DISSERTATION

IMPROVING PREDICTIONS AND GENERATING ACTIONABLE FORECAST INSIGHTS FOR DOWNSLOPE WINDSTORMS WITH MACHINE LEARNING

Submitted by

Casey L. Zoellick

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2025

Doctoral Committee:

Advisor: Russ Schumacher

Kristen Rasmussen
Elizabeth Barnes
Peter Nelson

ABSTRACT

IMPROVING PREDICTIONS AND GENERATING ACTIONABLE FORECAST INSIGHTS FOR DOWNSLOPE WINDSTORMS WITH MACHINE LEARNING

Downslope windstorms are an extreme weather phenomenon characterized by accelerating winds down the lee slope of a mountain with gusts often exceeding 45 m s^{-1}. These impact society through damage directly related to the high winds, ground transportation concerns in the vicinity of the windstorm, aviation impacts through the accompanying mountain wave turbulence, and fueling the rapid intensification and spread of wildfires such as the 2018 Camp Fire, the 2021 Marshall Fire, and the 2023 Lahaina Fire. Despite improvements in numerical weather prediction and observational datasets, predictability of these windstorms still rarely exceeds 12 hours further exacerbating their impacts. Recent advances have made machine learning (ML) more accessible to researchers and have shown promise in improving forecasts of other extreme weather phenomena.

We first present models driven by two different types of ML architectures that classify wind events as moderate or high at three locations along the Rocky Mountain Front Range: Cheyenne, Wyoming; Fort Collins, Colorado; and Boulder, Colorado. The first type of architecture is the random forest (RF), which is comprised of multiple decision trees, and the second type is the convolutional neural network (CNN), which is a deep learning method that excels at image recognition. These models make forecasts at the Day 1 and Day 2 lead times based on predictors derived from a 12-km version of the WRF operated at Colorado State University. The results show improvement over the direct weather model forecasts. CNNs show enhanced event detection capability compared to the RFs but with a higher false alarm rate limiting their utility in some cases.

Next, explainable artificial intelligence (XAI) techniques are presented. Feature importances indicate that the ML models rely on predictors at geographic locations that align with known atmospheric variables important to downslope wind forecast along the terrain. Also, a framework

for reducing the dimensions of the predictor data and clustering these data with a Gaussian mixture model yields insights to the forecast ML models' performance and the synoptic conditions in which downslope windstorms along the Front Range occur. The ML models perform better in regimes characterized by prominent synoptic features such as cold air advection or the presence of the jet stream aloft.

Lastly, we investigate whether increasing the resolution of the traditional weather model creating the ML predictors results in performance improvements. We use NOAA's High Resolution Rapid Refresh (HRRR) model to derive input predictors for newly trained CNNs and observe a decrease in false alarms that results in an overall performance boost over the direct HRRR forecasts. A case study on the Marshall Fire is conducted and indicates that the HRRR-based CNN is able to correctly forecast the subsequent downslope wind event before the wind event is explicitly depicted in the HRRR output itself. This study is an example of how ML fused with current weather models closes the forecast gap in these impactful weather phenomena with incomplete physical understandings.

ACKNOWLEDGEMENTS

I would like to acknowledge the Air Force Institute of Technology (AFIT) for sponsoring my degree program and providing me the opportunity to teach and continue my research as faculty following graduation. I would like to thank Col Rose Tseng and Lt Col Bob Tournay (retired) for mentoring me after my master's degree and encouraging me to pursue my PhD at Colorado State University. I would also like to thank Dr. Sam Childs at the 16th Weather Squadron (16 WS) and former Schumacher Group member for collaborating with me on my research and the ongoing efforts at the 16 WS, which I used to be affiliated with. Finally, I would like to thank all the members of the Schumacher Group during my tenure for providing stimulating discussion on research problems and an avenue for mental breaks.

I would like to thank my family as without their love and support I would not be able to continue to achieve my goals in academia and my Air Force career. My parents, Jeff and Cindy, have always supported me and continue to assist with moves and other emerging situations when I have been out of the country in the past. My wife, Sarah, and son, Zephyr, provide the reason I keep going every day working on the current goal and looking ahead to the next. I love you both very much!

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Downslope windstorms are an extreme weather phenomenon characterized by strong, gusty, and violent winds that descend the lee slope of a mountain capable of achieving EF1 to EF2 equivalent damage on the Enhanced Fujita scale (American Meteorological Society 2025). The conditions required for such windstorms generally include a deep layer of air flowing perpendicular to a terrain barrier with a stable layer present near the top of the ridgeline (Markowski and Richardson 2010). These events are observed around the world including in Europe, Japan, and South America (Klemp and Lilly 1975; Otero and Araneo 2021). Besides the risk posed by the damage caused these windstorms, these onset of these events is difficult to forecast, and the predictability of the intensity beyond 12 hours remains low despite improvements in observations and numerical weather prediction (Reinecke and Durran 2009).

Additionally, these windstorms pose threats to transportation both on the ground and in the air. These high winds can impact vehicles on roadways in the vicinity of these events that is especially hazardous to high-profile lightweight trucks (Brothers and Hammer 2023). In the aviation sector, downslope windstorms are often accompanied by mountain waves that produce severe turbulence that all types of aircraft must avoid. After many aircraft incidences over mountainous terrain occurred in the middle of the 20th century, in-situ observations of these events and subsequent modeling studies characterized the threat to aircraft through most of the troposphere ensure aircraft avoided these areas by flying around them and not attempting to fly above the mountain wave (Lilly and Kennedy 1973; Klemp and Lilly 1975; Lilly 1978). Further advancements in numerical weather prediction and observations allows for the forecasting and identification of mountain waves, especially the more severe events (Smith 2019). However, the existence of a mountain wave does not guarantee that high winds are reaching the surface. Thus, the forecast advancements of mountain waves has not directly translated to an increase in the predictability of downslope windstorms.

Two of the biggest threats associated with downslope windstorms are wildfire spread and intensification. In December 2017, the Santa Ana winds in Southern California fuel the Thomas

Fire, which briefly became the largest fire by total area burned in California's history (Fovell and Gallagher 2018). Less than a year later in November 2018, a downslope windstorm in the western Sierra Nevada (sometimes referred to as the North wind to distinguish it from the Diablo winds that occur on the other side of the mountains in the San Francisco Bay area) rapidly propelled the Camp Fire into the communities of Concow, Paradise, and Magalia, CA that destroyed 19,000 structures and became the deadliest wildfire in California's history. The 2021 Marshall Fire in Colorado was pushed into a suburban area by severe winds descending the lee slope of the Front Range near Boulder, CO. This resulted in the most destruction cost-wise caused by a wildfire in the state's history (Fovell et al. 2022; Benjamin et al. 2023). These downslope wind-driven fires are not contained to the midlatitudes as evidenced by the 2023 Lahaina Fire that killed at least 100 people and caused an economic loss of 4-6 billion dollars. This fire quickly spread down the western slopes of the West Maui Mountains into the town of Lahaina before reaching the ocean (Juliano et al. 2023a; Mass and Ovens 2024). While downslope windstorms are generally associated with the fall, winter, and spring months due to the synoptic conditions necessary to produce these windstorms, they can occur throughout the year. These events then potentially coincide with the drought-prone months as was the case with the June 2012 High Park Fire and 2013 Black Canyon Fire in Colorado (Coen and Schroeder 2015). Concerns for these fires exacerbated by windstorms continue to grow as the wildfire-urban interface rapidly expands and conditions favorable for fire weather become more common due to climate change (Bowman et al. 2020; Juliano et al. 2023b). Many other cases of these dual downslope wind and fire events exists, the ones discussed are some of the more destructive examples in recent years that motivated the research in this study.

1.1 Downslope Windstorm Theory

The main focus of downslope wind theory concerns hydraulic jumps and their application to the real atmosphere in proximity to steeply sloped mountains. As a layer of fluid suddenly encounters a barrier and a decrease in elevation hydraulic jumps are possible. Hydraulic jumps occur as subcritical flow transitions to supercritical flow as it encounters a barrier and accelerates down the lee slope of the obstacle. This creates the jump and converts the potential energy of the subcritical

flow into kinetic energy (Durran 1990). The Froude number (Fr), which can be used to diagnosed the flow regime, is the ratio between the mean flow and the gravity wave phase speed in a fluid:

$$\mathrm{Fr}^2 = \frac{u^2}{c^2} = \frac{u^2}{gD}, \tag{1.1}$$

where u is the mean flow, c is the intrinsic shallow-water gravity wave phase speed ($c = \sqrt{gD}$ for shallow water systems), g is gravity, and D is the depth of the fluid (Durran 1990; Markowski and Richardson 2010). When Fr is close to unity, hydraulic jump conditions are possible. However, this theory assumes a free surface at the top of the fluid, which is not the case in the real atmosphere that does not behave as a two-layer fluid and is continually stratified (Markowski and Richardson 2010; Smith 2019). Thus, this theory is too restrictive in its assumptions to explain the observed phenomena (Klemp and Lilly 1975).

Atmospheric stratification and standing waves also play a role based on observations and numerical modeling of downslope windstorms. Specifically, early numerical models showed severe wind orientations when waves would break aloft and create amplification through downward reflection from a critical level (Smith 2019). This condition is favorable when a layer of strong static stability exists near the mountaintop with a layer of lesser stability above (Markowski and Richardson 2010). The Scorer parameter, l, is used to diagnose whether the atmosphere at and downstream of a terrain barrier might produce trapped lee waves:

$$l^2 \equiv \frac{N^2}{\bar{u^2}} - \frac{1}{\bar{u}}\frac{d^2\bar{u}}{dz^2}, \tag{1.2}$$

where N is the Brunt-Väisälä frequency, $\bar{u}$ is the mean cross-barrier wind velocity, and z is the height coordinate (Scorer 1949; Markowski and Richardson 2010; Smith 2019). Since the ratio of the vertical perturbance of the mean horizontal flow with height to the mean flow itself is generally small before wave breaking occurs, l is predominantly dictated by the ratio of N, which quantifies stability, to $\bar{u}$. When two layers of the atmosphere have two different Scorer parameters such that $l^2_{upper} < l^2_{lower}$ and l^2_{lower} allows vertically propagating waves in that layer, it is possible for these

waves to duct or be trapped when encountering the interface with the upper layer. The conditions are favorable at mountain barriers where the mean flow suddenly encounters another layer of atmosphere in the lee of the peak with different properties. In practice, trapped lee waves are occur in an atmosphere where l^2 decreases strongly with height (Smith 2019). However, downslope windstorms can occur without this mean-state critical level, and recent research has shown these types of events may be less predictable (Metz and Durran 2023). The combination of standing waves, the physics of hydraulic jumps, and how they present in the real atmosphere makes forecasting downslope windstorms difficult (Durran 1990). Furthermore, finite wave amplitude effects, the nonlinear portion of the Scorer parameter, are not accounted for in gravity wave drag parameterizations, and these effects are more significant before wave breaking occurs than once thought (Metz and Durran 2021).

1.2 Forecasting Downslope Windstorms

Forecasters look for a number of indicators to anticipate the formation of downslope windstorms. First, these events occur in areas of terrain with a steep slope in the lee of the predominant wind pattern. Second, synoptic conditions with high pressure upstream and low pressure downstream create a favorable pressure gradient across the mountain barrier to pull the winds to the lower elevations. Third, the wind component perpendicular to the mountain should exceed 15 m s^{-1}. This means that flows whose directions exceed 30° out of perpendicular with the terrain are not favorable for downslope windstorms. Fourth is the presence of a stable layer near the mountaintop that acts to promote the trapping of lee waves and push the winds down the slope of the mountain. This may include a level with weak cross-barrier flow or flow reversal. Lastly, cold air advection and anticyclonic vorticity advection promote downward vertical motion in the atmosphere that aid in the formation of a downslope windstorm though not required (Markowski and Richardson 2010). Despite advancements in downslope windstorm theory, numerical weather prediction, and observations, forecasting these events remains difficult. The criteria listed above for forecasting downslope windstorms are simple enough to diagnose in any observational or model forecast data. However, these conditions exists far more often than downslope windstorms occur,

so these criteria only point to the possibility of a windstorm occurring. To forecast the timing an intensity of a possible downslope windstorm, forecasters look to other data.

1.3 Motivation

Recent advancements in machine learning (ML) combined with the ongoing improvement of traditional weather models have made new numerical predictions method more obtainable to researchers. The motivation for this study includes four parts. The first part is the application of ML techniques to physics-informed weather models to improve the forecasts and extend the predictability of an extreme weather phenomenon, downslope windstorms. The impacts of these storms listed above combined with the ongoing struggle to accurately forecast their occurrence make them a compelling phenomenon for the applications described in this study. We aim to utilize ML to close the gap in our physical understanding of downslope windstorms represented in traditional weather models and reality.

The second part of the motivation is that the ML techniques themselves must be adaptable and flexible. The methodology used is designed to be able to be applied anywhere in the world and potentially for other extreme phenomenon. The first ML models described learn from predictors supplied by a 12-km weather model. This is in the range of global weather model resolution so that the ML models do not require high resolution models with specifically tuned parameterizations available in only one geographic area. This is potentially useful to military forecasters who often must forecast for new locations based on emerging operational requirements that are data sparse. By using a global model resolution, these data are available across the globe.

The third part of the motivation is that the ML models must be trustworthy. Part of trustworthiness includes understanding how the models arrive at their predictions, but it also encompasses the relevance of the models' forecasts. This is accomplished through verification, and these methods must also be transparent and understandable to the end user. If a model can accurately identify downslope windstorms but also forecasts them too often when they do not occur, this model is not trustworthy and of very little use to a human forecaster.

Generating actionable insights with ML techniques is the final part of this study's motivation. The insights extend beyond the output predictions of the ML models. We address how to use ML to categorize the meteorological scenarios in which these downslope windstorms occur. Also, insights into the ML models' performances in different scenarios can assist the forecaster on whether or not to trust the model's forecast on a particular day. We define actionable to mean that the insight provides guidance to a forecaster that can be used to make the final forecast decision. This also implies that the insight could conceivably be generated in real time during the forecast process, and not something uncovered through analysis after the fact.

Throughout the study, we often refer to the results and conclusions in the context of the forecast problem from the perspective of a forecaster. The forecast problem is predicting the occurrence of downslope windstorms and is further defined in the next chapter. This is not a formal research-to-operations (R2O) study, and no formal relationships or feedback are gathered from operational forecasters. However, the forecaster's perspective is necessary to frame the efficacy of the insights to applying the ML model predictions to the forecast problem. Thus, the forecaster is referenced in the general sense, and not all the techniques described below are applicable to every end user even within the impacts from downslope windstorms. Formal work in the R2O realm would be required to validate these techniques in an operational environment. This study aims to present techniques that scope future work.

This study is structured as follows: Chapter 2 defines the forecast problem and presents the ML models developed to forecast downslope windstorms along the Front Range of the Rocky Mountains. This chapter also describes the verification of these models and concerns regarding the application of these verification techniques. Chapter 3 outlines a framework for generating actionable forecast insights through the use of XAI techniques including methods described in the literature and another method that leverages unsupervised ML and cluster analysis. A case study is presented detailing how the application of these insights is useful in a forecast setting. Chapter 4 investigates whether increasing the resolution of the input predictors to that of a mesoscale model improves from the forecasts supplied by the lower resolution-driven ML models. A second case

study on the Marshall Fire shows the potential utility of the ML models compared to the traditional weather model forecasts available at the time. The conclusion of the study encompasses the final chapter.

CHAPTER 2: FORECASTING DOWNSLOPE WINDSTORMS WITH RANDOM FORESTS AND CONVOLUTIONAL NEURAL NETWORKS

2.1 Introduction

As discussed in the introduction, downslope windstorms threaten significant property damage and utility disruptions for populations living in the lee of mountainous terrain. Additionally, these windstorms rapidly intensify and spread wildfires at paces never previously observed as evidenced by the 2017 Thomas Fire and 2018 Camp Fire in California, the 2021 Marshall Fire in Colorado, and the 2023 Lahaina Fire in Hawaii (Fovell and Gallagher 2018; Brewer and Clements 2020; Fovell et al. 2022; Mass and Ovens 2024). High winds also pose threats to motorists on roadways impacting any commerce conducted along interstates in the vicinity of these events (Brothers and Hammer 2023). Despite advances in the understanding of the underlying mechanisms of these events and increases in the accuracy and resolution of numerical weather prediction models, the onsets of the windstorms remains difficult to predict and the confidence in the intensity forecasts beyond 12 hours remains low (Reinecke and Durran 2009). Furthermore, the mountainous terrain itself presents challenges as these regions tend to be observation-sparse and poorly resolved in forecast models.

