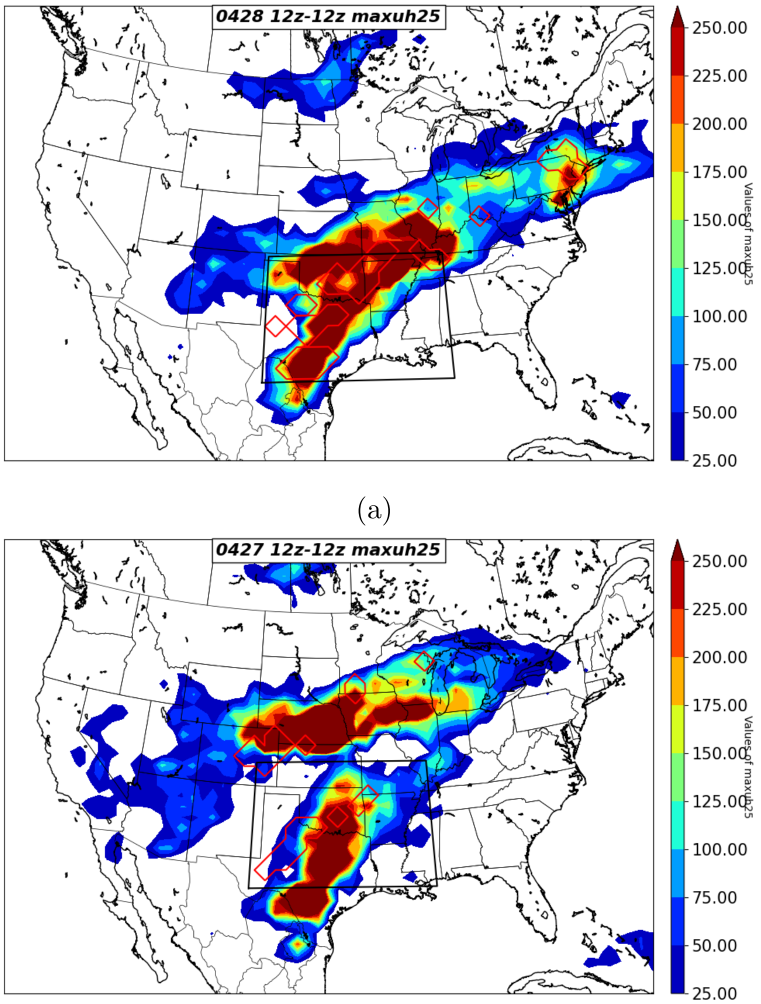


(a)

(b)

Figure 3.11: The (a) high importance and (b) low importance cases output probability from the NFD model.

First, examining the actual $maxuh25$ values, big differences between the high importance and low importance case become apparent. The high importance case $maxuh25$ as seen in Figure 3.12a actually co-aligns with the majority of the storm reports (with values of 125 $m2/s2$ or higher for the majority of storm reports). The low importance case, however, appears to miss a good number of storm reports with only minimal values of $maxuh25$ co-aligned with the majority of severe reports as seen in Figure 3.12b. This demonstrates that $maxuh25$ helps improve the model forecast for the high importance case while worsens the model for the low importance case.

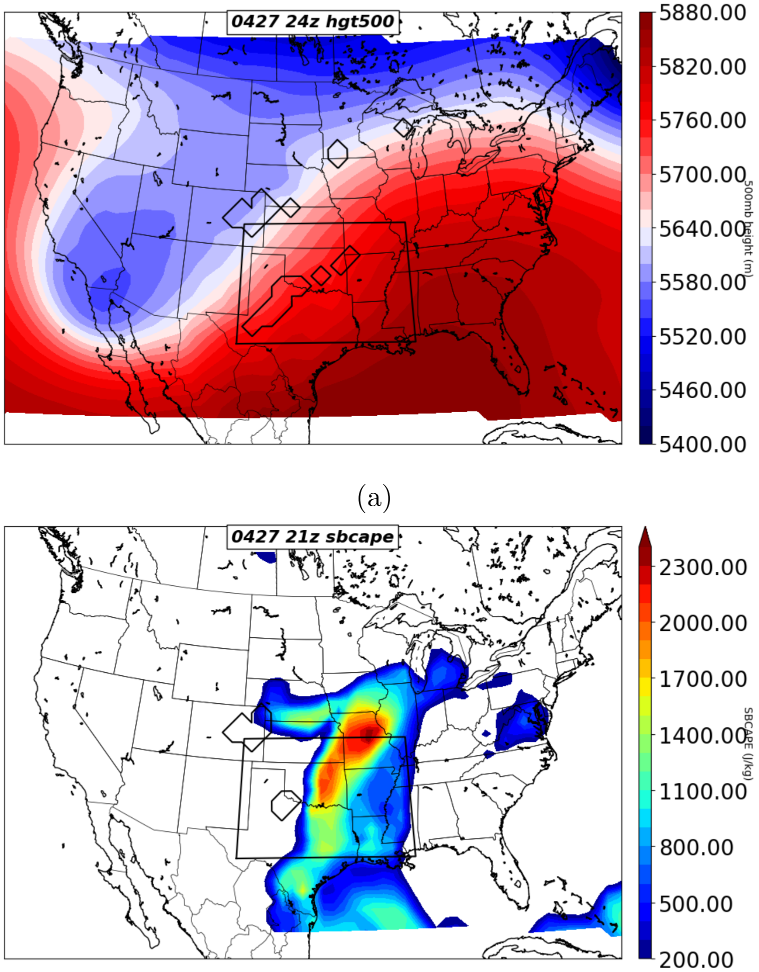


(a)

(b)

Figure 3.12: The (a) high importance *maxuh*25 maximum value from the 10 member ensemble and (b) low importance maximum value from the ten-member ensemble cases output probability from the NFD model.

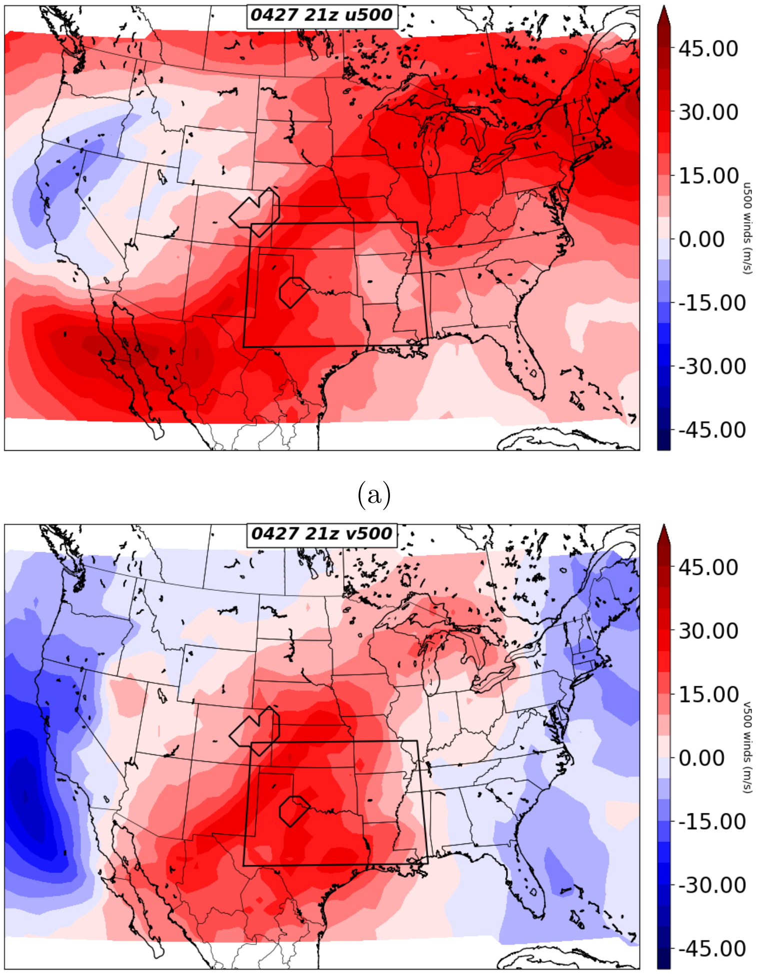
Looking at the environment of the two cases, we notice that the low importance feature has a large trough over the western CONUS. The high importance case has more of a subtle, low-amplitude feature over the central/northern portion of the domain as seen in Figure 3.15a. This reflects what is expected with the 500 mb height pattern that was observed in the composite differences. The instability patterns also reflects the composite differences previously plotted, with the high importance case having predominantly higher instability over the southern portion of the domain seen in Figure 3.15b except for the far western portion of the domain. The low importance case having higher instability more in the northern portion of the domain as seen in Figure 3.13b. The $u500$ and $v500$ winds also reflect the patterns seen in the composite differences, as u500 is maximized over the northern portion of the domain for high importance case (Figure 3.16a) and is maximized in the southern/western portion of the domain for low importance case as seen in Figure 3.14a. The $v500$ winds follows a similar pattern with the high importance case having very low values of $v500$ which demonstrates a low-amplitude feature as seen in Figure 3.16b while the low importance case has strong $v500$ winds which suggests a long-amplitude feature as seen in Figure 3.14b.



(a)

(b)

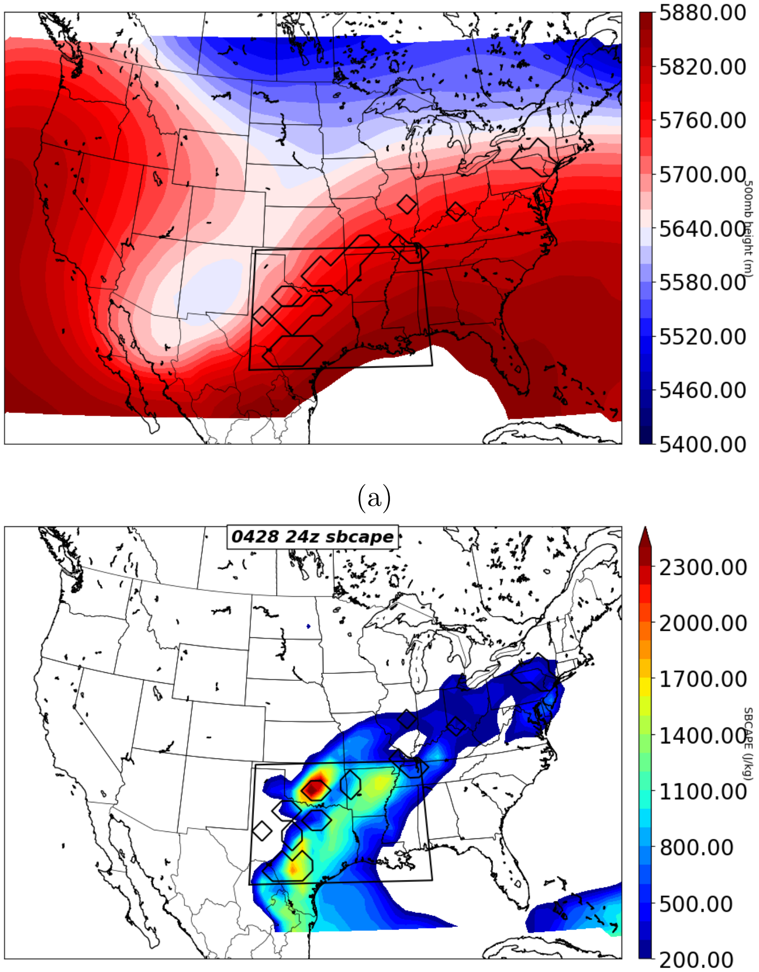
Figure 3.13: The low importance case (a) 500 mb heights, (b) SBCAPE environment.



(a)

(b)

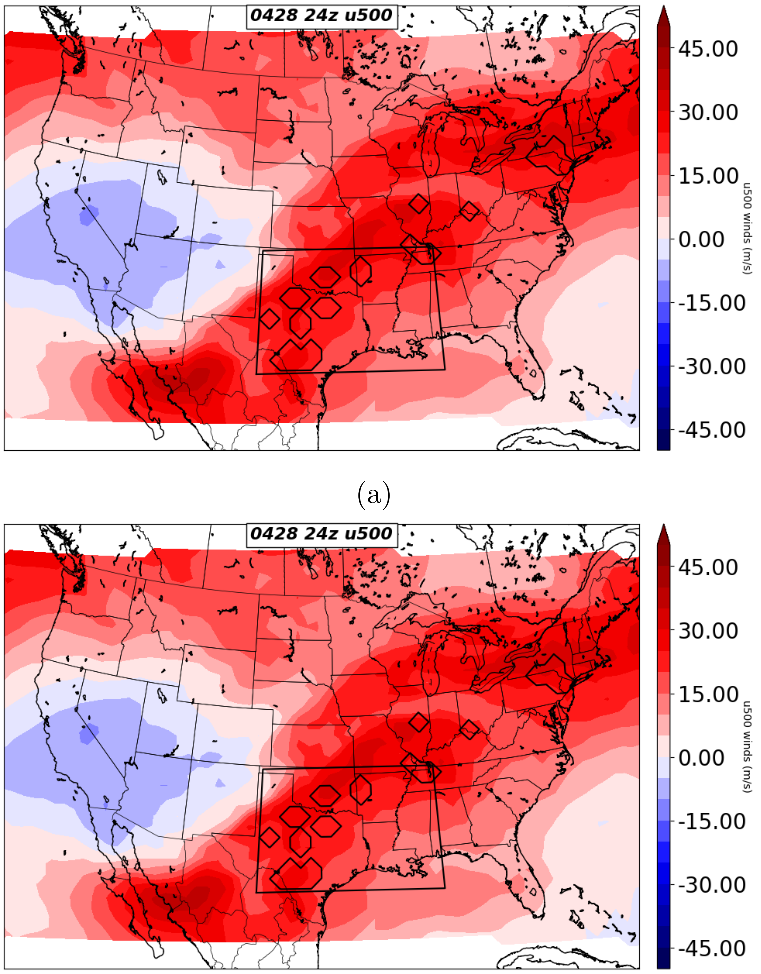
Figure 3.14: The low importance case (a) u500, (b) v500 environment.



(a)

(b)

Figure 3.15: The high importance case (a) 500 mb heights, (b) SBCAPE environment.



(a)

(b)

Figure 3.16: The high importance case (a) u500 and (b) v500 environment.

3.1.4.6 Summary of Composite Differences

The composite difference results have demonstrated that the relationship among predictors change as physically useful and meaningful signals were noticed on the composite difference. The general pattern of the high importance cases was to have more height falls or lower heights compared to the low importance case, were the cases for the model forecasts did the best, as that would suggest stronger forcing for storms to initiate. The u500 and v500 winds reflect this, however, demonstrating a low-amplitude feature about the size of HWT domain or smaller. Additionally, stronger instability over the southern portion of the domain supports stronger storms, which would have a higher tendency to develop updraft helicity (or rotation) in the mid-levels in combination with stronger u500 winds (particularly over the northern portion of the domain), meaning more organized convection moving off of a forcing boundary. This demonstrates the apparent qualitative relationship between the 80-km grid predictors and the synoptic-scale pattern, further motivating a quantitative approach to FD model training.

3.2 Impacts of Flow Dependent Training

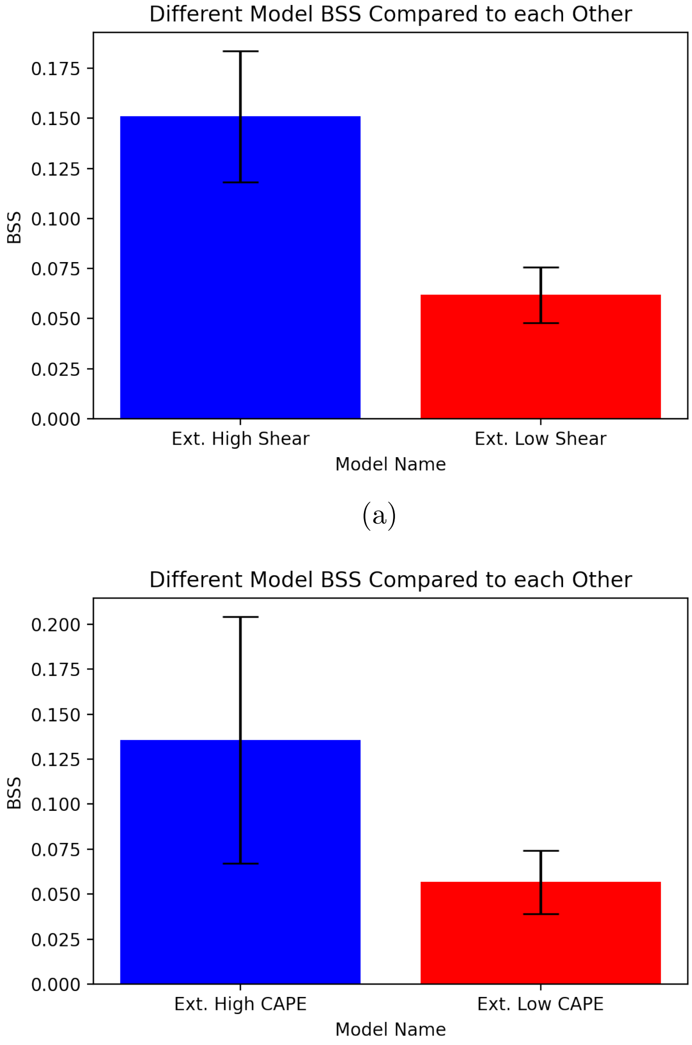
In this section, I will explore the impacts on model performance by training on the flow-dependent classifications. Additionally, I will examine a case study for the PS model to see if the skill makes physical sense. Additionally, I will explore the second flow dependent classification, the spatial pattern classification performance. Finally, I will compare these methods to each other to see if one method performs better compared to the other.