Advancements in recent decades in ML have made methods for improving the forecasts of extreme weather phenomena more accessible to researchers. Herman and Schumacher (2018) successfully trained skillful random forests to forecast extreme precipitation above specific average recurrence intervals using reforecasts from the Global Ensemble Forecast System (GEFS/R). Studies have used also RFs to forecast large hail using predictors derived from convection-allowing ensembles (Gagne et al. 2017) and ERA5 reanalysis data (Czernecki et al. 2019). Trafalis et al. (2014) compared different ML classifiers' performances on detecting tornadic mesocyclones based on characteristics of the storm and its environment and found that the support vector machines were the most successful. Furthermore, RFs have utilized to synthesize radar information during tornado outbreaks to increase warning lead time by simplifying the forecaster decision-making process (Sandmæl et al. 2023). Lastly, weather phenomena impacting aviation operations have

also been targeted including fog and visibility forecasts (Herman and Schumacher 2016) and upper level turbulence (McGovern et al. 2014; Muñoz-Esparza et al. 2020). This has also led to the implementation of more machine learning-driven forecasts into forecast operations. For example, in recent years, RF-generated probabilistic forecasts for severe weather and excessive rainfall have been implemented into the forecast operations of the Storm Prediction Center and the Weather Prediction Center, respectively (Hill et al. 2023; Schumacher et al. 2021).

There has been particular focus on applying ML to the Warn-on-Forecast System (WoFS) output, which itself is an advancement in forecasting severe weather through the use of convection-allowing ensembles that update up to four times an hour (Clark and Loken 2022). These efforts have mainly been post-processing routines driven by logistical regression or RFs to create probabilistic forecasts for severe weather phenomena at the storm scale or to predict the skill of the WoFS ensemble itself within the current meteorological environment (Flora et al. 2021; Clark and Loken 2022; Flora et al. 2024; Potvin et al. 2024). Due to the speed of the ML algorithms at inference time, they are able to run after rapidly updating numerical weather prediction output without delaying the end product to the forecaster. This makes ML-based algorithms attractive within the warning-decision timeframe compared to other post-processing routine that require more computation time.

Forecasts enhanced or created with ML have focused on wind prediction as well. Lagerquist et al. (2017) employed a variety of ML methods to forecast damaging straight-line convective winds and achieved improvements in the critical success index (CSI), Brier Skill Score (BSS), and the area under the receiver operating characteristic (ROC) curve though the best values occurred closest to the storms themselves and at very short lead times (0-15 min). Wind gust probabilities at three airports derived from artificial neural networks (ANNs) trained on ERA5 data proved to outperform similar techniques with logistical and multiple linear regression especially during the cold season (Coburn and Pryor 2022). More specific to the Front Range, downslope windstorms in Boulder, CO were forecast using ANNs, support vector regression (SVR), and stepwise linear regression, and SVR was determined to be the optimal technique (Mercer et al. 2008). In addition,

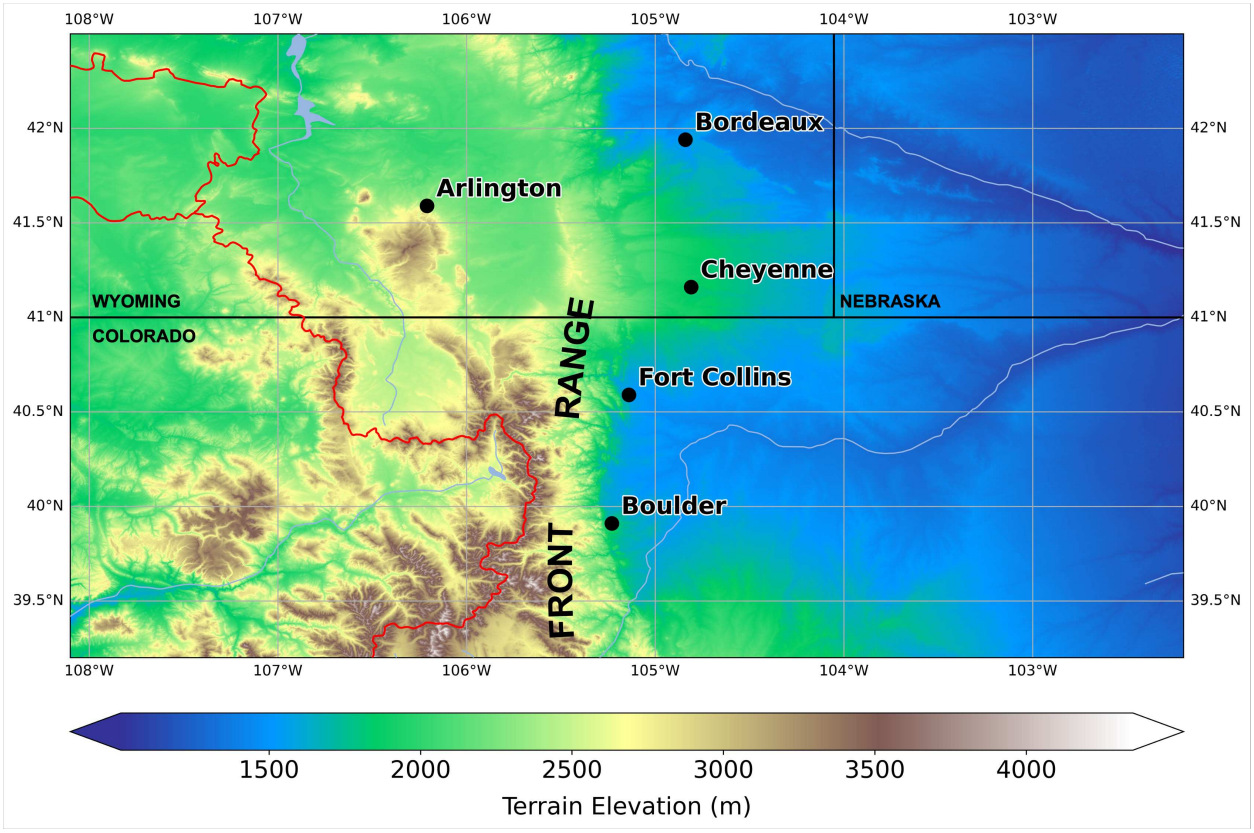

Figure 2.1: Terrain map depicting the Front Range locations discussed in Section 2.1. The continental divide is outlined in red. Terrain data from the United States Geological Survey (2021).

the National Weather Service (NWS) tested RF high wind forecasts at three locations: Arlington, Bordeaux, and Cheyenne, WY. The former two locations experience frequent gap winds due to pressure gradients across the mountains while Cheyenne experiences downslope winds in addition to pressure gradient driven winds from midlatitude cyclones transiting across the high plains. The results showed improvements over more traditional forecasting techniques used by the NWS Cheyenne forecast office (Brothers and Hammer 2023). Figure 2.1 presents a map of these locations relative to the terrain of the Front Range.

Most of the aforementioned studies applied RFs or statistical regression to their forecast problems. When forecasting extreme weather phenomena in the mesoscale, there appears to be less emphasis on using ANNs and even less so on convolutional neural networks (CNNs). CNNs excel in image recognition, which may seem less relevant when creating forecasts (LeCun et al. 2015). However, satellite, radar, and traditional model output fields can all be thought of as images making

CNNs highly suited to process these images. The use of these networks has seen wider uptake in climatologic studies including detection of extremes in climatological datasets and front recognition in climate simulations (Molina et al. 2023; Chase et al. 2023). Studying the performance of CNNs on downslope wind forecasting is valid given the use-case applicability and similar nature of the forecast problem as seen in climate studies specifically extreme detection.

Work by Lindsey et al. (2011) who created a statistical regression model for forecasting high winds at Christman Airfield in the foothills west of Fort Collins (shown in Figure 2.1) motivates this study, which aims to implement RF and CNN models to improve upon the previous techniques. This study expands the forecasting scope by including forecasts for moderate and high wind events at three locations along the Front Range of the Rocky Mountains. We hypothesize three main results after evaluating these ML models:

1. The ML models will be able to extend the lead time or predictability of these high wind events beyond 24 hours compared to traditional weather model output.

2. The ML models will perform poorest when a location has a small sample size of wind events in the climatology and the dataset is highly imbalanced.

3. CNNs will perform better than the RFs due to their image processing strength and ability to adapt to nonlinearities inherent to this forecasting problem.

The rest of this chapter is structured as follows: Section 2.2 provides background on location selection, creation of the predictors and labels used to train the models, and the architecture and training process of the models themselves. Section 2.3 discusses the metrics employed to verify the performances of the ML models and the baseline numerical weather prediction model. Section 2.4 presents the results of the performances of the models over a two-year test period that is not included in the training or validation data. Finally, Sections 2.5 and 2.6 provide discussion and conclusions applicable to the models' performances and concerns with using these models in an operational forecast setting.

2.2 Data and Methods

2.2.1 Locations and Labels

This study applies specific criteria to the location selection as we focus on increasing the forecast lead-time of downslope windstorms in particular as opposed other high wind events caused by gap and convective winds. Thus, locations chosen are Cheyenne, WY, Fort Collins, CO, and Boulder, CO. First, these locations sit in the lee of the Front Range that rises dramatically from the high plains creating favorable terrain for these windstorms. Second, these locations boast a robust record of observations that allows for a training period limited only by the traditional weather model data available. Lastly, previous studies have attempted to forecast high winds at these locations and can be utilized for comparisons (Brothers and Hammer (2023) at Cheyenne, Lindsey et al. (2011) at Fort Collins, and Mercer et al. (2008) at Boulder).

Observations for Boulder are obtained from the 10-m sensor on the M2 tower on the Flatirons Campus of the National Renewable Energy Laboratory (NREL) south of the city (Jager and Andreas 1996). As in the Lindsey et al. (2011) study, observations for Fort Collins come from Christman Airfield located in the foothills west of the city via the Colorado Agricultural and Meteorological Network (CoAgMET). Cheyenne Regional Airport provides the observations for Cheyenne that are available from a variety of sources. This study pulled these observations from the Iowa Environmental Mesonet website (Iowa Environmental Mesonet 2022).

This study defines the nonconvective “wind season" as October through April to avoid summertime convective wind events. While convective winds are certainly possible during nonconvective wind season, severe convective winds are rare compared to nonconvective winds during these months at these locations. Similarly, downslope winds do occur during the summer but less frequently. Thus, we focus our attention on the high wind events that occur during the fall, winter, and spring.

The forecast day is defined as 06 UTC until 06 UTC the next day (06 UTC corresponds to 23 MST). As this is a supervised ML task, the models are trained against labels corresponding to thresholds of wind events (and non-events) derived from maximum sustained wind or gust

Table 2.1: Wind speed thresholds in m s^{-1} and miles per hour (mph) used to define non-event, moderate wind events, and high wind events.

Category	Wind Speed (m s^{-1})	Wind Speed (mph)	Wind Gust (m s^{-1})	Wind Gust (mph)
Non-event	≤ 11.1	≤ 25	≤ 15.6	35
Moderate	> 11.1 and ≤ 17.9	> 25 and ≤ 40	> 15.6 and ≤ 25.9	> 35 and ≤ 58
High	> 17.9	> 40	> 25.9	> 58

observation occurring within the forecast day. The problem is setup as a classification task in which the ML models forecast whether a given day will be a non-event, moderate wind event, or a high wind event. The thresholds for each wind category are given in Table 2.1, and each day is labeled accordingly. While this study focuses on high wind events, initial results suggested that adding a moderate category might improve forecast metrics by decreasing false alarms. This also allows the model to identify days where winds are elevated but not reaching speeds associated with the more catastrophic impacts of downslope windstorms.

By assigning each forecast day at each location one of the three possible labels, we can examine the frequency of occurrence of each event given the thresholds chosen. The wind seasons beginning in 2012 through 2018 are the data on which the machine learning models trained. The following two seasons, 2019 and 2020, are the validation data utilized for hyperparameter tuning. We combine the training and validation datasets and present histograms of moderate and high wind events by wind season for each locations in Figure 2.2. While Boulder clearly has significantly more high wind events across all seasons, Cheyenne tends to have more moderate events than the other two locations. Far fewer high wind events occur in Fort Collins that might cause difficulties when training the models with so few cases. Additionally, not shown are the days where no wind event took place that clearly outnumber the days with either wind event. For example, at Fort Collins across all the days in the training and validation period, there are 1461 non-event days, 413 moderate wind days, and 36 high wind days. This illustrates the high degree of class imbalance in the dataset that requires consideration when training ML models.

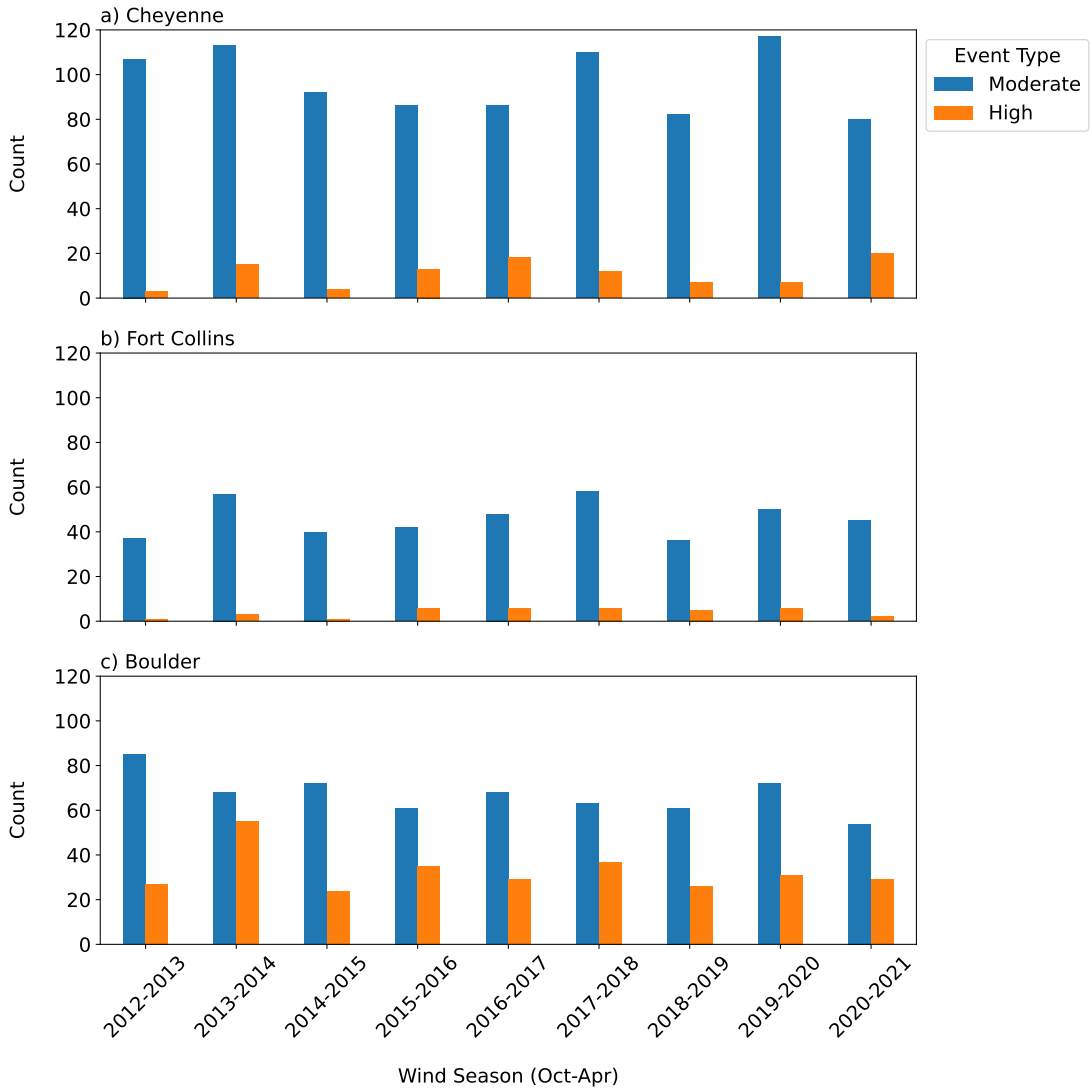

Figure 2.2: Histogram of moderate (blue) and high (orange) wind events at each location by year for the wind seasons beginning in 2012-2020.