3.2.1 Domain Average Classification

3.2.1.1 BSS Comparison to NFD Model

The BSS of the FD model forecast is evaluated using the baseline model BS evaluated on the same cases as the FD model as the reference forecast in Equation 2.2. Looking at the first FD model Classification of Domain Shear (Figure 3.17a) and CAPE (Figure 3.17b), we notice that the FD model outperforms the baseline model by 5-15%. Some important results to note are that the high versions of both shear and CAPE have the highest skill, but also the largest spread, likely due to the majority of cases happening in higher (or more conducive for severe) shear and CAPE. The low parameter skills are smaller compared to the high parameter model, but their corresponding spread is also much lower. Regardless of category, all models are statistically significantly better than the random model as the BSS was greater than two standard deviations above zero, which suggests that there is some benefit of training the RF model based on flow pattern rather than over multiple flow patterns.

$$BSS = 1 - \frac{BS_{FD}}{BS_{RM}} \tag{3.2}$$



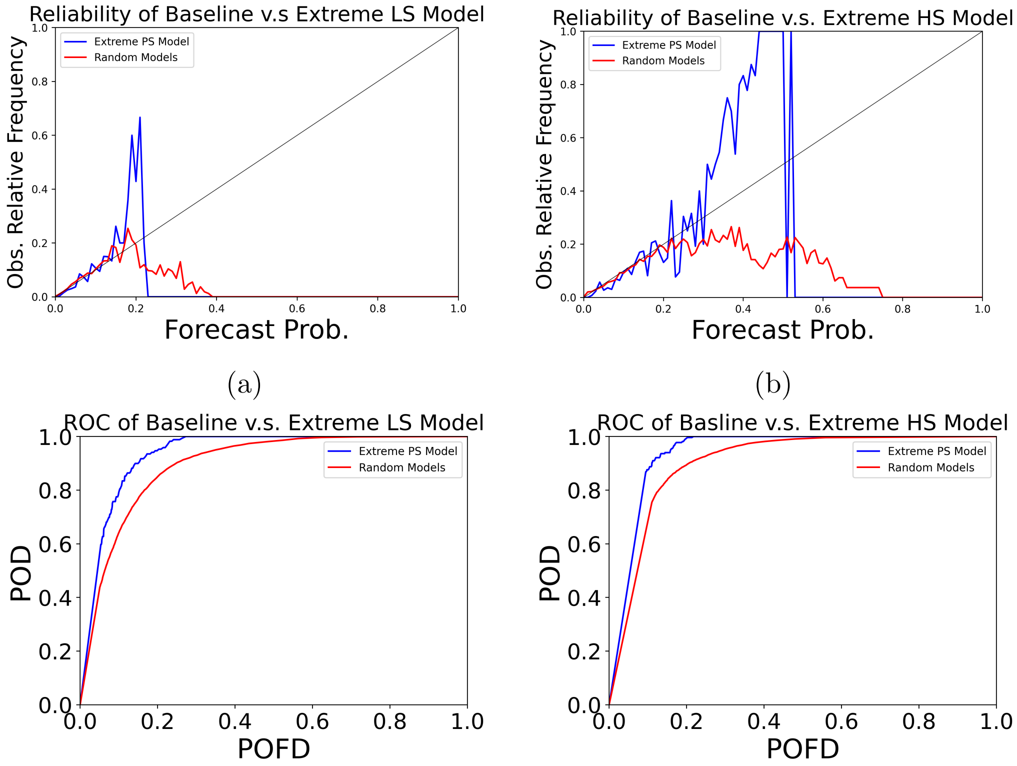
(a)

(b)

Figure 3.17: The BSS for the (a) PS shear and (b) PS CAPE model. A single SD is plotted for clarity.

3.2.1.2 AUC, Reliability Curves, and Model Diagnosis

Beginning with the high/low shear Models, the ROC curves drastically outperform the random model (Figure 3.18c and Figure 3.18d). This is further supported with the AUC values showing 3-4% higher values compared to the random model (Table 3.1), showing the model's ability to detect severe events is much better than the random model. The model's probabilities seem to be reasonably calibrated until probabilities exceed roughly 20% for the low shear model (Figure 3.18a) and probabilities exceed roughly 30% for the high shear model (Figure 3.18b). This is possibly due to limited sample size for those higher probabilities, leading to an underforecating bias at higher probabilities. The random model, however, exhibits an overforecasting bias, with higher probabilities occurring at a low frequency. However, the random model's reliability value for all categories is better than the FD model's mainly due to some calibration issues that could be fixed in the future with an additional calibration step. These results are also consistent for the high/low CAPE FD model. However, two differences are noticeable with the high/low CAPE FD model. The high CAPE AUC is only 1-2% higher than the baseline model from Table 3.1 compared to 3-4% for other FD models. Additionally, the high CAPE reliability is actually slightly lower than the baseline model unlike all other categories. Overall, the PS Model is very good at detecting severe events compared to the random model, though some calibration is needed to improve the probability magnitudes.



(a)

(b)

(c)

(d)

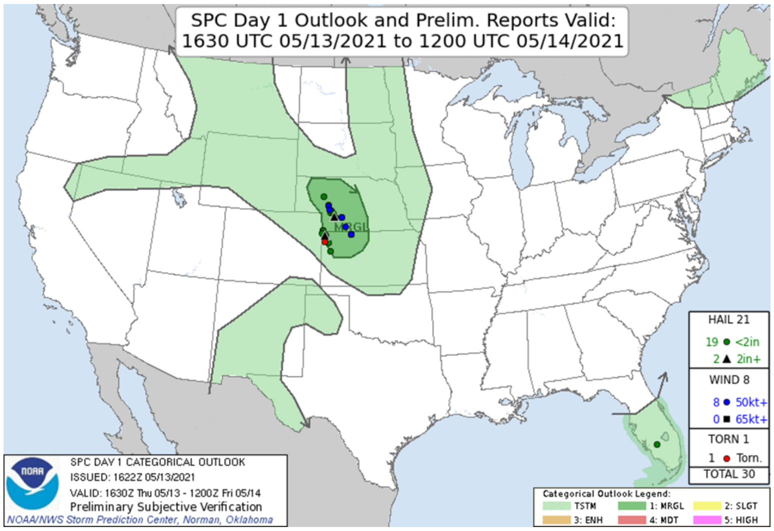
Figure 3.18: On the left, the (a) Low Shear Reliability and (c) ROC curves. On the right, the (b) High Shear Reliability and (d) ROC curves.

Table 3.1: The Reliability, Resolutions, and AUC for each PS Model.

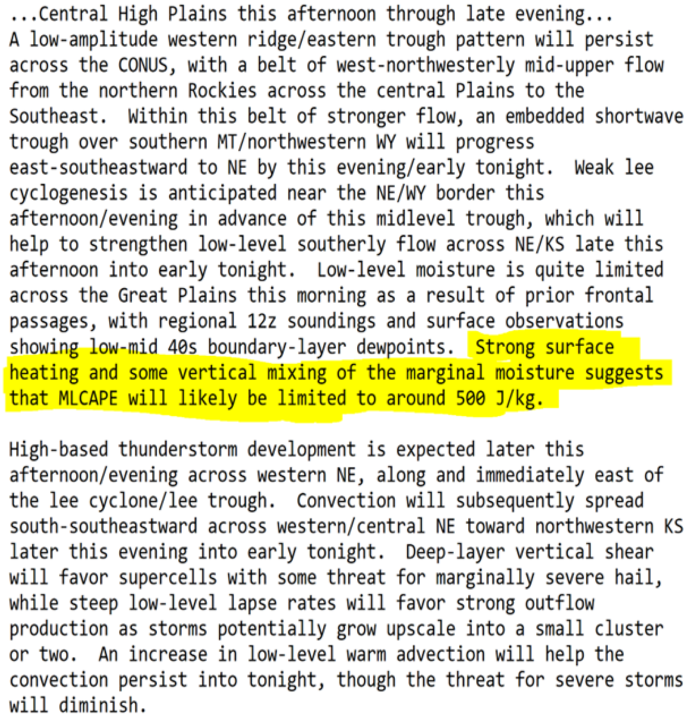
Model Name	Reliability	Resolution	AUC
Low Shear	Model: 0.001350 Baseline: 0.0000824	Model: 0.00459 Baseline: 0.00286	Model: 0.93669 Baseline: 0.89696
High Shear	Model: 0.002984 Baseline: 0.002289	Model: 0.01334 Baseline: 0.00771	Model: 0.94463 Baseline: 0.91275
Low CAPE	Model: 0.001430 Baseline: 0.000716	Model: 0.00423 Baseline: 0.00249	Model: 0.94275 Baseline: 0.90670
High CAPE	Model: 0.002993 Baseline: 0.003271	Model: 0.01895 Baseline: 0.01429	Model: 0.93559 Baseline: 0.92331

3.2.1.3 Subjective Differences between NFD Model and FD Model

While objective differences are always good in determining which model is better, it is important to see how these differences qualitatively appear in individual cases. As a result, a subjective view of the differences between the FD model on 13 May 2021 and the random model can give insights on what differences might be occurring. First, we can confirm that this case is indeed Low CAPE per the SPC outlook (Figure 3.19a) having low probabilities and the SPC discussion (Figure 3.19b) only mentions low instability of 500 J/kg of ML CAPE. The FD probability output demonstrates that the model has skill as the highest probabilities occur within the region of storm reports as seen in Figure 3.20. This when compared to multiple random model outputs, shows a different output than the random models. Particularly, the random models have higher probabilities compared to the FD model and also higher probabilities outside of the regions of storm reports. To investigate the differences between these two different outputs, we can look at Feature Importance (not PFI) of the models to see that predictor importance does change from the RM model, particularly SBCAPE, mucape, u500, and 0-1 km SRH (Figure 3.24). Investigating these predictors with a tree interpreter, one predictor's change from the random model and FD model is drastic is the contributions to the 0-1 km SRH. The FD model contribution(s) (Figure 3.25a and Figure 3.26) is highest from this predictor compared to the random model (Figure 3.25b and it includes a lot of the region where storm reports occurred. A possible reason for this is that in low instability setups, storms can overcome the shortcoming in instability with increased dynamics, in this case 0-1 km SRH. This suggests that the FD model identifies different relationships among predictors that the random model is otherwise unable to determine.