Examining the frequency of wind events by month across the training and validation years allows an additional look at the annual pattern of our label dataset, and this is shown in Figure 2.3. The number of high wind events in Boulder and Cheyenne shows a clear ramp up into the peak of the winter season before decreasing as spring emerges. In Fort Collins, the moderate wind events follow this pattern. Due to the fewer amount of high wind events at Fort Collins in the period, no clear signal is evident in that distribution. Interestingly, the moderate wind days at Boulder and Cheyenne are somewhat evenly distributed across the months, except for an increase at the

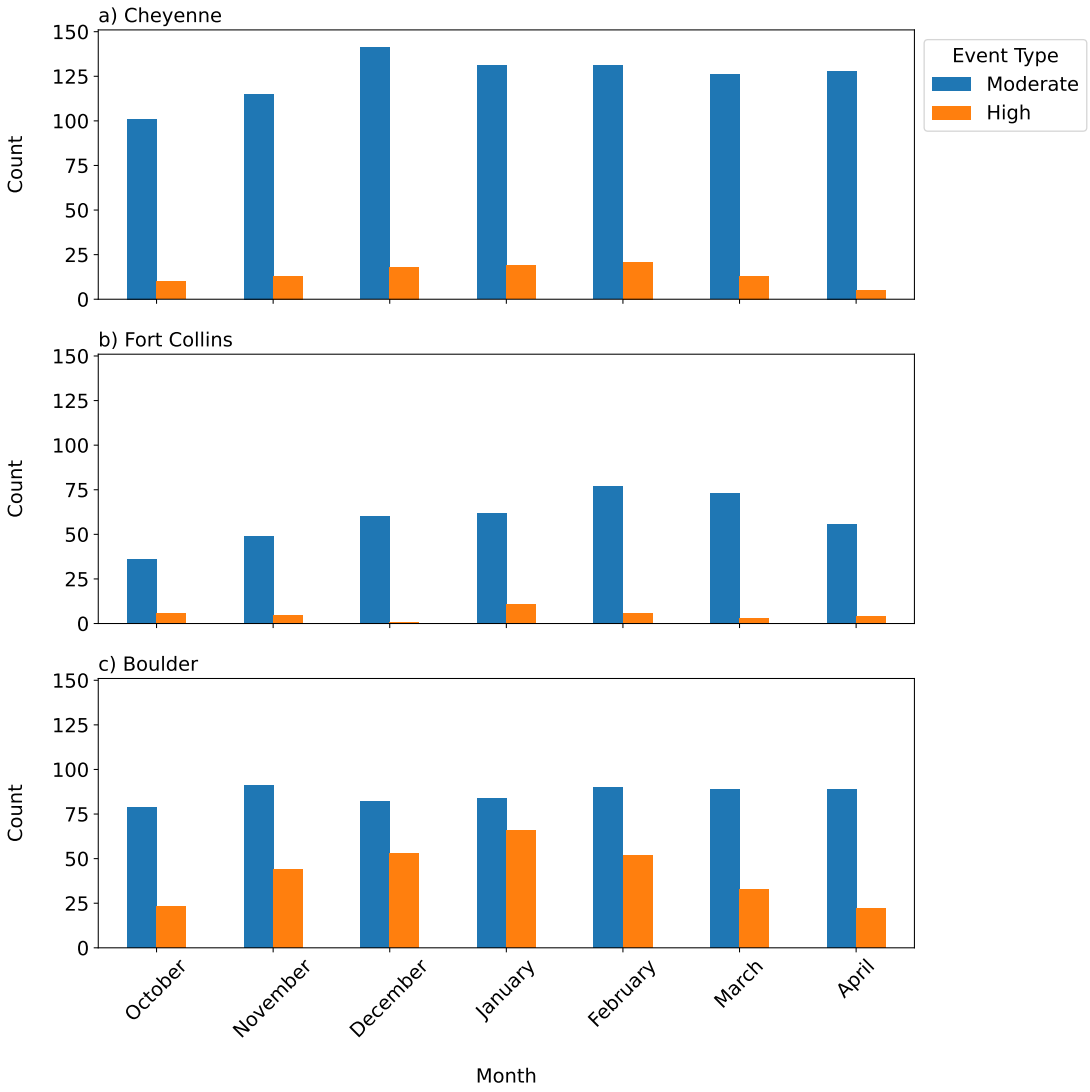

Figure 2.3: Same as in Figure 2.2 with event counts for each month across the wind seasons beginning in 2012-2020.

beginning of the season in Cheyenne. This implies that the driver of the windiest days is correlated to the progression of the annual season but not necessarily for the more moderate windy days. It appears some of the high wind days in Boulder and Cheyenne remain moderate in Fort Collins indicating high wind events might require additional forcing. Thus, the annual pattern seen in the Boulder and Cheyenne high wind events appears in the moderate wind events in Fort Collins.

2.2.2 Input Features

The the input features used to train the ML models are derived from the 12-km WRF model run daily at 00 UTC at Colorado State University (CSU-WRF) consisting of four ensemble members. Specifically, the input training features are derived from the ensemble mean. Figure 2.4 shows the CSU-WRF domain in addition to the domain from which the input features for the ML models are extracted. These features are captured from the original CSU-WRF output GRIB files and regridded to a 10-km latitude-longitude grid (50 by 80 points) before conversion to a NetCDF file. Subsequently, the xarray and numpy python packages are utilized to prepare the input data to be compatible for ingest into the machine learning models. Each day's forecast receives five forecast hours as input data. The Day 1 input data consists of the 06, 12, 18, 24, and 30-hour forecasts from model while Day 2 comprises the 30, 36, 42, 48, and 54-hour forecasts. There are 13 variables that are directly extracted or derived from the 12-km CSU-WRF data, and these are listed in Table 2.2. Thus, there are 65 total forecast maps (13 variables times 5 forecast hours) that make up the input features. These are concatenated together into a single three-dimensional feature cube with dimensions of 50 by 80 by 65. Comparing this two more conventional RGB image recognition applications, this input cube can be thought of as a 50 by 80 pixel image with 65 channels. This results in 260,000 individual predictors that are supplied to the ML models.

The atmospheric variables chosen for inclusion in the input features are motivated by the Lindsey et al. (2011) study. Those statistical regression models ingested point-based or point-differenced data. While this provides simplicity for the regression fit, the skill of the model is likely sensitive to the points chosen. Inputting an entire two-dimensional domain as shown in Figure 2.4 into the model removes this sensitivity by allowing the model itself to optimize the important areas. This creates possibilities for additional analyses on how the ML models make their forecasts. As each point's physical location in the domain has geographic significance in this problem (meaning possible areas of importance would correspond to specific geographical locations), it is possible to better understand which atmospheric variables and locations matter for high wind events at a given location. While this study uses full-gridded data, the variables chosen

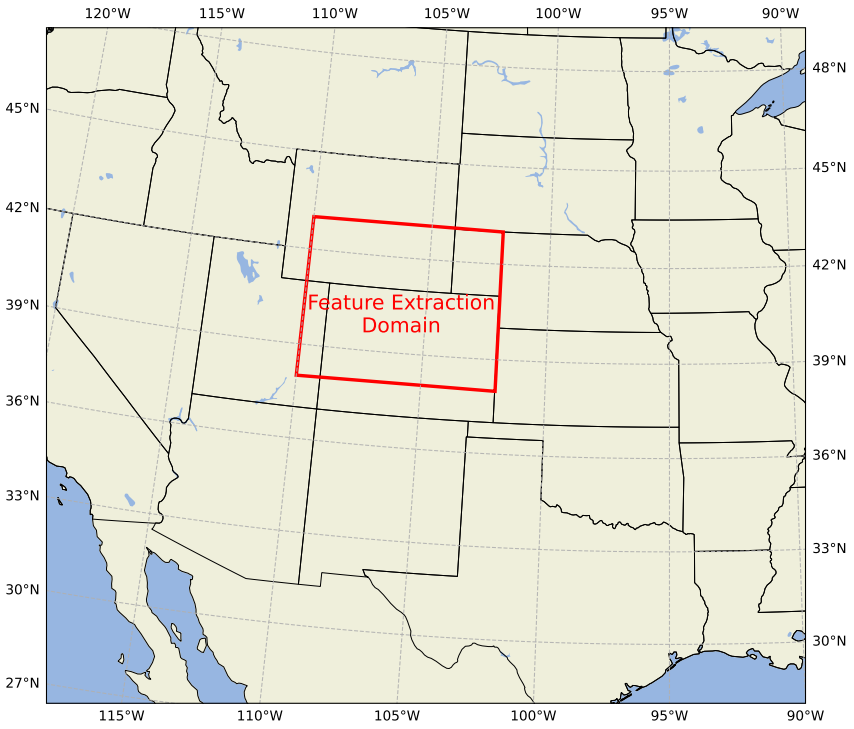

Figure 2.4: Domain from which input features are extract (red box) in the context of the larger 12-km CSU-WRF domain.

are the same as what one would use to calculate the point-based values used in the Lindsey et al. (2011) study. The only exception is the inclusion of the 300-hPa wind field as previous forecast experience indicates that understanding the strength and region of the jet stream over a location is helpful when gauging the strength of the winds able to mix down to the surface on a given day.

Although the choice of feature inputs is important, which data are excluded is also notable. Even though the models may differentiate between where points lie in the given domain, no latitude or longitude data are included. Additionally, the models do not ingest any details about the terrain so they cannot use elevation or the relative locations of mountains and valleys that human forecasters typically rely on when forecasting downslope windstorms. Finally, even though the distribution of these events throughout the year is discussed above, the models are not given the date. For example, the model does not know that in January a high wind event is more likely in Boulder, a heuristic it could potentially use when making a forecast. Including one-dimensional

Table 2.2: List of forecast atmospheric variables collected as input features to the machine learning models.

Input Features
300-hPa zonal (u), meridional (v), and vertical (w) wind
650-hPa potential temperature
Potential temperature difference between 700-hPa and the surface
700-hPa u, v, and w wind
700-hPa geopotential height
10-m u and v wind
2-m temperature
Mean sea level pressure

dates with other two-dimensional data presents training difficulties so the dates remain left out of the input features.

2.2.3 RF Models

The first goal of the study is to create forecasts for the three forecast locations with RFs. The Scikit-learn package in Python provides the ability to quickly train RFs with user-provided input features and hyperparameters (Pedregosa et al. 2011). Hyperparameters are settings the user can tune that determine how the RF behaves as it is trained. Essentially, any decision a user has to make before training the model can be considered a hyperparameter. A RF is a collection of decision trees that iteratively split data based on values of the predictors such that the input predictors associated with the samples are correctly matched to their labels. The RF learns to optimize these values to minimize error in the predictions. Probabilities are determined based on the number of decision trees within the forest that arrive at a particular prediction. For example, in an RF containing 10 decision trees where three predict a high wind event, the probability of a high wind event occurring is taken to be 30%.

Because of how RFs operate, the input features must be flattened before model ingest. Thus, the RF receives the 260,000 points contained in one sample in a one-dimensional array and treats them as independent predictors, although we recognize that nearby points in the array are highly correlated. As the RF receives each sample during training, a given decision tree splits the data

Table 2.3: Hyperparameters chosen for RFs at each location for each forecast day.

Model	Tree Depth	Minimum Samples Split	Minimum Samples Leaf	Number of Estimators
Cheyenne Day 1	25	5	20	200
Cheyenne Day 2	25	5	20	200
Fort Collins Day 1	10	4	8	5000
Fort Collins Day 2	5	12	2	2000
Boulder Day 1	10	10	6	2000
Boulder Day 2	25	18	6	100

based on a predictor value such that Gini impurity is minimized meaning the probability that a sample would be misclassified based on the chosen predictor threshold is decreased (Scikit-learn Developers 2025). Additionally, four other hyperparameters dictate how the RF builds decision trees during training. The hyperparameter “maximum depth" or “tree depth" tells the RF how many levels each decision tree can grow when splitting the data. The hyperparameters “minimum samples split" and “minimum samples leaf" control how a split, also called a leaf, is created. The former tells the RF how many samples must be within a leaf to create a split, and the latter specifies how many samples must exist on both sides of the resulting split to allow such a split. Lastly, the “number of estimators" sets the maximum number of decision trees that each RF may contain. Because each location and forecast day are separate RFs, each of the resulting six RFs have specifically tuned hyperparameters that differ from each other. The validation data are used to tune the hyperparameters in a way that maximizes the forecast metrics of each RF without overfitting the models to the training data, which would result in models that will not generalize well to new data. The hyperparameter values for each RF are shown in Table 2.3.

As discussed above, this problem presents a significant class imbalance between the number of non-events, moderate wind events, and high wind events. This impacts model training as the RFs will see far more non-event samples than samples of either intensity of wind event. Without considering this, the model solution would likely converge on the RF forecasting non-event every day. This is not helpful when we are attempting to improve the forecast of the extreme, rare events. Thus, we introduce different weights to the samples based on which event category they

belong. The balanced class weighting scheme within Scikit-learn is employed by assigning weights inversely proportional to the class frequencies of the input data. This is given by:

$$w_s = \frac{n_s}{n_c n_y} \tag{2.1}$$

where w_s is the weight assigned to the sample, n_s is the number of total samples in the input data, n_c is the number of classes, and n_y represents the number of samples belonging to a particular class (Scikit-learn Developers 2025). This ensures that moderate wind events are weighted more than non-events and high wind events are weighed even greater than the other two classes. Thus, the RF during training is punished more severely for miscategorizing high wind events pushing the solution towards forecasting these events correctly.

2.2.4 CNN Models

The second aim of the study is to create CNN models to forecast wind events using the same input and label data as the RFs. This allows comparison between the two ML methods. A CNN uses convolutions to process an image by changing the data within filters that move across an image (LeCun et al. 2015; Chase et al. 2023). In this way, the CNN extracts features from the images that it deems useful in making its final classification. The CNNs also contain pooling layers following the convolutional layers that extract the maximum or mean value within a 2x2 grid box and propagate that value to the next layer. This achieves dimensionality reduction within the model and allows it to hone in on higher resolution features within the image (Chase et al. 2023). We utilize Keras on top of Tensorflow to construct and train the CNNs in this study (Chollet et al. 2015; Abadi et al. 2022).

The input images for the CNNs are the output from the CSU-WRF. The CNNs ingest the same 50x80x65 cube of input data though the cube need not be flattened into one dimension as with the RFs. This leverages a strength of CNNs: image processing and recognition. Instead of using the CNN to tell us what the images contain, we are asking it to recognize wind events based on the output of a traditional numerical weather prediction model. In contrast with the

class imbalance mitigation method employed by the RFs, the training samples for the CNNs are bootstrapped and weighted equally. This is accomplished by randomly resampling the moderate and high wind event samples until their respective numbers matched the amount of non-event samples. This provides more stability during training as well as assisted in preventing the model from immediately overfitting to the non-event class.

As with the RFs previously, each of the six CNNs has specifically tuned hyperparameters to maximize its performance for its location on the given forecast day. However, the overall architecture of each CNN is the same. After the input layer, the data encounter the first “channel dropout" layer. This layer is similar to a standard dropout used in a dense, fully connected layer where a specified percentage of neurons are randomly turned off by assigning them zero weight (Géron 2019). This applies regularization to the model by preventing overfitting because some of the learned features are lost during dropout requiring the model to continue to look at more samples before converging on a solution. While normal dropout can be applied to a convolutional layer, this does not achieve the same strength of regularization because the neurons are highly spatially correlated in an image. So while the model loses a pixel’s value it still retains effectively the same information by looking at the neighboring pixels. Channel dropout applies a stronger form of regularization by randomly dropping out a specified percentage of entire channels of the previous layer. In our case, the input data flowing through first channel dropout layer yields that not all 65 maps output from the CSU-WRF reach the first convolutional layer. For example, the model may not have the 12-hour forecast of the 700-hPa u wind for a particular sample. However, the data missing changes randomly for each sample so the model still learns from the 12-hour forecast of the 700-hPa u wind on a holistic basis. The channel dropout helps extend the training process by preventing overfitting its solution to the features contained within a few samples.