SPC Day 1 Outlook and Prelim. Reports Valid:
1630 UTC 05/13/2021 to 1200 UTC 05/14/2021
MRGL
HAIL 21
19 ● <2in
2 ▲ 2in+
WIND 8
8 ● 50kt+
0 ■ 65kt+
TORN 1
1 ● Torn.
TOTAL 30
SPC DAY 1 CATEGORICAL OUTLOOK
ISSUED: 1622Z 05/13/2021
VALID: 1630Z Thu 05/13 - 1200Z Fri 05/14
Preliminary Subjective Verification
NOAA/NWS Storm Prediction Center, Norman, Oklahoma
Categorical Outlook Legend:
TSTM
1: MRGL
2: SLGT
3: ENH
4: MDT
5: HIGH

(a)



...Central High Plains this afternoon through late evening...
A low-amplitude western ridge/eastern trough pattern will persist across the CONUS, with a belt of west-northwesterly mid-upper flow from the northern Rockies across the central Plains to the Southeast. Within this belt of stronger flow, an embedded shortwave trough over southern MT/northwestern WY will progress east-southeastward to NE by this evening/early tonight. Weak lee cyclogenesis is anticipated near the NE/WY border this afternoon/evening in advance of this midlevel trough, which will help to strengthen low-level southerly flow across NE/KS late this afternoon into early tonight. Low-level moisture is quite limited across the Great Plains this morning as a result of prior frontal passages, with regional 12z soundings and surface observations showing low-mid 40s boundary-layer dewpoints. Strong surface heating and some vertical mixing of the marginal moisture suggests that MLCAPE will likely be limited to around 500 J/kg.

High-based thunderstorm development is expected later this afternoon/evening across western NE, along and immediately east of the lee cyclone/lee trough. Convection will subsequently spread south-southeastward across western/central NE toward northwestern KS later this evening into early tonight. Deep-layer vertical shear will favor supercells with some threat for marginally severe hail, while steep low-level lapse rates will favor strong outflow production as storms potentially grow upscale into a small cluster or two. An increase in low-level warm advection will help the convection persist into tonight, though the threat for severe storms will diminish.

(b)

Figure 3.19: (a) The SPC 1630 UTC Outlook overlayed with storm reports and (b) the 1630 UTC SPC Discussion for the 14 May 2021.

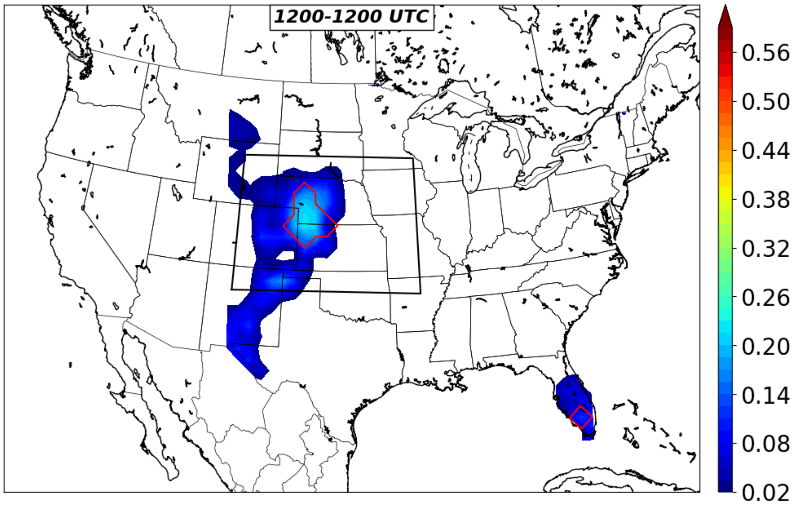
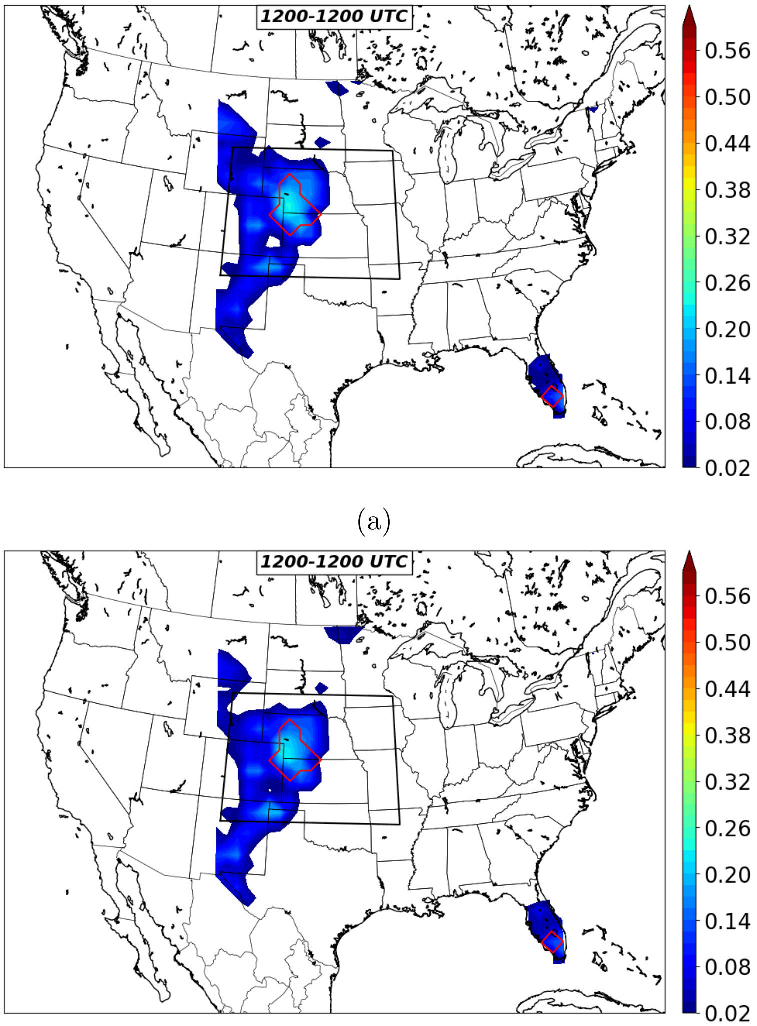


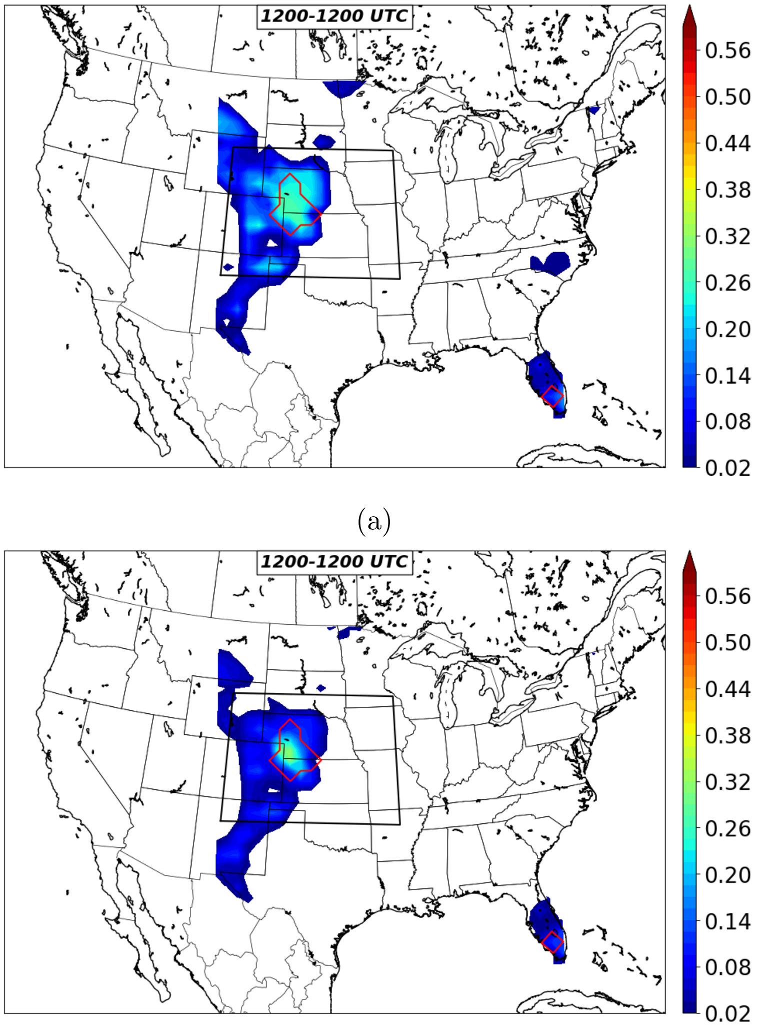
Figure 3.20: FD model output for 13 May 2021. Probability covers the storm report region well, but note a locally higher probability region on the southern portion of the domain.



(a)

(b)

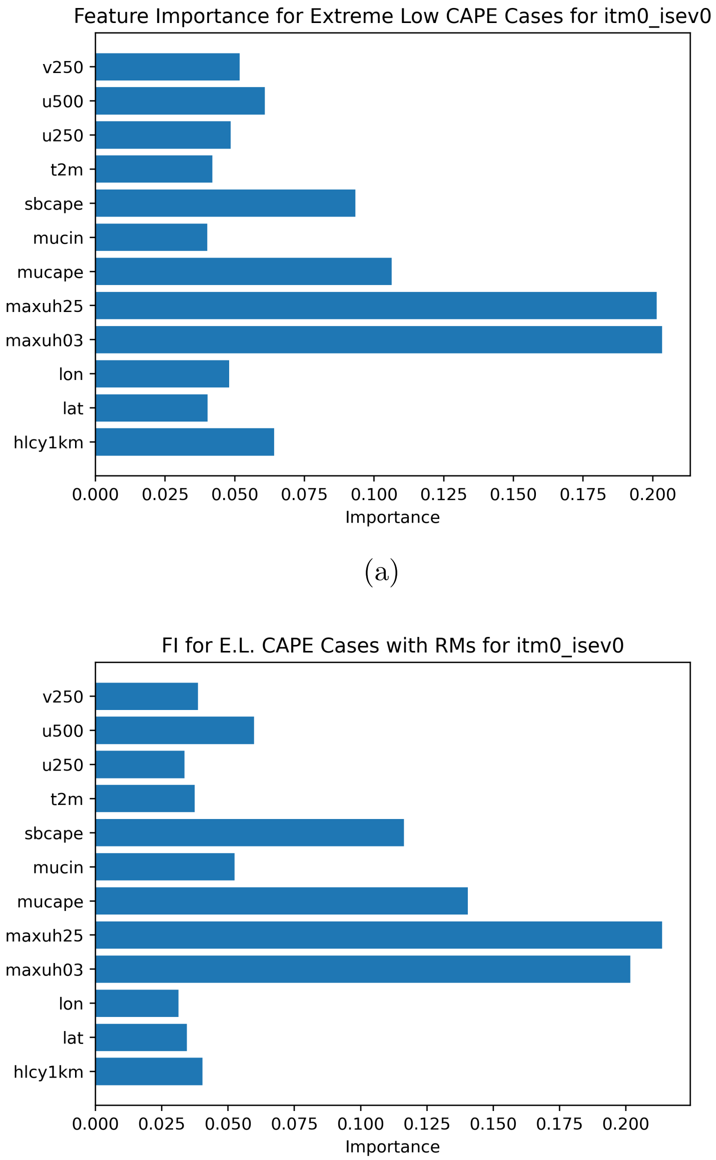
Figure 3.21: Two different iterations of the NFD model baseline and the probability generated for each iteration.



(a)

(b)

Figure 3.22: Two more different iterations of the NFD model baseline and the probability generated for each iteration.



(a)

(b)

Figure 3.23: (a) The FD model FI, (b) the average FI of the NFD model.

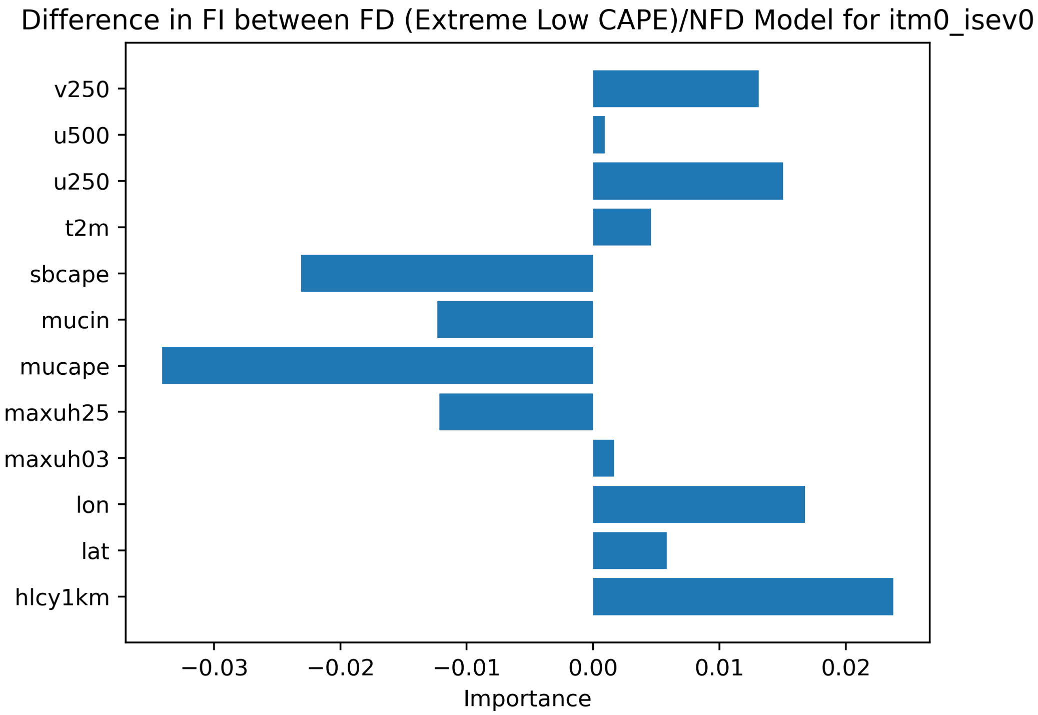
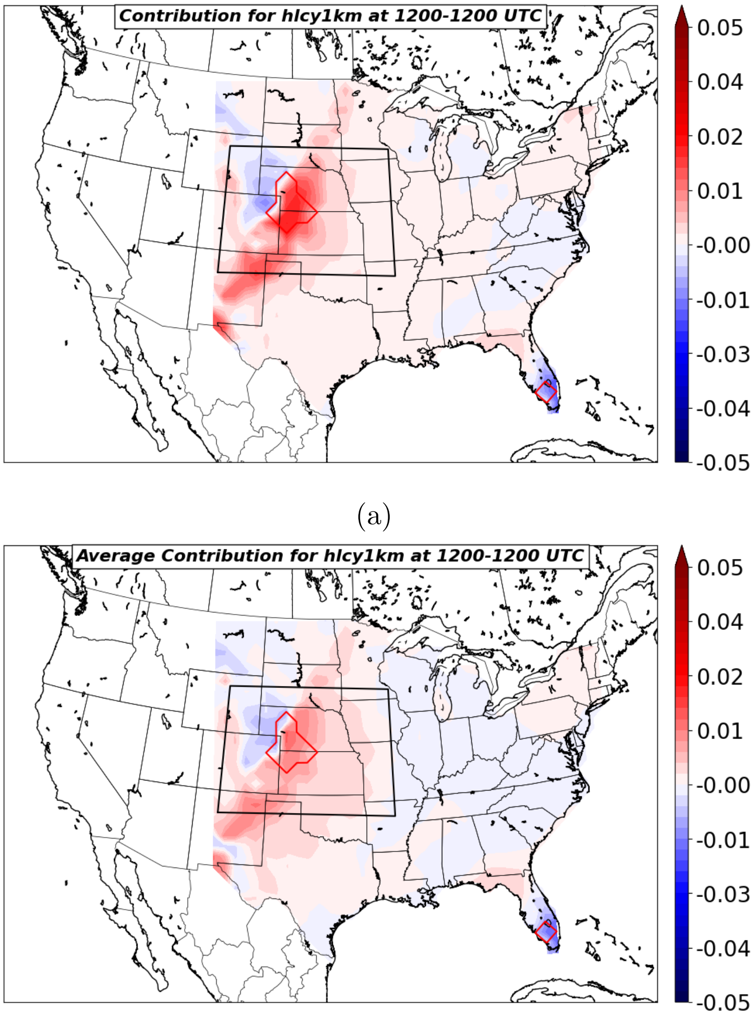


Figure 3.24: The difference between the FD model FI and the average FI of the NFD model. More emphasis was place on low-level shear parameters and less on instability parameters, which reflects what you would expect in a Low CAPE setup.


Contribution for hlcy1km at 1200-1200 UTC
0.05
0.04
0.02
0.01
-0.00
-0.01
-0.03
-0.04
-0.05
Average Contribution for hlcy1km at 1200-1200 UTC
0.05
0.04
0.02
0.01
-0.00
-0.01
-0.03
-0.04
-0.05

(a)

(b)

Figure 3.25: (a) The 0-1km Helicity Contribution to the FD model, (b) the average contribution of 0-1km Helicity to the NFD model.