After the first channel dropout layer, the models contain a standard convolutional layer. This layer scans a 3x3 kernel with a specified number of filters over each map in input data that passed through the channel dropout layer. The maximum pooling layer discussed previously then follows this convolutional layer. This three-layer pattern repeats twice more except that the final iteration

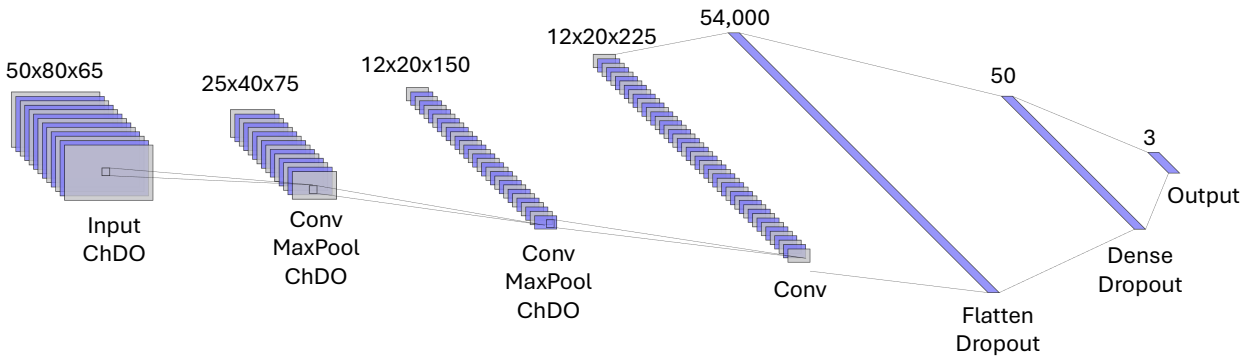

Figure 2.5: Diagram depicting the architecture of the Boulder Day 1 CNN. The other five CNNs follow the same layer architecture with small differences in hyperparameters such as number of filters and dense nodes. Above each layer are the output dimensions, and the channels shown in each layer are downscaled by a factor of five. Abbreviations: channel dropout (ChDO), convolution (Conv), maximum pooling (MaxPool).

does not contain a pooling layer. The number of filters applied during each convolution is a factor of the number of filters applied during the first convolution. This first number varies at each forecast location, and the second convolution utilizes double the original number of filters while the third convolution triples the amount of filters used in the first convolution. Thus, the two-dimensional domain of the features shrinks as it passes through the pooling layers, but the number of channels increases as the filters are applied. The convolutional layers utilize exponential linear unit (ELU) activation functions. Initial testing highlighted the “dying neuron" problem when the more common rectified linear unit (ReLU) activation function is used. This happens when the weighted sum of inputs to a neuron is negative resulting in that neuron’s output being set zero in accordance with the ReLU activation (Géron 2019). This means that the neuron stops learning going forward in the training process. ELU permits negative values to remain during training, allowing these neurons to continue learning.

Following the convolutional portion of the model, the features are flattened into one dimension before encountering fully connected layers. This allows for the transition from feature images to the final output consisting of three neurons representing the likelihood that a sample belongs to each of the three forecast classes. Specifically, after flattening, a standard dropout layer is applied before the data encounter an ELU-activated dense layer. The number of neurons in this dense layer is again a hyperparameter specifically tuned for each model, as is the percentage of neurons

switched off by the dropout layer. The data pass through another dropout layer before reaching the previously described output layer. Importantly, both the channel and standard dropout layers are turned off during model inference meaning the models utilize all the supplied data when making forecasts. Figure 2.5 depicts the architecture of the Boulder Day 1 CNN whose overall architecture matches the other CNNs as well.

The model forecasts the event classification associated with the highest value output neuron. The softmax activation function used in the output layer normalizes these values and ensures that they sum to one. These output values can be thought of as probabilities; however, there is nothing inherent within the model that guarantees that these probabilities are calibrated in a way that would be meaningful to the forecaster. Thus, these probabilities directly output from CNNs are not investigated further in this study.

2.3 Verification Metrics

2.3.1 Dichotomous Metrics

This study verifies the forecasts made by the RFs and CNNs with standard contingency table metrics described in Chapter 9 of Wilks (2019) throughout the test period, which are the wind seasons beginning in 2021 and 2022. Each forecast is verified in one of four ways: a hit (h), a miss (m), a false alarm (f), or a correct negative. A hit is defined as when an event (or non-event) occurs and the model correctly forecasts its occurrence. A miss is when an event (or non-event) occurs but the model does not forecast it or forecasts a different event severity. A false alarm occurs when the model forecasts an event (or non-event) to occur but that event (or non-event) is not observed. A correct negative constitutes an event (or non-event) not occurring and the model correctly forecasting its non-occurrence. We ignore correct negatives since we are mainly interested in the models' ability to forecast each event (or non-event). Each event (or-non event) category is verified in a binary sense meaning the two other categories containing the non-observance or non-forecasts of the category being verified are binned together. For example, if we are verifying an observed moderate wind event, the model forecast will be scored a hit if and only if the moderate category is forecast. The model receives a miss whether it forecasts a non-event or a high wind event on that

day. Thus, the metrics presented are focused on the event category being verified and are agnostic to the actual meaning of the other two categories.

The rate at which the model correctly forecasts a particular category is the probability of detection (POD) calculated by:

$$\text{POD} = \frac{h}{h+m}, \tag{2.2}$$

and similarly the false alarm rate (FAR) is the rate at which the model forecasts a particular category that is not then observed where:

$$\text{FAR} = \frac{f}{h+f} \tag{2.3}$$

(Wilks 2019). To assess the models hits, misses, and false alarms together in one metric, we use the critical success index (CSI) where:

$$\text{CSI} = \frac{h}{h+m+f}. \tag{2.4}$$

The CSI works well when verifying infrequent events such as high winds as it is not dominated by correct negatives (Taggart et al. 2022; Wilks 2019).

Because the RFs output probabilities along with a deterministic forecast, we assess the skill of these probabilistic forecasts. To do this, we apply the Brier Skill Score (BSS) that measures the skill of a forecast against a reference forecast, which in this study is the climatological occurrence of each wind category within the training and validation period. The BSS is first calculated from the Brier Score (BS):

$$\text{BS} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2, \tag{2.5}$$

where n is the number of forecasts being verified, y_k is the probabilistic forecast of the wind category occurring, and o_k is either one or zero based on whether the event occurred or not, respectively.

Subsequently, the formula for the BSS follows:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{ref}}, \tag{2.6}$$

where BS_{ref} is the reference BS calculated using the climatology as previously discussed. Thus, BSS values above zero represent forecast skill relative to a climatological forecast while negative BSS values indicate the forecast is not skillful (Wilks 2019).

2.3.2 Multicategorical Metric

In addition to the dichotomous metrics above, other verification methods address multiple categories simultaneously so that all possible combinations of the forecast and observed categories are taken into account during scoring. While other skill scores can be generalized to include more than two categories, they do not penalize forecasts that miss by multiple categories more heavily than forecasts that miss by only one category (Wilks 2019). The Gerrity skill score takes into account the distance the forecast is from the observed category in addition to adjusting the scoring weights to account for the sample climatology of the events (Wilks 2019). Two concerns arise from this approach: first, when forecasting rare events, the weights determined from the sample climatology are such that the score can be maximized by forecasting the rare events excessively, leading to the proliferation of false alarms. This happens because of the high reward for correctly forecasting rare events, and it overtakes the penalty for overforecasting the more common events at the lower thresholds. Second, false alarms and misses are penalized equally, although in reality these errors rarely carry the same impact in most forecast applications (Taggart et al. 2022). In order to weight forecast misses with respect to the distance of the observed events and to also appropriately balance punishing false alarms and misses, this study adopts the FIxed Risk Multicategory (FIRM) framework from Taggart et al. (2022).

The FIRM score creates a scoring matrix that takes into account both the difference in weight between the different forecast categories and that false alarms and misses should not be punished equally. In our case with three event categories, there is one weight associated with moderate (w_1),

and high (w_2) wind events. Additionally, the risk parameter, α, determines the cost of a miss and in turn that of a false alarm $(1-\alpha)$. For a 3x3 contigency table, this yields the following matrix representing the penalties based on the observed and forecast categories:

$$\begin{bmatrix} 0 & \alpha w_1 & \alpha(w_1+w_2) \\ (1-\alpha)w_1 & 0 & \alpha w_2 \\ (1-\alpha)(w_1+w_2) & (1-\alpha)w_2 & 0 \end{bmatrix}, \tag{2.7}$$

where the observed categories increase to the right across the columns and the forecast categories increase down the rows. As the zero diagonal represents correct forecasts, the penalties associated with misses are above the diagonal and, likewise, the false alarm penalties exist below the diagonal (Taggart et al. 2022).

Both the weights and the risk parameter are chosen based on the user's needs and the forecast application. We set the weights [w_1,w_2]=[1, 3] meaning that forecasting the highest category correctly is three times as important as correctly forecasting the moderate category. Because this study is not tied to a specific operational forecast directive, we set α=0.7, which is the value used in the first example given in Taggart et al. (2022). This means a miss cost 2.33 times as much as a false alarm and a forecaster should forecast the highest category for which the probability of that category verifying exceeds 30% $(1-\alpha)$. This allows for higher weighting the more extreme, rarer events since their real-world impacts are much greater without over-rewarding these forecast through penalizing false alarms appropriately with respect to the misses. But because false alarms are penalized less than misses, there is still incentive to forecast the higher categories as long as the forecast probabilities warrant it. With these values, this study applies the following matrix of penalties derived from the matrix in Equation 2.7 when utilizing the FIRM framework:

$$\begin{bmatrix} 0 & 0.7 & 2.8 \\ 0.3 & 0 & 2.1 \\ 1.2 & 0.9 & 0 \end{bmatrix}. \tag{2.8}$$

The FIRM scores are calculated for each day with the scores Python package that includes the FIRM calculation by accepting forecast and observed xarray DataArrays (Leeuwenburg et al. 2024). The mean of these scores is then calculated and presented below. While the weights and risk parameter chosen above seem arbitrary, they do allow us to assess each model at each location under a consistent scoring framework. Changing these values does indeed affect the FIRM scores, and subsequent sections will address how changing the risk parameter affects how the models performances are perceived relative to each other.

2.4 Results

2.4.1 Dichotomous Metrics

We first present the metrics traditionally associated with the two-class contingency table, including POD, FAR, and CSI as described above. We also display the miss rate for ease of comparison, though it does not provide additional information beyond the POD since the miss rate is simply $1 - \text{POD}$. Recall that each class's metrics treat the other two possible classes as the same class (non-event relative to the event class being scored), thus producing the two possible classes. Probabilistic metrics for the RFs are shown following the deterministic forecast metrics.

Deterministic Forecast Metrics

The following figures compare dichotomous metrics for the RFs and CNNs for both Day 1 and 2. Additionally, the direct output from the CSU-WRF is scored and presented. In order to score the CSU-WRF, the 10-m wind is extracted from the model grid point nearest each location for every three-hourly forecast hour, within the 06-06 UTC verification window on each forecast day for all ensemble members. We consider the member with the highest 10-m wind value occurring during each forecast day's verification window as the CSU-WRF forecast for that day. This value is categorized according to the sustained wind event criteria in Table 2.1 and scored similarly to the forecasts from the ML models. The 10-m wind output from the CSU-WRF is not a true sustained wind forecast as it represents the instantaneous 10-m wind speed at the time the forecast hour output is written. Thus, it likely overestimates the true sustained wind during the forecast hours.

However, no gust output is available from the CSU-WRF, and comparing the 10-m wind to the gust criteria is unfair given the large gust factors often observed during downslope windstorms. This study aims to give the CSU-WRF the best opportunity to forecast high winds. Therefore, this verification methodology likely yields more optimistic CSU-WRF performance metrics than how the model performs in reality.

Figure 2.6 displays the contingency table metrics for high wind forecasts by the RFs, CNNs, and the CSU-WRF for all locations on both forecast days. When comparing the two types of ML models, the differences in the POD and FAR standout between the RFs and the CNNs. This is less evident in Boulder, and especially pronounced in Fort Collins. The CNNs are able to detect more high wind events in general, but this comes with a higher FAR. Thus, if detection alone is particular important, the CNNs offer a better forecast. However, at some point FAR becomes problematic especially as some models produce FARs at or above 70%. For example, in Fort Collins on Day 1, the CNN achieve a POD of 100% for this two-year verification period. It would be tempting to put high confidence when presented with a high wind forecast from this model, however, with a 82.5% FAR, the vast majority of these high wind forecasts do not verify. That implies great uncertainty to a forecaster though no events are missed during this verification period. This uncertainty is reflected within the relatively low 0.175 CSI value for the Day 1 Fort Collins CNN. Ultimately, the POD-FAR imbalances observed between the RFs and the CNNs are taken into account in the CSIs. We see the RFs tend to edge the CNNs on Day 1 with the CNNs beating the RFs on Day 2. Boulder is the exception to this where the RFs maintain their CSI advantage in Day 2.

Fort Collins itself is an outlier due to the extremely small sample size of high wind events in comparison to the other two locations. All three model types struggle, but the RF metrics are particularly poor due to the fact that they do not produce a high wind deterministic forecast. This means the decision trees “voting" for high winds within the RF never outnumber either of the other two classes. The CNNs achieve some success at detecting high wind events, but as previously discussed, this comes at a significant inflation of the FAR.

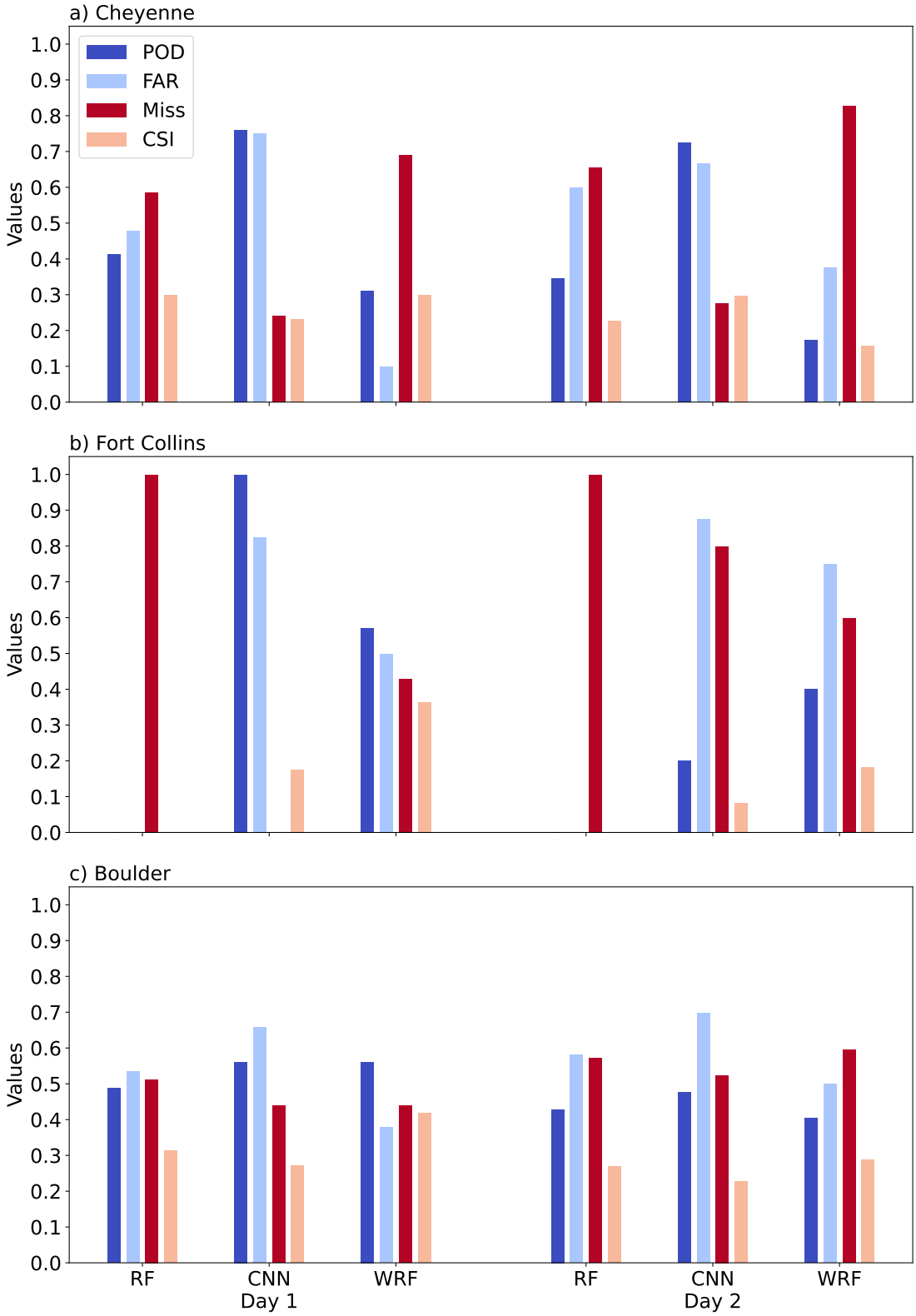

Figure 2.6: High wind event contingency table metrics for the Day 1 and Day 2 RFs, CNNs, and direct CSU-WRF output for (a) Cheyenne, (b) Fort Collins, and (c) Boulder. POD (dark blue), FAR (light blue), miss rate (dark red), and CSI (light red) and displayed.

On Day 1, the CSU-WRF performs on par or better than the ML models. In fact, the CSU-WRF is the better model in Fort Collins and Boulder scoring a higher CSI mainly through a lower FAR compared to the ML models. While the FAR remains quite low in Cheyenne, the miss rate is much higher especially compared to the CNN. Despite the CSU-WRF having a CSI near the RF and better than the CNN, the higher miss rate should be considered.

Day 2 is where the utility in the ML models becomes apparent. Focusing on Day 2, the CSU-WRF performs worse across the board compared to the ML models except at Fort Collins. Specifically, the detection decreases significantly, and this is accompanied by an increase in FAR as well. Essentially, the CSU-WRF continues to forecast high winds on Day 2, but on incorrect days with very few correct days “redeeming" the false alarm forecasts. Comparing each model’s performance between Day 1 and 2, we note the CSU-WRF’s metrics worsen in particular. Setting Fort Collins aside for the previously mentioned sample size concerns, the RF and CNN Day 2 forecasts perform closer to their respective Day 1 forecasts. This implies the ML models are increasing the predictability of these high wind events by increasing the lead time their forecasts making them still useful to a forecaster. The CSU-WRF forecast performances decrease dramatically in Day 2 to where they provide less value to a forecaster.

Next, the contingency table metrics for moderate wind events are presented in Figure 2.7. Not unexpectedly, the ML models and CSU-WRF perform better on forecasting the moderate wind class likely due to increase in moderate wind events compared to the high wind events despite the class imbalance measures taken when training the models. Once again, we note a larger decrease in CSI from Day 1 to Day 2 for the CSU-WRF than seen in the ML models implying the ML models provide value at longer lead times. However, the RF performances exceeds that of the CNN or CSU-WRF with the exception of Day 1 Boulder where the CSIs are similar across all three models. The RFs achieve this with higher PODs and lower FARs meaning it has more successfully optimized detection without sacrificing false alarms compared to the CNNs and CSU-WRF. While the RFs struggled with detection compared to the CNNs for high wind events, the opposite appears to be true for the moderate category. A forecaster would benefit from utilizing the RF deterministic

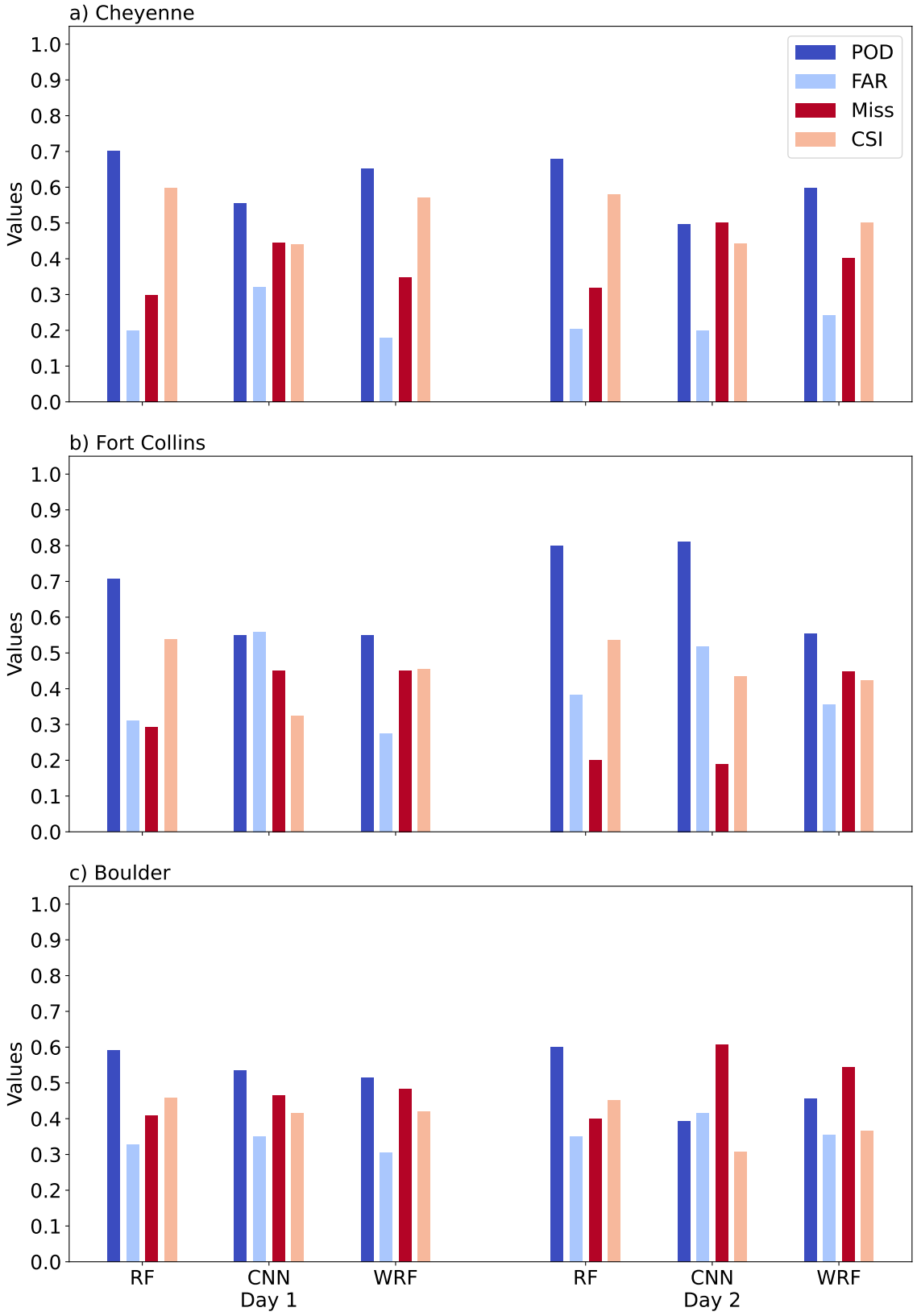

Figure 2.7: Moderate wind event contingency table metrics same as in Figure 2.6.

forecasts at any location for both forecast days. Despite the RF struggles forecasting high winds, RFs also output probabilistic forecasts that provide information to the forecaster. These metrics are presented next subsection.

Probabilistic Forecast Metrics

As the RFs output probabilities for each forecast class based on the number of decision trees within the RF forecasting each wind event category, a forecaster can consider this probabilistic information when making a forecast in addition to the deterministic forecast. In order to determine whether these probabilistic forecasters are useful to the forecaster, we need to ensure they are calibrated. As CNNs do not produce calibrated probabilities and the CSU-WRF does not output ensemble probabilities, only probabilistic forecasts from the RFs are considered in this subsection.

Figure 2.8 shows the RF calibration curves for all three wind classes at each location for both forecast days. Each curve shows how often a bin of forecast probabilities actually verifies (fraction of positives). This figure is created by binning the forecast probabilities into ten bins. Ideally, a 40% forecast of high winds should verify 40% of the time. The 1:1 forecast probabilities to fraction of positives shown in Figure 2.8 by the black dashed line represents a perfectly calibrated probabilistic forecast system.

For moderate wind probabilities, all six RFs tend to overforecast the lower probabilities and then underforecast the higher probabilities, though, the range of probabilities where this transition occurs varies by location and forecast day. Overall, the probabilities remain relatively close to the 1:1 line, therefore, we conclude the moderate wind probabilities are reliable especially if a forecaster keeps in mind the over-to-under forecast tendency transition as the probabilities increase.

High wind probabilities in Cheyenne follow the same trend noted in the moderate wind with more variability in the position of the bins about the 1:1 line. Additionally, above the 70% probability threshold, the probabilities are particularly underforecasted on both forecast days implying a forecaster should put more confidence in a high wind event than even these higher probabilities indicate. Looking at the high wind probabilities in Boulder, we note a consistent overforecasting tendency, but the curve still increases similarly as the 1:1 line. Subtracting 5-10% from the

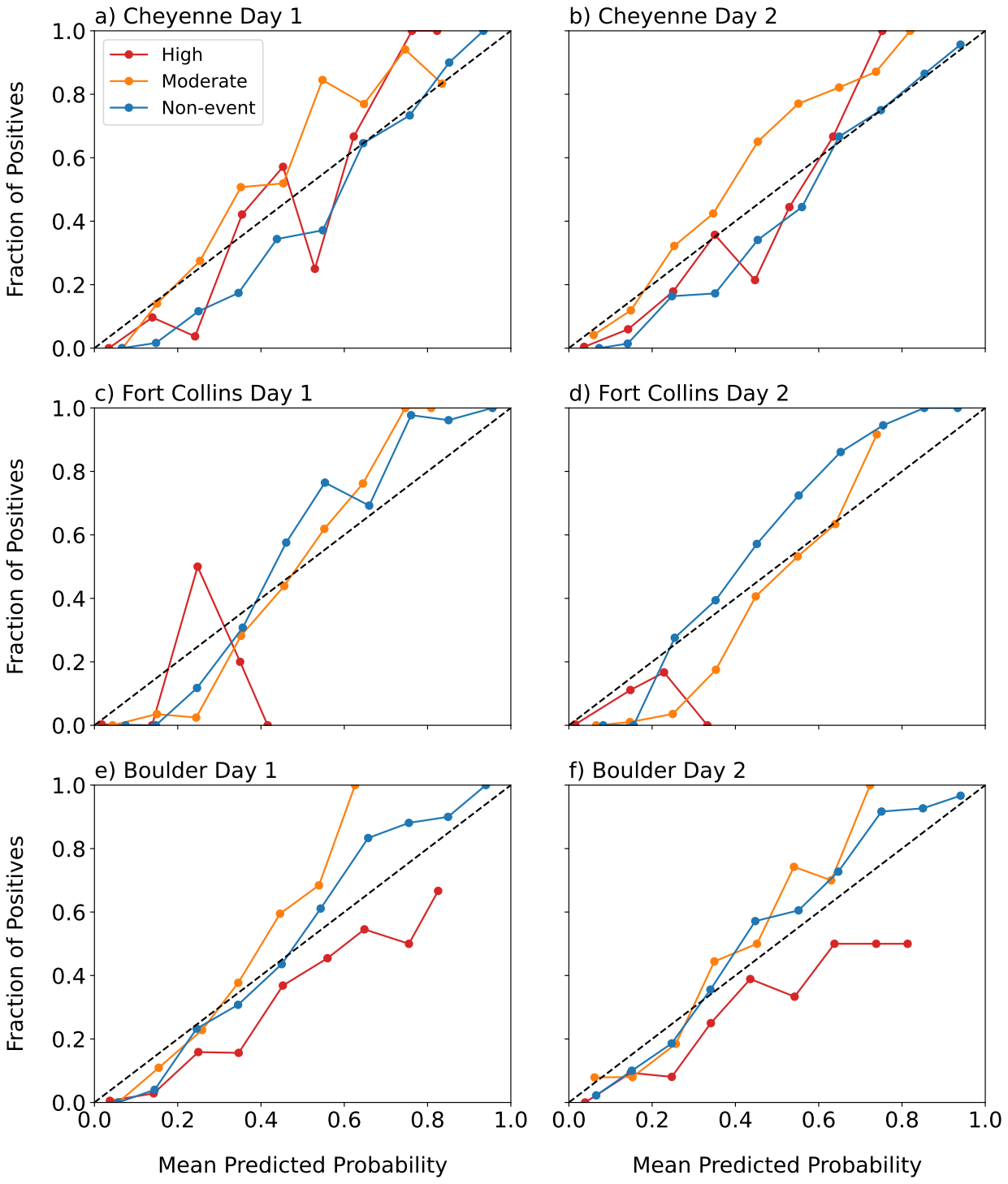

Figure 2.8: Calibration curves for the (a-b) Cheyenne, (c-d) Fort Collins, and (e-f) Boulder Day 1 and 2 RFs. Red, amber, and blue curves represent high wind events, moderate wind events, and non-events, respectively. The black dashed line shows the 1:1 or perfectly calibrated forecast line.

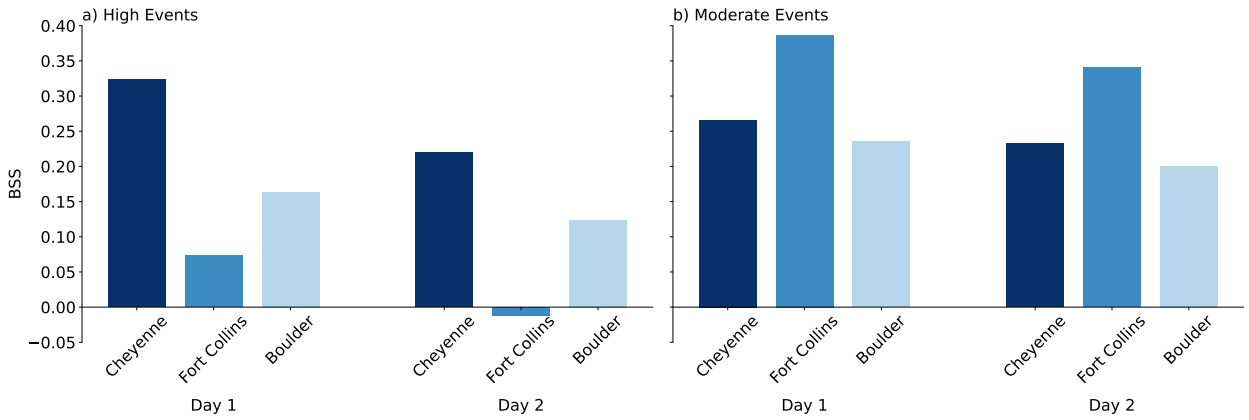

Figure 2.9: BSS for RFs at all three locations on both forecast days for (a) high wind events and (b) moderate wind events. Positive values indicate forecast skill relative to a climatological forecast while negative values indicate the forecast has no skill.

output probability would actually produce a well-calibrated forecast. The main disadvantage is the Boulder RFs did not forecast high wind probabilities above 86% and 81% for Day 1 and Day 2, respectively, during the two-year test period so no reliability information can be gleaned above those thresholds.

Despite the probabilistic forecasts discussed so far showing reasonable calibration, the high wind probabilities in Fort Collins are not reliable. As with the deterministic forecasts, the issue is small sample size so the RFs do not see enough unique training cases to produce a reliable forecast. In fact, the Fort Collins RFs do not forecast high wind probabilities above 43% and 38% for Day 1 and Day 2, respectively (these lower probabilities are why the high wind classification is never the plurality resulting in no high wind deterministic forecasts). However, the Day 2 probabilities below 20% are well-calibrated, thus, a forecaster might use this as an indicator to consider the possibility of high winds at the Day 2 lead time. In practice, a forecaster should entertain the possibility of high winds at any probability above 20% as well, just understanding the forecast probability does not correlate to a true probability of the event occurring, and that there is still no guarantee an event will occur even at the highest probabilities the RFs actually produce.

Having established confidence in the reliability of the RF probabilistic forecasts given the caveats mentioned, the next step entails grading the performance of these forecasts. As discussed in the previous section, the BSS compares the probabilistic forecasts to the baseline climatology

of each event occurring at each location. The BSS values for the probabilistic forecasts for all six RFs are shown in Figure 2.9.

For the high wind probabilities, we note positive forecast skill at each location on both forecast days with the exception of Fort Collins on Day 2. This means the probability forecasts are skillful when compared to a climatological forecasts in addition to being calibrated. Although the lower high wind probabilities on Day 2 in Fort Collins are calibrated, the negative BSS indicates those forecasts may be no better than forecasting the probability of high winds occurring based on the climatology of the dataset. Both Cheyenne RFs show particularly strong forecast skill when compared to the other locations. Above in Figure 2.6, the CSIs for the deterministic RF forecasts for Cheyenne and Boulder competed with each other, but now we see that the probabilistic forecasts in Cheyenne are superior.

Moving onto the moderate wind probabilities, all six RFs produce skillful forecasts with no exceptions. Figure 2.7 previously showed that these RFs performed closely when comparing the CSIs. However, the higher BSSs on both days in Fort Collins mean these probabilistic forecasts may outperform the forecasts at the other two locations. While the Fort Collins RFs struggled with forecasting high wind probabilities, the models forecast moderate wind probabilities much more successfully likely due to the higher number of unique training cases available.