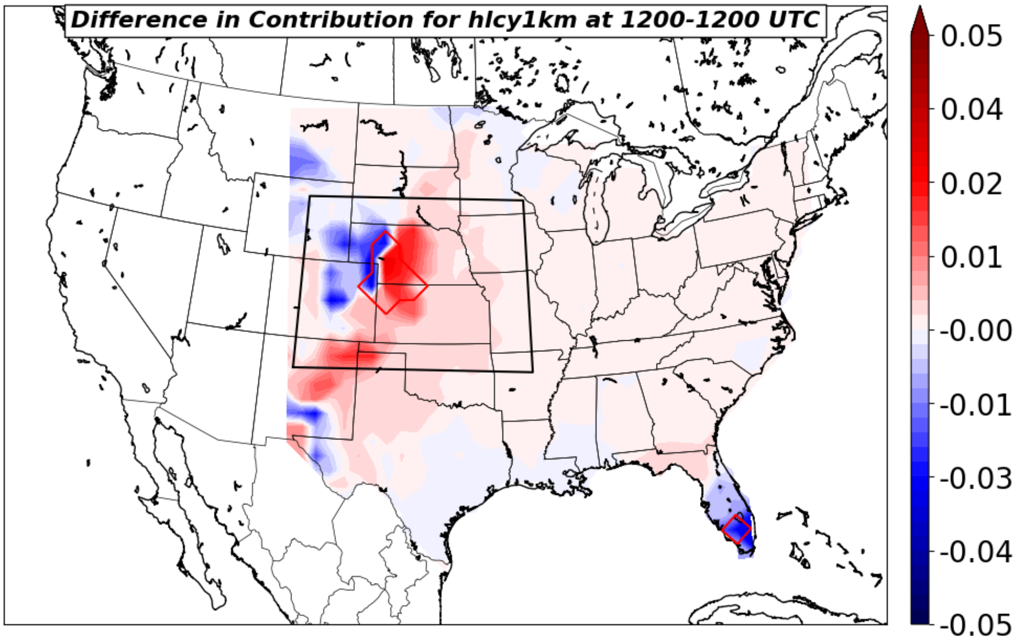
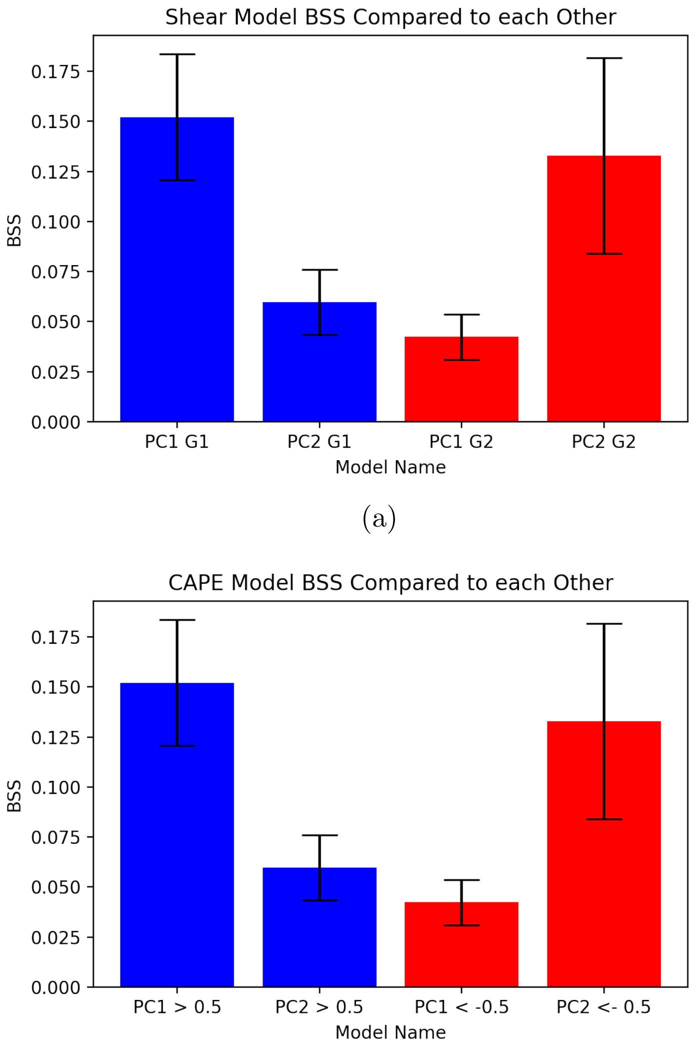


Figure 3.26: The difference between the two. This is the largest difference in contribution between the NFD and FD model, and helps increase probabilities over the storm report region and decrease it to the NW and N of the reports region.

3.2.2 Spatial Pattern Classification

3.2.2.1 BSS Comparison to NFD

Looking at the BSS for all the PC Models, we notice similar trends to that of the PS models. We do notice that particularly from PC1 G1 and PC2 G2 for both CAPE/Shear has the best skill (Figure 3.27a and Figure 3.27b). Their skills do have a lot of spread, but similar to the Domain Average Classification, all models are statistically better than the baseline model. It appears that both methods have improved skill compared against the random model baseline, demonstrating that FD model training provides some skill that training on random cases cannot provide.



(a)

(b)

Figure 3.27: (a) BSS plot for shear PC model and (b) BSS plot for the CAPE PC model.

3.2.2.2 Reliability and Resolution

To avoid unnecessary amounts of plots, below is the reliability, resolution, and AUC values as this will provide similar information as those plots. Similar to the Domain Average Classification, we see the similar trend of better AUC, better resolution, but slightly worse reliability as seen in Table 3.2. This would support the Spatial Pattern classification like the Domain Average Classification can identify severe events at a higher rate than the baseline models. However, like the Domain Average method, some calibration of the probabilities is necessary to address some of the reliability issues.

Table 3.2: The Reliability, Resolutions, and AUC for each PC Model.

Model Name	**Reliability**	**Resolution**	**AUC**
CAPE PC1 G1	Model: 0.002831 Baseline: 0.004163	Model: 0.002309 Baseline: 0.001734	Model: 0.94303 Baseline: 0.93446
CAPE PC1 G2	Model: 0.001050 Baseline: 0.000780	Model: 0.00385 Baseline: 0.00219	Model: 0.94616 Baseline: 0.92129
Shear PC1 G1	Model: 0.003187 Baseline: 0.002022	Model: 0.01318 Baseline: 0.00771	Model: 0.94062 Baseline: 0.90391
Shear PC1 G2	Model: 0.00119 Baseline: 0.000678	Model: 0.00370 Baseline: 0.00252	Model: 0.94528 Baseline: 0.91555
CAPE PC2 G1	Model: 0.002096 Baseline: 0.002015	Model: 0.01311 Baseline: 0.00982	Model: 0.94210 Baseline: 0.92467
CAPE PC2 G2	Model: 0.00138 Baseline: 0.000969	Model: 0.004994 Baseline: 0.00348	Model: 0.95271 Baseline: 0.91915
Shear PC2 G1	Model: 0.00112 Baseline: 0.000872	Model: 0.005567 Baseline: 0.00387	Model: 0.9347 Baseline: 0.89988
Shear PC2 G2	Model: 0.00219 Baseline: 0.00245	Model: 0.01430 Baseline: 0.1084	Model: 0.94702 Baseline: 0.92937

3.2.2.3 Comparison of Both Methods

While both methods have shown statistical significantly better models than the random models, which method would be considered the best method? While it would be easy to try and compare the model diagnosis above to determine which model is better, it is not a direct and equal comparison. Besides the cases not being the exact same in each category, the number of cases within each category are not the same. However, the combined model allows for all models to be forecasting the same number of cases, and thus a direct and equal comparison. Examining the BSS of the combined model, we notice that all models are statistically better than the random model baseline. This means that both methods appear to improve the training of the RF model compared to the random model or more of a traditional method. Additionally, all model's spread encompass every other model, meaning that no real model or classification method appears better than the other. Both methods of Flow Dependent model training improves the forecast compared to the baseline, however, no clear advantage is observed between the PS and PC methods. Future studies, potentially using even larger sample sizes, may find more effective ways of using the spatial pattern to improve the PS Model even further.

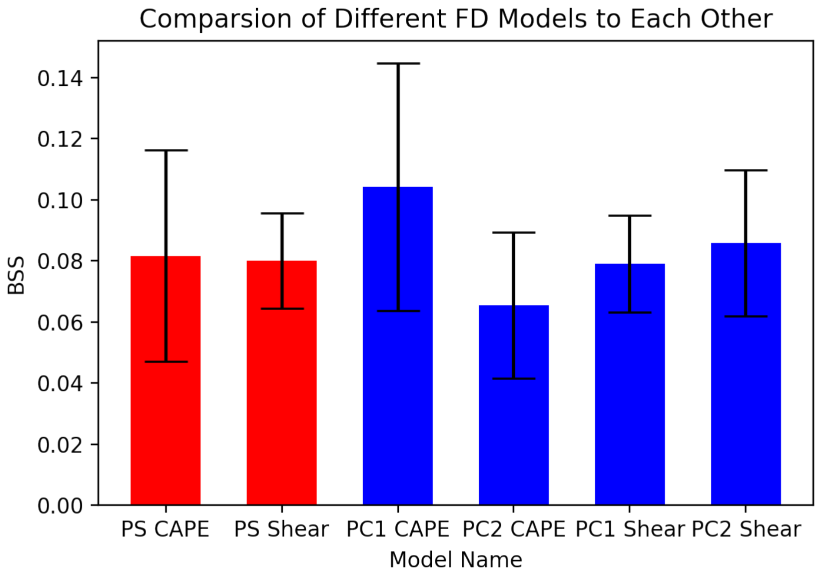


Figure 3.28: A comparison of both the PS and PC Models using the Combined BSS as mentioned earlier. No FD model has a clear advantage over the other, but all FD models are statistically significantly better than the baseline. For each model, one SD is plotted for clarity.

3.3 Impacts of Sample Size on Model Performance

In this section, I will discuss the impacts of sample size on the FD model performance. I will compare both the PS and PC model performance with the increased sample sizes. Any substantial model performance degradation will be explored as it arises. Finally, any conclusions of how a FD model does will be briefly stated with implication of findings discussed further in Chapter 4.

3.3.1 Combined BSS Comparison

While the conclusions reached were statistically significant, the dataset was a relatively small dataset (e.g., only 8 cases in each PS category). Would this strong conclusion extend if more samples were included in the model training or would it begin to fall apart? For this study, the additional dataset of the 2019 HWT cases were added, increasing the case size to 53 cases. Furthermore, most FD model's case sizes became approximately double with the exception of CAPE PC1 as that did not really project well (particularly in the negative) into 2019. Another important thing to note is that the original PCs/EOFs were used, meaning the new cases would be projected onto the leading EOFs to see which group that would fall in to mimic what could be an operational version of the model in the future, and keep definition of flow categories the same while only changing the number of cases in each category.

Looking again at the combined BSS output for the different FD models as seen in Figure 3.29, we notice that there are some differences. For example, all PC Models except for PC1 Shear are no longer statistically better than the random model, but both PS models are statistically better. In fact, the PS model skill remains roughly the same for PS Shear or slightly improved in the case of PS CAPE. This only occurs with a slight adjustment to the original PS cutoffs in order to double the case sizes of each model (from 8 cases to 16 cases).