Up until this point, these dichotomous metrics treat the two classes not being verified equally as “non-events". In reality, missing a high wind event when a moderate event is forecast does not have the same impact to the end-user as missing a high wind event when no wind event is forecast at all. Therefore, a verification system that considers all potential classifications simultaneously with different impacts provides benefits when assessing these models in a more realistic framework. The next subsection presents these results.

2.4.2 Multicategorical Metrics

Confusion Matrices

Before presenting the FIRM metrics introduced in the Data and Methods section above, confusion matrices provide conclusions by analyzing the hit, miss, false alarm, and correct negative

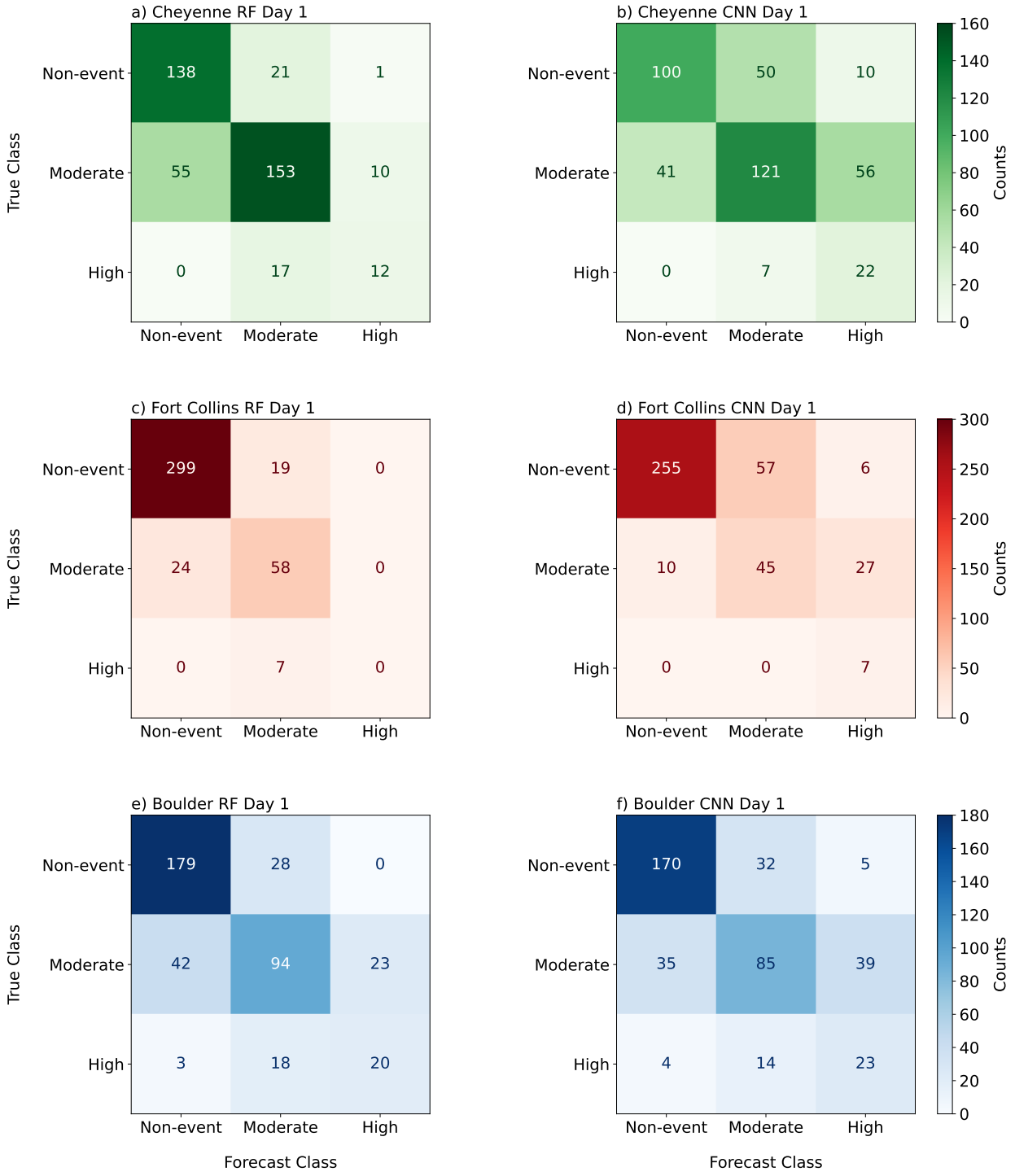

Figure 2.10: Confusion matrices for the Day 1 RFs and CNNs for Cheyenne (green), Fort Collins (red), and Boulder (blue). Darker shading indicates a larger proportion of the forecasts belonging to that square. Note the color scale changes for each location.

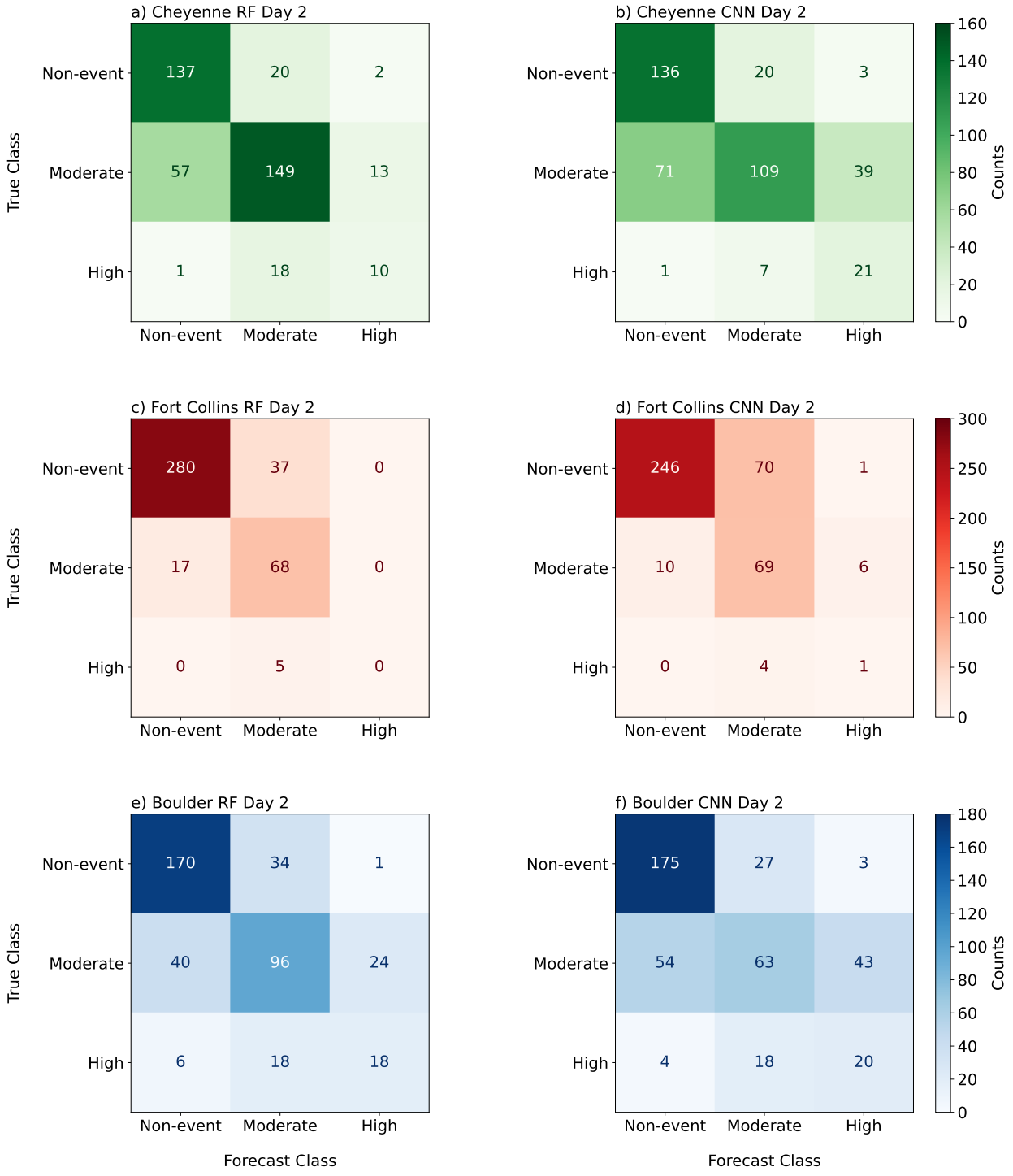

Figure 2.11: Confusion matrices as in Figure 2.10 for the Day 2 ML models.

forecast counts relatively to the three categories. Indeed, this is how the dichotomous metrics are calculated by essentially collapsing the 3x3 confusion matrix into a 2x2 contingency table, but by looking at the 3x3 confusion matrices independent of the dichotomous metrics above, we can assess how close the model is to making correct forecasts in all three categories.

The confusion matrices for both the RFs and CNNs at each location are shown in Figures 2.10 (Day 1) and 2.11 (Day 2). These matrices display the number of days belonging to each box based on what the ML model forecast for that day (the columns) and which category verified on that day (the rows). A perfect model would forecast such that all of the cases exist along the diagonal where the forecast class and the true class are the same. For our imperfect models, however, studying where the cases end up in the matrices provides insight on the models' performances without calculating any metrics. Cases above the diagonal represent false alarms while cases below the diagonal are missed forecasts.

After looking at these confusion matrices, we still draw many of the same conclusions as we did when analyzing the dichotomous contingency table metrics. The CNNs tend to have more false alarm cases, but also tend to detect more high wind events than the RFs. For the moderate class, the RFs perform more reliably. However, in an operational forecast setting, these squares are not all equal in end-user impact. A high wind false alarm may not be as concerning when moderate winds verify versus when no wind event occurs. If the latter case happens consistently, that indicates the model is performing poorly and might not be of much use to a forecaster. The confusion matrix allows for these distinctions to be made.

A model forecasting incorrectly by two categories is a strong marker of poor performance. Beyond optimizing specific metrics, this study aims to minimize the model forecasting high winds when no wind event verifies (top-right corner of the matrix) or it forecasting no wind event when high winds verify (bottom-left corner of the matrix). Missing the high wind event entirely is certainly more impactful, but the two-category false alarm also warrants concern as it implies the model likely has not learned the high wind event signature in the predictors. For the models in this study, we see these two-category misses and false alarms are quite rare. The Day 1 CNN in

Cheyenne does have a larger number of two category false alarms compared to any other model, but these still only account for 15% of the high wind false alarms and the CNN overforecasting tendency has been previously noted. Therefore, we are confident that these models learn from the predictors due to the rarity of these two-category incorrect forecasts.

FIRM Scores

Next, the FIRM score extends the confusion matrix analysis by assigning penalties to each off-diagonal box based on the weight of each event class and the risk threshold of the end user as described in Section 2.2. This allows quantification of the models' performances with respect to all three categories simultaneously. The mean FIRM scores for the ML models and the CSU-WRF for all locations and forecast days are depicted in Figure 2.12. Because the FIRM calculation assesses penalties for each forecast based on the verification matrix in the previous section, the score itself can be broken down into overforecast penalties for false alarms and underforecast penalties for misses. Figure 2.12 shows the breakdown of these penalties with blue representing the component of the score due to underforecasting and the red due to overforecasting. Thus, a lower overall FIRM scores indicator better forecasts.

Comparing each type model, similar conclusions stand out as seen with the dichotomous metrics. The CNNs are prone to false alarms more than the other two models. At each location, the RFs tend to perform better on both days relative to the other two models at the same location. This is likely tied to their better performance on moderate wind events, which might outweigh their struggles in detecting high wind events. The models' FIRM scores do not increase dramatically into Day 2. This includes the CSU-WRF's performance that contrasts with the sharp decline in the dichotomous metrics on Day 2. Lastly, the RFs and the CSU-WRF have a larger proportion of their FIRM scores derived from underforecasting. So while these models' FIRM scores are lower than the CNNs FIRM scores with some exceptions, the FIRM metric is rewarding the RFs and CSU-WRF for minimizing false alarms. If detection is important to the forecaster, this is an important distinction to bear in mind.

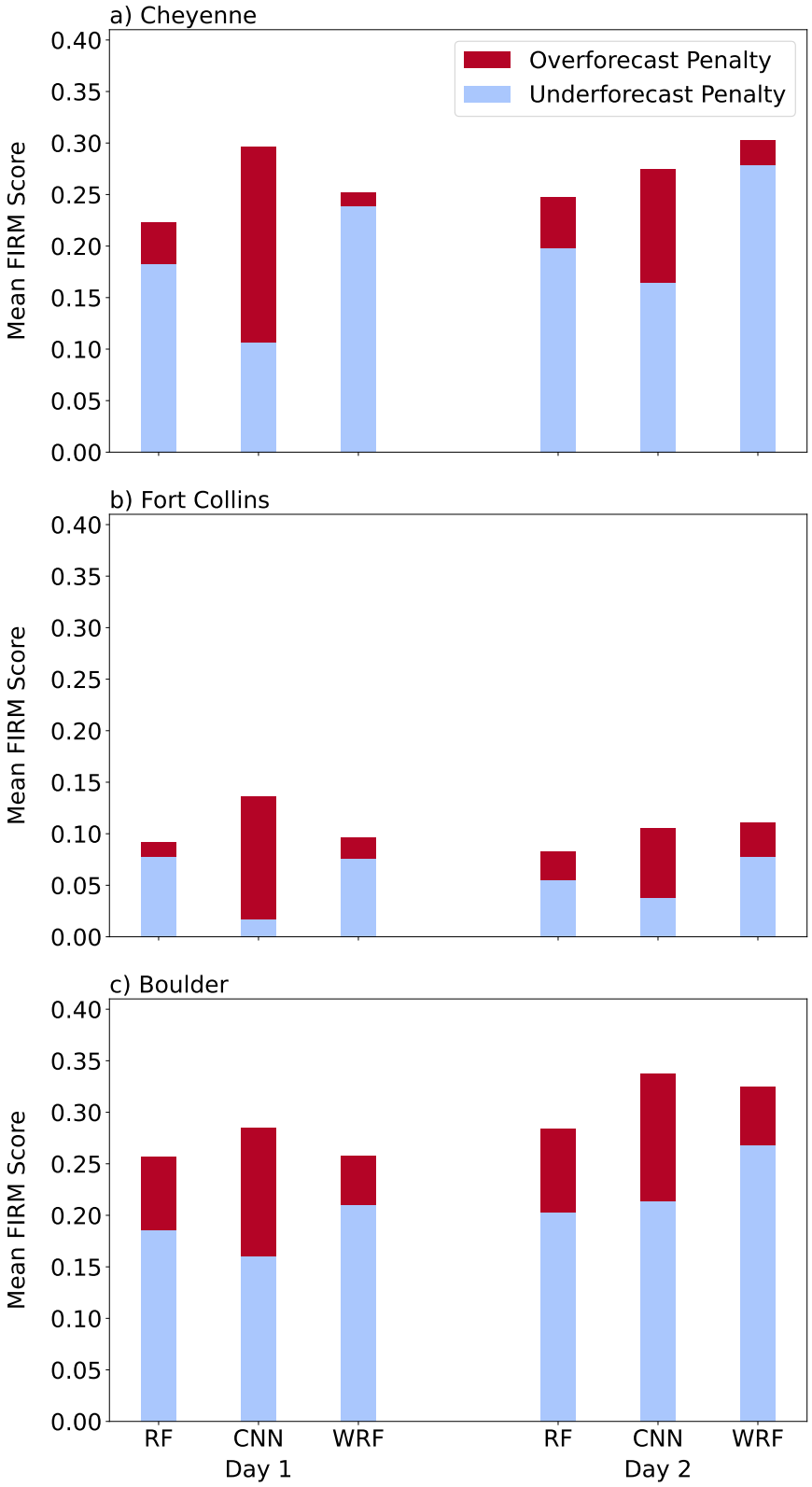

Figure 2.12: FIRM scores at (a) Cheyenne, (b) Fort Collins, and (c) Cheyenne for the Day 1 and Day 2 RFs, CNNs, and CSU-WRF. The blue and red bars represent the contribution of underforecast and overforecast penalties, respectively, to the FIRM score. Lower FIRM scores indicate better forecast performance.

The advantage of the FIRM score is that we can validate these model comparisons as these scores take into account all three event classes. The CNNs' high wind false alarm tendency, for example, is not redeemed by their better performance on moderate wind events, at least under the scoring framework chosen. Also, for these weights and risk threshold, misses are penalized more severely than false alarms so the blue and red bars in Figure 2.12 do not correlate 1:1 to the number of events. This helps remove subjectivity when comparing dichotomous metrics or looking at cases numbers in a confusion matrix and deciding how many false alarms is too many at the expense of increasing misses.

Initially, this study aimed to use the FIRM score as a method to compare the models across locations, perhaps even as a way to definitively assess whether one model at one location outperformed the rest everywhere else. This would lead us to conclude that all of the models perform better in Fort Collins even though previously it seemed as though Fort Collins was the worst location. However, once again, the small sample size must be taken into account. There are fewer moderate and high wind days in Fort Collins compared to the other two locations. When calculating the mean of the daily FIRM scores, Fort Collins has more days where the models correctly forecast no winds, which they excel at as evidenced by the confusion matrices in Figures 2.10 and 2.11. A correct forecast is scored the same (assessed zero penalty) no matter the category that verifies, and Fort Collins has more zero-penalty days that are easier to forecast that lower the mean FIRM score. Therefore, this study cautions against comparing FIRM scores across locations with different event climatologies. Similarly, it appears the models perform slightly worse overall in Boulder, but this could be due to the larger number of wind events of both intensities.

In order to combat the issue of correct negative forecasts artificially minimizing a model's FIRM score, Figure 2.13 shows the mean FIRM scores only including days where high winds verified or the model forecast high winds. This focuses the analysis on only the highest impact correct forecasts, misses, and false alarms. While this is similar to the approach used for the dichotomous metrics, the models verifying incorrectly by one or two categories is still penalized accordingly.

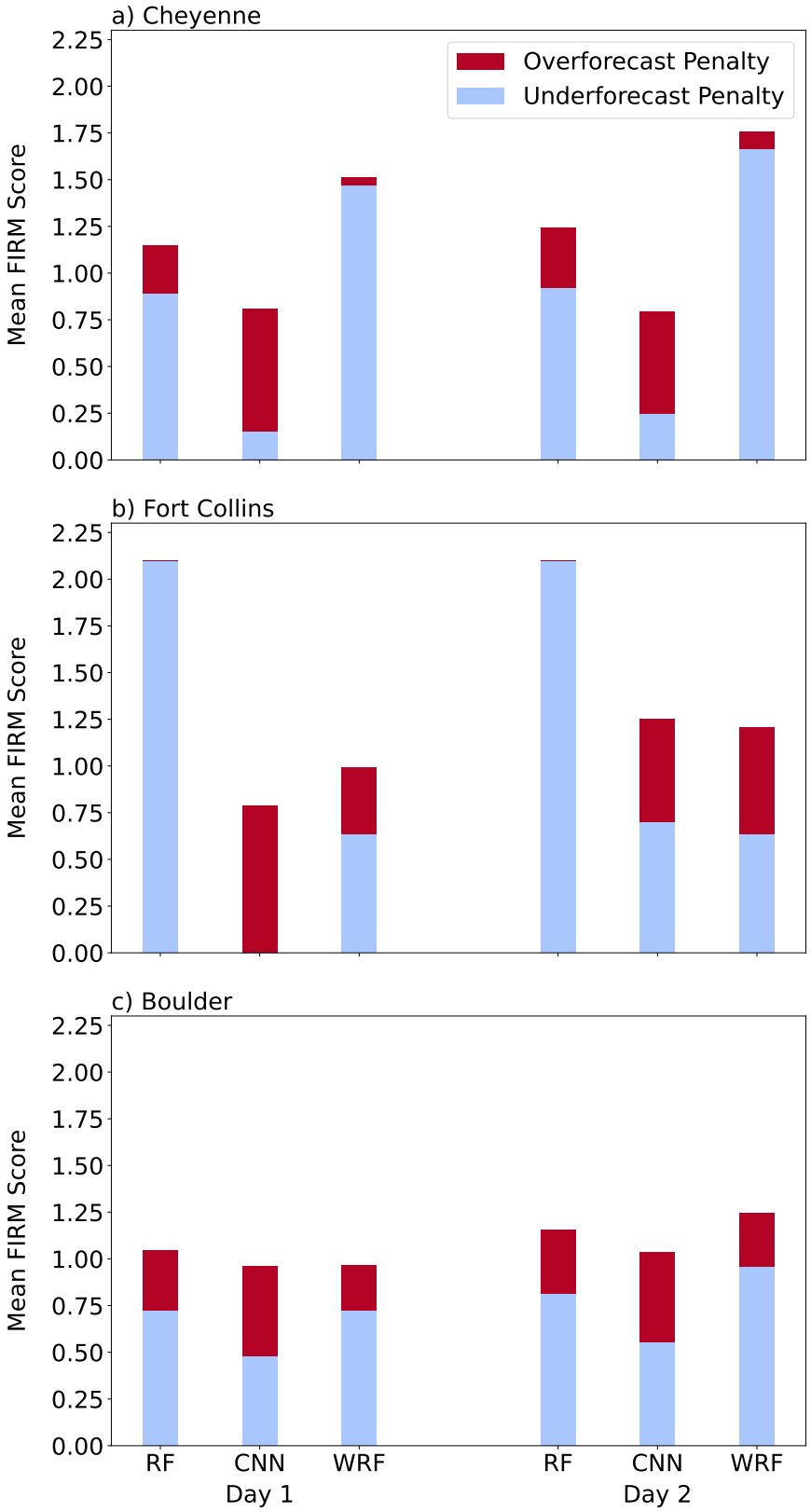

Figure 2.13: Same as in Figure 2.12 except only days where high winds verified or the model forecast high winds are included in the mean calculation. Note the range change on the Mean FIRM Score axis.

Unsurprisingly, the models perform worse on these high wind days and forecasts. The CNNs still overforecast worse relative to the other two models, though, the CSU-WRF actually receives a higher overforecast penalty in Fort Collins on these days than it did when looking at all of the days. However, the CNNs detection capabilities are rewarded at this threshold as they now outperform the RFs at every location on both forecast days. The CNNs also have lower FIRM scores than the CSU-WRF in all cases except for Fort Collins on Day 2. Also, by removing the correct negatives moderate wind events and forecasts, the models perform significantly worse at Fort Collins as expected, which aligns with the conclusions derived from the dichotomous metrics. The increased sample size of high wind events in Boulder and Cheyenne benefits the RFs the most as evidenced by the significant reductions in mean FIRM score at those two locations compared to Fort Collins.

FIRM Score Sensitivity to the Risk Parameter (α)

As alluded in the previous section, the weights assigned to each event and the risk threshold chosen impact the FIRM score calculation. Despite the acronym, Taggart et al. (2022) describe the FIRM framework as flexible in its application. This is useful when a user has a clear forecast directive from which to derive a risk threshold or a long record of metrics from a previous forecast system to derive the risk threshold. Choosing these parameters is more difficult in this study that is not tied to a specific operational user, which is why values given in the Taggart et al. (2022) study are selected.

To see the impact of varying the risk threshold, α, on the FIRM scores, we simply recalculate the FIRM scores for different thresholds. Figure 2.14 presents these results for the RFs (solid lines) and the CNNs (dashed lines) on Day 1 (blue) and Day 2 (red) for all three locations. The range of risk thresholds tested varies from 0.4 to 0.95 likely well beyond any practical use at the extremes. The low end of the range actually penalizes false alarms more than misses meaning the use case tolerates a high number of misses as long as false alarms are minimized. At the other end of the α range, misses are weighted more heavily, though when $\alpha = 0.95$, the use case approaches a point where any number of false alarms is acceptable as long as events are not missed. This jeopardizes the credibility of the model with the end user. As with comparing standard contingency table

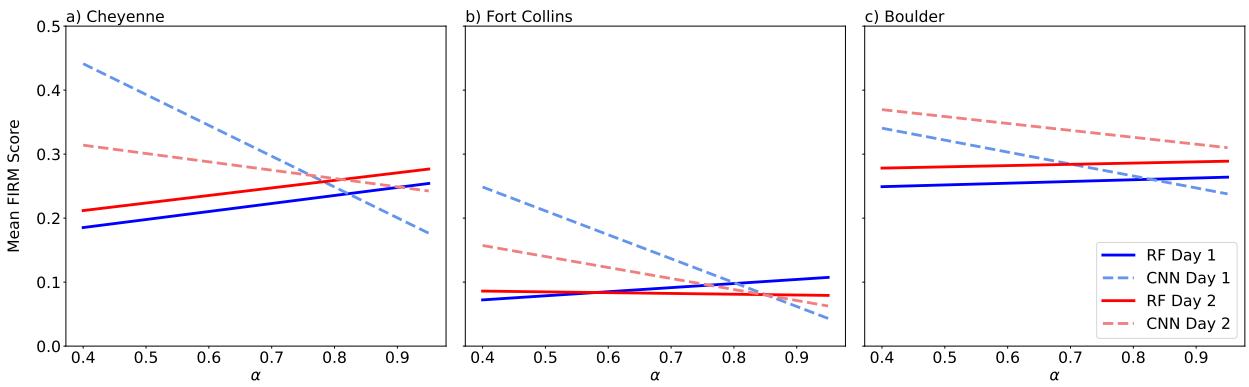

Figure 2.14: Mean FIRM scores with varying risk thresholds, α, at (a) Cheyenne, (b) Fort Collins, and (c) Boulder for the RFs and CNNs on both forecast days. The blue and red correspond to Day 1 and Day 2, respectively. The RFs' FIRM scores are shown with solid lines and the CNNs' scores with dashed lines.

metrics, difficulty exists when choosing the threshold that accurately quantifies the impact of a miss versus a false alarm.

The results presented above used a risk threshold of 0.7, and Figure 2.14 again shows the RFs tend to have lower (better) FIRM scores at this threshold. However, by looking at each forecast day, between 0.8 and 0.85 the RF and CNN FIRM scores intersect with the exception of Boulder on Day 2 where the RF outperforms the CNN for all α's tested. Above this intersection at these higher risk thresholds, a forecaster should use the CNN forecast. The CNNs better detection capabilities outweigh their overforecasting weakness at these higher thresholds. Therefore, it is important to determine the risk threshold associated with the forecast application.

While the FIRM score allows a forecaster to assess model performance without neglecting the difference in impact between false alarms, misses, and the severity of the event categories, the results only exist in the context of the weights assigned to each event class and the risk threshold chosen. Changing any of these will change the results perhaps even in a way that may lead the end user to determine a different model performs better. This highlights the importance of choosing these parameters carefully within the context of the forecast problem. The next section further discusses these parameters and their impact on the FIRM score.

2.5 Discussion

Any user of these ML models must keep a few things in mind or risk extending the application of the models inappropriately. First, these ML models are trained independently and the event climatology varies significantly between each location. Thus, inter-model comparison between locations should not have much bearing on assessing forecast confidence, but rather understanding each model's performance metrics on their own merits when making forecasts. For example, a forecaster should not necessarily trust a high wind forecast from a Cheyenne RF more than a forecast from a Boulder RF solely because the Cheyenne RFs scored higher BSSs than the Boulder RFs. A forecaster should, however, put less confidence in a probabilistic high wind forecast from the Day 2 Fort Collins RF on the basis that this model has negative forecast skill. The comparison between models speaks more to the challenge of the forecast problem as a whole and to the efficacy of the study's methodology to solve the problem. In this case, due to the infrequency of high wind events in Fort Collins, the ML models struggle more compared to the ML models forecasting for Cheyenne and Boulder. As a result, we surmise that the high wind forecast problem is more difficult in Fort Collins. The possibility exists that the methodology for developing both the RFs and CNNs is not adequate for the high wind forecast problem in Fort Collins and that a better method of training these models or training models with different architecture exists. Despite these varying event climatologies, the ML models show promising in extending the forecast lead time of high wind events. At each location, the ML models achieved smaller reductions in forecast performance on Day 2 than the CSU-WRF. As expected, the CSU-WRF performs worse on Day 2 in line with the challenge of forecasting these events using traditional numerical weather prediction. Interestingly, these same worsening forecasts are training the ML models meaning the ML models are learning more from worse training data on Day 2 than on Day 1. Because they are trained independently, the Day 2 models outperform the Day 1 models though the head-to-head metrics do not indicate this. These ML training methods are able to bridge the larger gap in the Day 2 training data compared to Day 1. This may mean there are some elements in the 30 to 54-hour CSU-WRF that are correct though not explicitly forecast in the numerical output. At least, there are patterns

in these forecast hours that the ML models learn to skillfully forecast these wind events at longer lead times. Whether or not these patterns are meaningful to a forecaster and whether they should be trusted is the subject of the next chapter.

Next, we see that in both the dichotomous and multicategorical metrics that the increase in event sample size benefits for the RFs greater than the CNNs. This is especially evident when comparing Fort Collins high wind mean FIRM scores to the other two locations' FIRM scores. At least for the ML architectures used in this study, it appears that CNNs are better suited for situations with smaller sample sizes due to imbalance in the event climatology or limited training period. CNNs also have better event detection so in applications where misses are highly detrimental, CNNs may be more appropriate. In this study, the deterministic RF forecasts are not useful for high wind events leaving only the CNN forecast. However, the RFs do provide readily available probabilistic forecasts so its important to weigh the tendency of the model architectures' performances and the information they provide when designing a new model for a new location or forecast problem.

Next, careful consideration of the risk parameter, α, is necessary when applying or interpreting FIRM score verification results. To minimize the FIRM score, a forecaster issues a forecast for the highest threshold containing at least a portion of the forecast probability distribution equal to the risk threshold (Taggart et al. 2022). So in this study when $\alpha = 0.3$, if a calibrated forecast probability greater than 30% exists for high winds, the forecaster should forecast high winds regardless of the probabilities in the other two categories. Over time, this would maximize the performance for this forecaster. The Taggart et al. (2022) study notes this provides a consistent scoring framework that cannot be hedged as easily as other metrics. This is true as long as the forecast directive is well-known and specific. In research studies like this one where the forecast problem is more ambiguous, the FIRM scores can be manipulated to make a model's performance appear better than another model. We include the results shown in Figure 2.14 varying the risk parameter not only to highlight the importance of selecting the parameters associated with the FIRM framework, but

Table 2.4: High wind event counts within the two-year test period for each location and the number of each type of wind criteria observed verifying.

Location	Event Count	Sustained Only	Gust Only	Sustained and Gust
Cheyenne	29	0	16	13
Fort Collins	7	2	0	5
Boulder	41	14	1	26

also to show choosing which models outperform others using the FIRM framework is necessarily more straightforward then relying solely on dichotomous metrics.

One final analysis elucidates how the 10-m wind chosen as a proxy for the CSU-WRF wind forecast may have impacted the Day 2 CSU-WRF metrics. As described above, the maximum 10-m wind is extracted from the CSU-WRF ensemble members from the gridpoint nearest to each location and verified against the sustained wind criteria. At both thresholds of wind events, either the sustained wind criteria, the gust criteria, or both criteria can verify for an event to register. Table 2.4 shows the number of high wind events verified by each criteria. The results differ by location significantly. Cheyenne did not have high wind events verified by the sustained criteria only while Fort Collins and Boulder had 29% and 34% of their high wind events verified by the sustained criteria only, respectively. Differences could exist in the meteorological dynamics unpinning the events at each location that dictates the reason Cheyenne did not observe any sustained-only high wind events during this two-year test period. However, there are also certainly differences in the observing equipment at each location and the algorithms logging sustained winds and gusts. As an automated surface observing system (ASOS), the Cheyenne sensor observation must maintain consistency in wind reporting with other airports across the country. The weather sensor at Christman Airfield in Fort Collins is not an ASOS and does not need to necessarily follow these observing rules, and the Boulder observations from the M2 Tower operated by NREL certainly do not either. Differences in observing equipment and data logging algorithms are beyond the scope of this study, but it is important to acknowledge them and investigate potential impacts on the results.