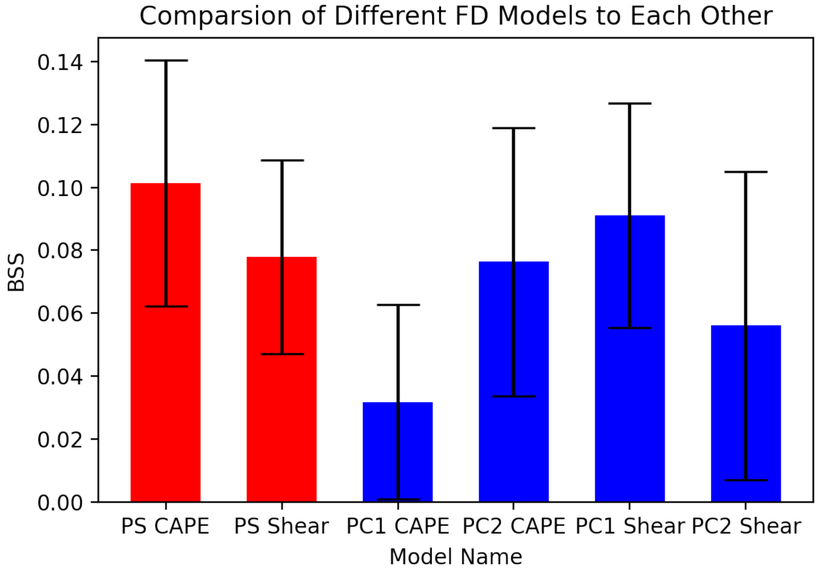


Figure 3.29: A comparison of both the PS and PC Models using the Combined BSS as mentioned earlier on the larger sample size. PS Model remain statistically significantly better than the baseline, but all PC Models (except for PC1 Shear) do not remain statistically better than the baseline.

3.3.2 Reasons for Differences

While it is important to see how the models react to a larger number of cases, it is also important to speculate on why the number of cases (particularly for the PC Models) drastically decreases model skill. One reason for this is that 2019 does not project onto the original 2021 EOFs well (particularly for PC 1 CAPE). This means that the case may not truly represent the pattern that it is projecting onto well. Another reason could be that the Spatial Pattern Method is not a great way of classifying Flow Pattern. However, this seems to not be the case, as there were significant results with the smaller sample size. As a result, it is possible that the poor project of the spatial patterns into the 2019 is to blame for the poor performance with increased sample size. However, the PS FD model still has a statistically better model than against the baseline even with the increase in sample size. This suggests that FD model training can still be better than the baseline model even with more samples to add, suggesting that the PS method benefits can persist even into other seasons or with larger sample sizes.

Chapter 4

Conclusion

4.1 Summary of Work and Findings

The purpose of this study was to explore the possibility of FD model training of RF to improve probabilistic severe weather forecasts. First step, I explored if RF predictors performed better or worse (high/low importance) based on the flow pattern. Considering the results from the RF predictor experiments, this study introduced, implemented and evaluated two different methods of classifying the flow pattern based on the domain-average parameter space and spatial pattern as quantified using principal components. The goal of these experiments was to see if models trained on certain flow patterns could perform better than a baseline random model, and determine how much these different classification methods affect the relative skill of FD vs baseline. Finally, this study explored the impacts of increasing sample size on the FD model performance.

The different flow patterns were evident for each predictor in the composite differences between cases where a given predictor has relatively high versus relatively low importance. Three unique flow patterns arose from the four different predictors from their composite differences. Predictors $maxuh03$ and $maxu25$ had similar flow patterns, with high PFI cases showing more strongly forced setups. However, some limitation exists as the domain cannot include the amplitude and depth of the trough with a unstable warm sector for high importance cases. The $u500$ composite difference favored a more compacted trough within northwest flow, but shifted east and with a much smaller warm sector compared to the previous two predictors. The final predictor, *mucape*, shows the most atypical pattern, with the trough moving out of the domain, but a stronger warm sector present throughout the entirety of the domain for high importance cases. A case study based on the $maxuh25$ importance demonstrated that high and low importance cases not only make physical sense, but reflect elements of their composite differences, further justifying the idea that FD model training may provide skill that might not otherwise be obtained using traditional model training methods.

Using the Domain Average Classification method, examinations of PS shear show increased skill in forecasting the PS case(s) flow pattern compared to the baseline random model. Examining the different diagnostic metrics, PS shear model excelled in detection and resolution of severe events, but had worse reliability compared to the baseline model. The case study demonstrates that instability predictors (*sbcape*) contributed less positively. The 0-1km helicity predictor contributed more negatively to the probability forecast outside of the region of storm reports for the FD model compared to the baseline model. Since the instability predictors (*sbcape*, *mucape*) relates to the instability in the environment, placing less importance on the instability predictors helped decrease spurious high probability in this case for the FD model. The FI on the FD model shows that in general, decreased emphasis on instability and increased emphasis of low-level shear. The differences represented in this case are why the FD model outperforms the baseline model.

Exploring the PC models, I noticed a similar trend for individual models that was observed to that of the PS models. High resolution and detection of severe events seem to override the poorer reliability compared to the baseline model. However, the observed high reliability was likely due to low samples similar to the PS model. A comparison between models showed that all models provide statistically significant positive skills compared to the baseline model and that all models have similar skills to each other.

To confirm that these results were not sensitive to sample size, I increased the sample sizes for all models (including the baseline) by adding data from the 2019 severe season. As a result, the comparison of model skill showed that the PS model performed statistically better than the random baseline model. However, the PC models (with the exception of PC-1 of shear) did not remain statistically significant. One reason for this change is that the EOFs projected poorly onto the 2019 cases, particularly for the EOFs of CAPE, meaning that the leading modes of variability in 2021 were not necessarily representative of those in 2019. This led to an uneven split between positive and negative projections of the 2019 cases. This issue could be resolved by increasing the base sample size for the PCs or EOFs to account for more of the cases more accurately. However, the significance of the PS model remained consistent (or slightly higher skill) while increasing the sample sizes showed the promise of this method of training RF models.

4.2 Future Work

While the results are statistically significant, there are some limitations that need to be considered with the results. One limitation is that the sample sizes were relatively small compared to other studies on RF. Larger sample sizes could allow for better classification categories, such as high shear low cape (HSLC) and other traditional PS classification methods used in severe weather prediction. It would also potentially help the PC models help more accurately explain the variances within the dataset compared to projecting one year's EOFs onto another years. Additionally, the entire 24-hour period was average/use for the PS and PC model rather than the period associated with the highest severe occurrence. Some cases may have had higher instability or shear if a smaller time average (e.g., 4-hour average) than the 24-hour average, allowing for more accurate classifications of cases. Future work will include the impacts of this methodology on the forecasting of different severe hazards (tornado, wind, hail, and significant severe of the same type). Another piece of future work is to try to use 4-hour winds to forecast severe weather rather than the 24-hour period and see if model performance remains positive and significant. The potential improvements in RF forecast by FD model training could help the forecasting of severe weather, which in the end could help protect life and property from this most frequent hazardous weather.

References

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting Hail Using a One-Dimensional Hail Growth Model within WRF. *Monthly Weather Review*, **144 (12)**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1, URL https://journals.ametsoc.org/view/journals/mwre/144/12/mwr-d-16-0027.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

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. *Weather and Forecasting*, **31 (2)**, 581–599, https://doi.org/10.1175/WAF-D-15-0113.1, URL https://journals.ametsoc.org/view/journals/wefo/31/2/waf-d-15-0113_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Alvarez, F., 2014: Statistical calibration of extended-range probabilistic tornado forecasts with a reforecast dataset. URL https://www.semanticscholar.org/paper/Statistical-calibration-of-extended-range-tornado-a-Alvarez/2d862407774de393c679bfa0a665c6c82c383a74.

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2016: Investigation of Near-Storm Environments for Tornado Events and Warnings. *Weather and Forecasting*, **31 (6)**, 1771–1790, https://doi.org/10.1175/WAF-D-16-0046.1, URL https://journals.ametsoc.org/view/journals/wefo/31/6/waf-d-16-0046_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Berner, J., K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the Skill of Probabilistic Forecasts: Understanding Performance Improvements from Model-Error Representations. *Monthly Weather Review*, **143 (4)**, 1295–1320, https://doi.org/10.1175/MWR-D-14-00091.1, URL https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00091.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

Bhuiyan, M. A. E., E. I. Nikolopoulos, and E. N. Anagnostou, 2019: Machine Learning–Based Blending of Satellite and Reanalysis Precipitation Datasets: A Multiregional Tropical Complex Terrain Evaluation. *Journal of Hydrometeorology*, **20 (11)**, 2147–2161, https://doi.org/10.1175/JHM-D-19-0073.1, URL https://journals.ametsoc.org/view/journals/hydr/20/11/jhm-d-19-0073_1.xml, publisher: American Meteorological Society Section: Journal of Hydrometeorology.

Brimelow, J. C., G. W. Reuter, R. Goodson, and T. W. Krauss, 2006: Spatial Forecasts of Maximum Hail Size Using Prognostic Model Soundings and HAILCAST. *Weather and Forecasting*, **21 (2)**, 206–219, https://doi.org/10.1175/WAF915.1, URL https://journals.ametsoc.org/view/journals/wefo/21/2/waf915_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Bukovsky, 2011: Bukovsky Regions. URL https://www.narccap.ucar.edu/contrib/bukovsky/.

Burke, A., N. Snook, D. J. G. Ii, S. McCorkle, and A. McGovern, 2020: Calibration of Machine Learning–Based Probabilistic Hail Predictions for Operational Forecasting. *Weather and Forecasting*, **35 (1)**, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1, URL https://journals.ametsoc.org/view/journals/wefo/35/1/waf-d-19-0105.1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Clark, A. J., and E. D. Loken, 2022: Machine Learning–Derived Severe Weather Probabilities from a Warn-on-Forecast System. *Weather and Forecasting*, **37 (10)**, 1721–1740, https://doi.org/10.1175/WAF-D-22-0056.1, URL https://journals.ametsoc.org/view/journals/wefo/37/10/WAF-D-22-0056.1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, **32 (5)**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1, URL https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0010_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended Probabilistic Tornado Forecasts: Combining Climatological Frequencies with NSSL–WRF Ensemble Forecasts. *Weather and Forecasting*, **33 (2)**, 443–460, https://doi.org/10.1175/WAF-D-17-0132.1, URL https://journals.ametsoc.org/view/journals/wefo/33/2/waf-d-17-0132_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Gasperoni, N. A., X. Wang, and Y. Wang, 2022: Using a Cost-Effective Approach to Increase Background Ensemble Member Size within the GSI-Based EnVar System for Improved Radar Analyses and Forecasts of Convective Systems. *Monthly Weather Review*, **150 (3)**, 667–689, https://doi.org/10.1175/MWR-D-21-0148.1, URL https://journals.ametsoc.org/view/journals/mwre/150/3/MWR-D-21-0148.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

Hall, T. J., C. N. Mutchler, G. J. Bloy, R. N. Thessin, S. K. Gaffney, and J. J. Lareau, 2011: Performance of Observation-Based Prediction Algorithms for Very

Short-Range, Probabilistic Clear-Sky Condition Forecasting. *Journal of Applied Meteorology and Climatology*, **50 (1)**, 3–19, https://doi.org/10.1175/2010JAMC2529.1, URL https://journals.ametsoc.org/view/journals/apme/50/1/2010jamc2529.1.xml, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

Herman, G. R., and R. S. Schumacher, 2018: "Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation. *Monthly Weather Review*, **146 (6)**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1, URL https://journals.ametsoc.org/view/journals/mwre/146/6/mwr-d-17-0307.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. *Monthly Weather Review*, **148 (5)**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1, URL https://journals.ametsoc.org/view/journals/mwre/148/5/mwr-d-19-0344.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting Excessive Rainfall with Random Forests and a Deterministic Convection-Allowing Model. *Weather and Forecasting*, **36 (5)**, 1693–1711, https://doi.org/10.1175/WAF-D-21-0026.1, URL https://journals.ametsoc.org/view/journals/wefo/36/5/WAF-D-21-0026.1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Hohenegger, C., and C. Schar, 2007: Atmospheric Predictability at Synoptic Versus Cloud-Resolving Scales. *Bulletin of the American Meteorological Society*, **88 (11)**, 1783–1794, https://doi.org/10.1175/BAMS-88-11-1783, URL https://journals.ametsoc.org/view/journals/bams/88/11/bams-88-11-1783.xml, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind. *Weather and Forecasting*, **32 (6)**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1, URL https://journals.ametsoc.org/view/journals/wefo/32/6/waf-d-17-0038_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Ließ, M., J. Schmidt, and B. Glaser, 2016: Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. *PloS One*, **11 (4)**, e0153 673, https://doi.org/10.1371/journal.pone.0153673.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests. *Weather and Forecasting*, **35 (4)**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1, URL https://journals.ametsoc.org/view/journals/wefo/35/4/wafD190258.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Loken, E. D., A. J. Clark, and A. McGovern, 2022: Comparing and Interpreting Differently Designed Random Forests for Next-Day Severe Weather Hazard Prediction. *Weather and Forecasting*, **37 (6)**, 871–899, https://doi.org/10.1175/WAF-D-21-0138.1, URL https://journals.ametsoc.org/view/journals/wefo/37/6/WAF-D-21-0138.1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Marzban, C., and G. J. Stumpf, 1996: A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes. *Journal of Applied Meteorology*, **35 (5)**, 617–626, https://doi.org/10.1175/1520-0450(1996)035⟨0617:ANNFTP⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0450(1996)035⟨0617:ANNFTP⟩2.0.CO;2.

Marzban, C., and G. J. Stumpf, 1998: A Neural Network for Damaging Wind Prediction. *Weather and Forecasting*, **13 (1)**, 151–163, https://doi.org/10.1175/1520-0434(1998)013⟨0151:ANNFDW⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/13/1/1520-0434_1998_013_0151_annfdw_2_0_co_2.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Marzban, C., and A. Witt, 2001: A Bayesian Neural Network for Severe-Hail Size Prediction. *Weather and Forecasting*, **16 (5)**, 600–610, https://doi.org/10.1175/1520-0434(2001)016⟨0600:ABNNFS⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/16/5/1520-0434_2001_016_0600_abnnfs_2_0_co_2.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: DOES INCREASING HORIZONTAL RESOLUTION PRODUCE MORE SKILLFUL FORECASTS?: The Results of Two Years of Real-Time Numerical Weather Prediction over the Pacific Northwest. *Bulletin of the American Meteorological Society*, **83 (3)**, 407–430, https://doi.org/10.1175/1520-0477(2002)083⟨0407:DIHRPM⟩2.3.CO;2, URL https://journals.ametsoc.org/view/journals/bams/83/3/1520-0477_2002_083_0407_dihrpm_2_3_co_2.xml, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, **100 (11)**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1, URL https://journals.ametsoc.org/view/journals/bams/100/11/bams-d-18-0195.1.xml, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Melhauser, C., and F. Zhang, 2012: Practical and Intrinsic Predictability of Severe and Convective Weather at the Mesoscales. *Journal of the Atmospheric Sciences*, **69 (11)**, 3350–3371, https://doi.org/10.1175/JAS-D-11-0315.1, URL https://journals.ametsoc.org/view/journals/atsc/69/11/jas-d-11-0315.1.xml, publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

Munich, 2023: Hail, tornadoes, flash floods: Losses from thunderstorms on the rise | Munich Re. URL https://www.munichre.com/en/risks/natural-disasters/thunderstorms-hail-tornados.html.

NOAA, 2017: Calculating the Cost of Weather and Climate Disasters. URL https://www.ncei.noaa.gov/news/calculating-cost-weather-and-climate-disasters.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Weather and Forecasting*, **24 (2)**, 601–608, https://doi.org/10.1175/2008WAF2222159.1, URL https://journals.ametsoc.org/view/journals/wefo/24/2/2008waf2222159_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing Forecast Error in a Convection-Permitting Ensemble System. *Monthly Weather Review*, **142 (12)**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1, URL https://journals.ametsoc.org/view/journals/mwre/142/12/mwr-d-14-00100.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

Romps, D. M., 2016: Clausius–Clapeyron Scaling of CAPE from Analytical Solutions to RCE. *Journal of the Atmospheric Sciences*, **73 (9)**, 3719–3737, https://doi.org/10.1175/JAS-D-15-0327.1, URL https://journals.ametsoc.org/view/journals/atsc/73/9/jas-d-15-0327.1.xml, publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

Saabas, A., 2023: TreeInterpreter. URL https://github.com/andosa/treeinterpreter, original-date: 2015-08-02T20:26:21Z.

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and Optimizing Precipitation Forecasts from a Convection-Permitting

Ensemble Initialized by a Mesoscale Ensemble Kalman Filter. *Weather and Forecasting*, **29 (6)**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1, URL https://journals.ametsoc.org/view/journals/wefo/29/6/waf-d-13-00145_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective Modes for Significant Severe Thunderstorms in the Contiguous United States. Part I: Storm Classification and Climatology. *Weather and Forecasting*, **27 (5)**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1, URL https://journals.ametsoc.org/view/journals/wefo/27/5/waf-d-11-00115_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Sobash, R. A., G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016: Explicit Forecasts of Low-Level Rotation from Convection-Allowing Models for Next-Day Tornado Prediction. *Weather and Forecasting*, **31 (5)**, 1591–1614, https://doi.org/10.1175/WAF-D-16-0073.1, URL https://journals.ametsoc.org/view/journals/wefo/31/5/waf-d-16-0073_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

SPC, 2023: Storm Prediction Center Maps, Graphics, and Data Page. URL https://www.spc.noaa.gov/wcm/.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-Resolving Ensemble Simulations of Mediterranean Heavy Precipitating Events: Uncertainty on Initial Conditions and Lateral Boundary Conditions. *Monthly Weather Review*, **139 (2)**, 403–423, https://doi.org/10.1175/2010MWR3487.1, URL https://journals.ametsoc.org/view/journals/mwre/139/2/2010mwr3487.1.xml, publisher: American Meteorological Society Section: Monthly Weather Review.

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. *Weather and Forecasting*, **23 (3)**, 407–437, https://doi.org/10.1175/2007WAF2007005.1, URL https://journals.ametsoc.org/view/journals/wefo/23/3/2007waf2007005_1.xml, publisher: American Meteorological Society Section: Weather and Forecasting.

Weisman, M. L., and J. B. Klemp, 1982: The Dependence of Numerically Simulated Convective Storms on Vertical Wind Shear and Buoyancy. *Monthly Weather Review*, **110 (6)**, 504–520, https://doi.org/10.1175/1520-0493(1982)110⟨0504:TDONSC⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/110/6/1520-0493_1982_110_0504_tdonsc_2_0_co_2.xml, publisher: American Meteorological Society Section: Monthly Weather Review.