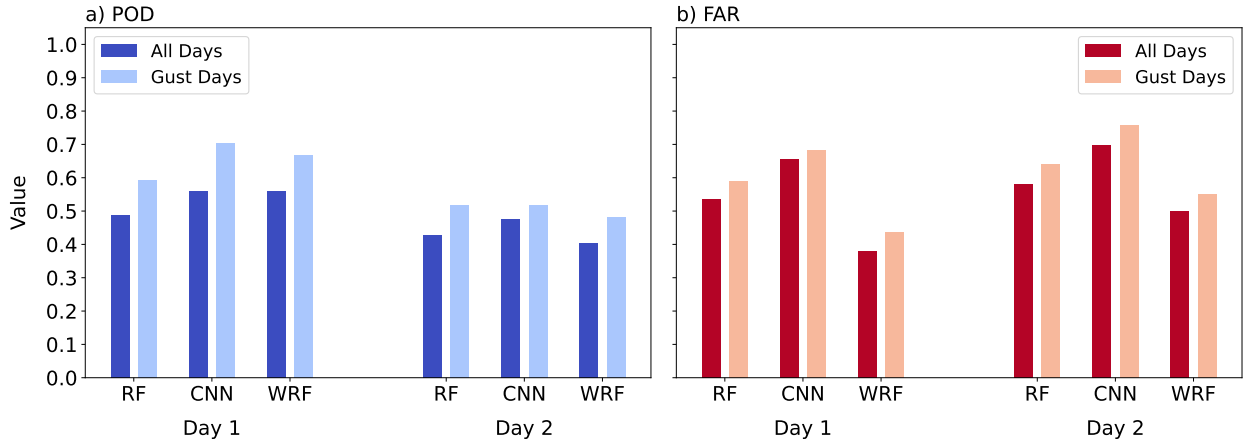

Figure 2.15: Day 1 and Day 2 high wind event (a) POD and (b) FAR for all models in Boulder including all days (darker shading) and gust days (lighter shading). Gust days do not include moderate and high wind events where only the sustained wind criteria verified. Non-events are still included in gust days in this analysis.

In the dichotomous metrics results above, the CSU-WRF's performance on Day 2 notable declined compared to the ML models. The differences in the sustained wind observations the locations provide motivation to recalculate the metrics for moderate and high wind events including non-event days and days where at least the gust criteria verified (gust days). This removes days where events verified with only the sustained wind criteria. While differences exist in the time period utilized when recording gusts, this study verifies against daily maximum winds so using the gust criteria at the three locations minimizes the impacts of the different observing platforms at each location.

Figure 2.15 compares the POD and FAR between all days and gust days specifically at Boulder because this location featured 14 sustained-only high wind events. With so few cases at the other two locations, any differences in metrics might only be due to characteristics of individual cases and not attributable to overall meteorological or observing platform characteristics. In Boulder, removing the sustained-only wind event days provides a boost in detection for all the models on both forecast days. However, we also note a corresponding increase in the FAR leading to minimal impact on the overall CSI. The increased POD indicates that the models detect these gust events more accurately when they occur, which is reassuring as these gust events would include the most

severe downslope windstorms. Even though we verified the CSU-WRF 10-m winds as sustained winds, the POD still improves on high wind gust events suggesting the 10-m winds represent a valid proxy for the stronger wind gusts as well. As the ML models are trained on the CSU-WRF forecasts, the ML model detection improvements could be solely due to the increased CSU-WRF detection especially since the magnitude of the improvement is very similar across all models.

2.6 Conclusion

In this chapter, we see that both the RF and CNN models are able to extend the forecast lead time of high wind events into Day 2 provided sufficient sample size of wind events in the training data. This is counterintuitive as the ML models' performances degrade less into Day 2 compared to the CSU-WRF on which the ML models are trained. We note that the CNN architecture detects more high wind events at the expense of a higher FAR. This requires the end users to make decisions on the relative impact of misses versus false alarms on their operations. Additionally, RFs benefit more from the increase in training data and events at Cheyenne and Boulder meaning implementing CNNs on sparse datasets such as Fort Collins proves to be a more viable strategy.

When verifying one wind event class, dichotomous metrics treat the other two non-event class equally requiring more subjectivity in the performance comparison between the models. Confusion matrices allow us to make these comparisons acknowledging the existence of all three categories, and we observe that the ML models are rarely incorrect by two categories especially for the most impactful cases of no wind event forecast on a day where high winds occurred.

The FIRM score quantifies the models' performances across all three categories simultaneously. With this multicategorical approach, we validate many of the conclusions drawn from the dichotomous metrics indicating their robustness. We highlight caution when implementing the FIRM verification as we demonstrate its sensitivity in the choice of risk threshold. In this study, a risk threshold above 80% tended to favor the CNN models' performances. Thus, a user must take care when selecting this parameter, which is not always a straightforward process especially in research scenarios with no documented forecast directive.

While the ML models show skill in certain situations in forecasting high wind events, none of the above metrics indicate that the models have learned anything physically meaningful about downslope windstorms themselves or that the way they arrive at their forecast solutions is trustworthy or meteorologically relevant to the forecaster. The next chapter aims to answer these questions by applying XAI techniques to these models to generate insights into the models' behaviors and the nature of the windstorms themselves.

CHAPTER 3: GENERATING INSIGHTS: APPLYING XAI TO ML MODELS TO ENHANCE FORECAST OPERATIONS

3.1 Introduction

In the previous chapter we presented RFs and CNNs that forecast moderate and high wind events at three locations along the Rocky Mountain Front Range: Cheyenne, WY, Fort Collins, CO and Boulder, CO. We showed that these models exhibit better capability at detecting high wind events in the 24-48-hour lead times than the 12-km CSU-WRF on which the ML models are trained. However, nothing inherent in the models' architectures or their performance guarantees that the models have learned anything about downslope windstorms or that their forecasts should be trusted. It seems unlikely that the models would achieve the preceding metrics due to pure chance due to their testing over two out-of-sample wind seasons. Physics-based weather models have equations derived from fluid dynamics, thermodynamics, and other disciplines that govern their behavior that, while still complex, can be understood by the operators even if only at a high level. ML models do not have physically meaningful equations underpinning their predictions leading to their common "black box" characterization (McGovern et al. 2019). There are equations and comprehensible relationships within their architecture, however, due to their complexity it is not intuitively apparent how a given model arrives at its output given the input features.

RFs are more readily understood especially when distilled down to a single decision tree. The mental exercise of starting at the top with a predictor value and following the decision tree down through its nodes and branches compares to human decision-making processes. This simplicity quickly breaks down in practice when considering the hyperparameters of the RFs presented in Chapter 2. These RFs receive 260,000 predictor points as their input, and while we understand that these points represent two-dimensional CSU-WRF output fields, the RFs treat each point as its own predictor. Furthermore, considering the RFs contain at least 100 and some over 1000 decision trees, keeping track of one variable value from a specific forecast hour at a specific gridpoint and drawing any meaningful conclusions about the model's behavior becomes impossible.

CNNs and neural networks in general present even further difficulties in understanding their behavior. Due to the amount of interconnected neurons and weights associated with these connections, following one piece of information through the network creates a compounding number of values to keep track of and computations to accomplish. Once again, drawing any physically relevant insight from this exercise would prove futile.

Researchers have created and applied specific techniques to ML models in order to explain their behaviors and understand how the arrive at their outputs. Any method that aims to visualize, explain, or interpret how ML-based models arrive at their predictions can be considered XAI (Samek et al. 2017). These XAI techniques allow users to gain trust in the models they operate and can also provide additional information about the models' use-cases. Examples of such efforts are discussed in the next sections.

This chapter discusses the application of XAI techniques to the RFs and CNNs previously presented and the resulting actionable insights relevant to the forecasting of downslope windstorms. Section 3.2 presents the feature importance, permutation importance, and saliency maps methods that are applied directly to the ML models to gain understanding of what the models deem relevant to their forecasts. In contrast, Section 3.3 discusses an indirect method of assessing model behavior through an unsupervised ML technique independent of the RFs and CNNs. Finally, a case study in Section 3.4 illustrates how to apply these XAI methods in a real forecast setting followed by the chapter conclusion.

3.2 Direct XAI Methods: Feature Importance, Permutation Importance, and Saliency Maps

Both feature and permutation importance are techniques applied to the ML model itself so the results are specific to each model and the input predictors. Feature importance is a property inherent to RFs trained in Scikit-learn and permutation importance is model agnostic so it can be applied to CNNs as well.

3.2.1 Feature Importance

Scikit-learn makes the feature importance readily available by running a python method on a trained RF classifier object. Thus, these feature importances are specific to each model and the predictors in the training set, not to data outside of the training dataset. This calculation attempts to quantify how well a given predictor divides the samples into their respective categories. This is accomplished by assessing at what depth within a given tree a predictor is used as a decision node as predictors used higher in the tree are splitting a greater proportion of the samples, which implies that this predictor is more important to the final prediction. Another way of thinking about it is to consider how many samples a given predictor contributes to the classification decisions. The more samples a predictor contributes to, the greater the importance of that predictor (Scikit-learn Developers 2025).

By averaging the estimates of the predictive ability of a feature over many trees within the RF, the variance of this estimate is reduced. This is known as the mean decrease in impurity, which Scikit-learn normalizes across all features and stores in an array accessible by the previously mentioned python method (Scikit-learn Developers 2025). These feature importance arrays can be analyzed in various ways, and next we present three different aggregations.

First, we analyze the importance of individual atmospheric variables and CSU-WRF forecast hours by summing the feature importances across all three locations' RFs. Figure 3.1 presents the summed feature importances for the Day 1 and Day 2 RFs by atmospheric variable. The zonal winds at 700-hPa and 10-m significantly outpace the other variables with the vertical motion at 700-hPa a clear third place on both forecast days. As these are summed across all three locations, we can generalize these observations to all of the RFs. Additionally, this figure reveals the importance of vertical motion at 300-hPa as this variable compares to the magnitude of importance with the meridional winds at 700-hPa and 10-m. This suggests the RFs recognize the role the jet stream plays in terrain-induced wind events as convergences at 300-hPa create downward vertical motion that aids in pushing elevated winds closer to the surface. Lastly, focusing on the thermodynamic variables, the RFs deem the potential temperature difference between 700-hPa and the surface as

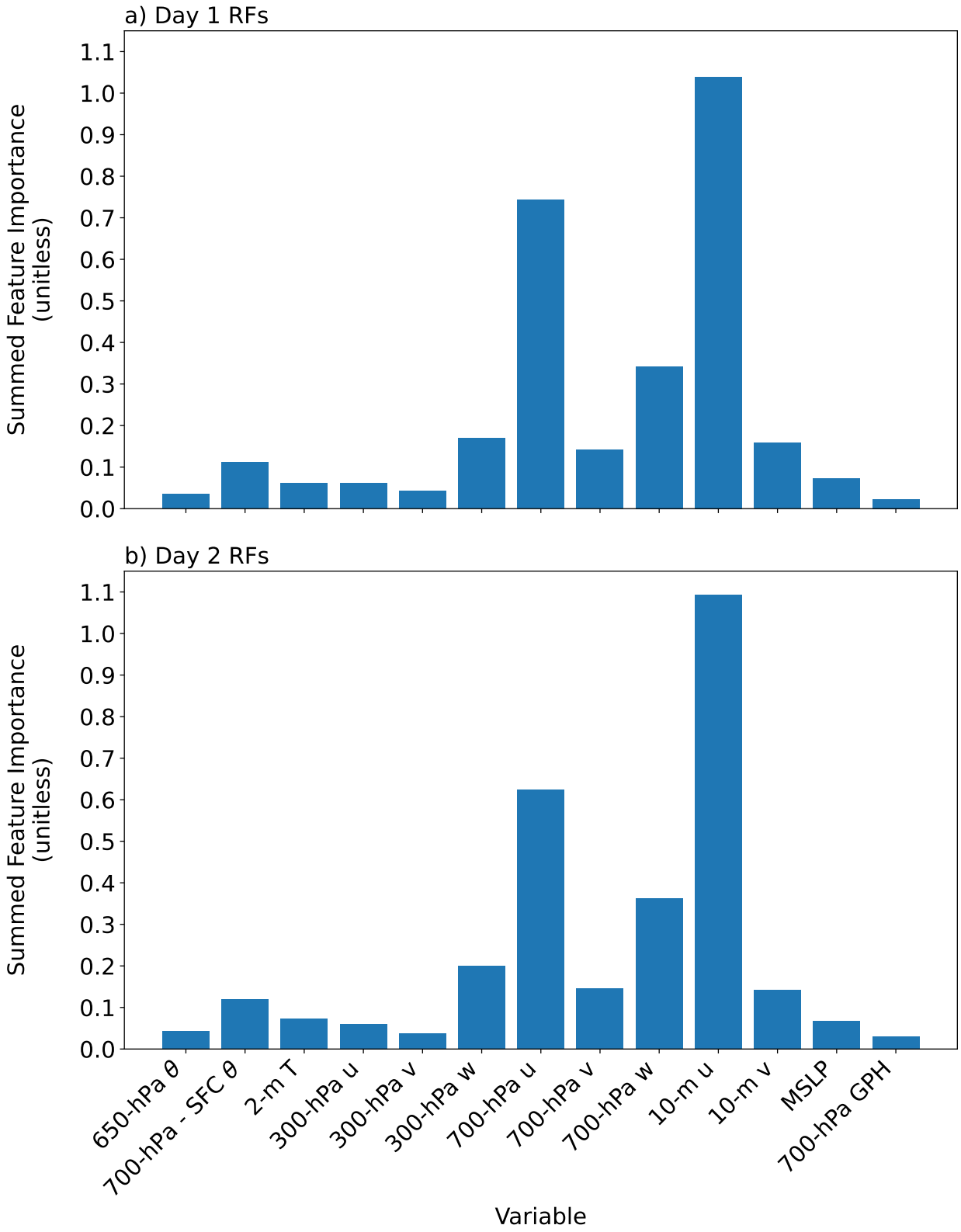

Figure 3.1: Summed feature importance values by atmospheric variable across all forecast hours for (a) Day 1 RFs and (b) Day 2 RFs for all three locations.

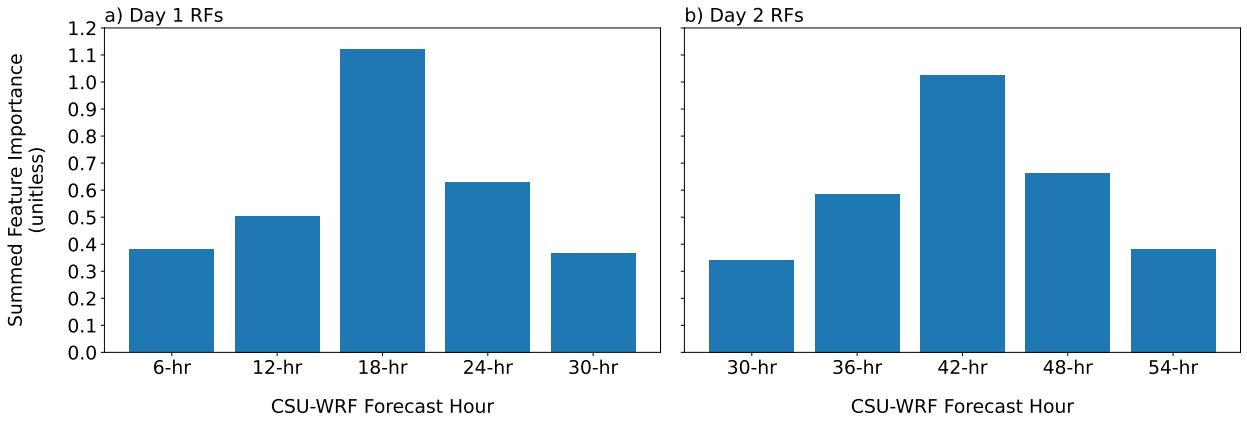

Figure 3.2: Summed feature importance values by CSU-WRF forecast hour across all atmospheric variables for (a) Day 1 RFs and (b) Day 2 RFs for all three locations.

the most important. As the proxy for stability, we know this is important for wave trapping and hydraulic flow characteristics over the ridgeline conducive to downslope windstorms.

Figure 3.2 displays the next feature importance plots that depict the summed feature importances for both forecast days for all atmospheric variables at all locations by each CSU-WRF forecast hour. The main observations is that the middle forecast hour provided on each forecast day receives the most feature importance, and in fact, the feature importances appear normally distributed from this midday peak. This could indicate a diurnal dependence in the downslope windstorms or more likely that the signatures in the input features the RFs learned are clearer in the 18 and 42-hr forecasts as these timesteps are farthest away from the verification window boundaries.

Because our features represent atmospheric variables on a geographic grid, we reshape the feature importance array back into our original three-dimensional predictor cube for plotting. This way we can assess the importance of the variables relative to their geographic location as well. This procedure results in 65 plots per RF: one plot for each of the atmospheric variables at each of the five forecast hours. We note that many of these plots contain features that are not important, and some plots contain features that are an order of magnitude more important than other features. Thus, we present only subsets of these plots based on the results from the aggregated feature importances above. All 65 plots for each RF are shown in Appendix A for completeness.

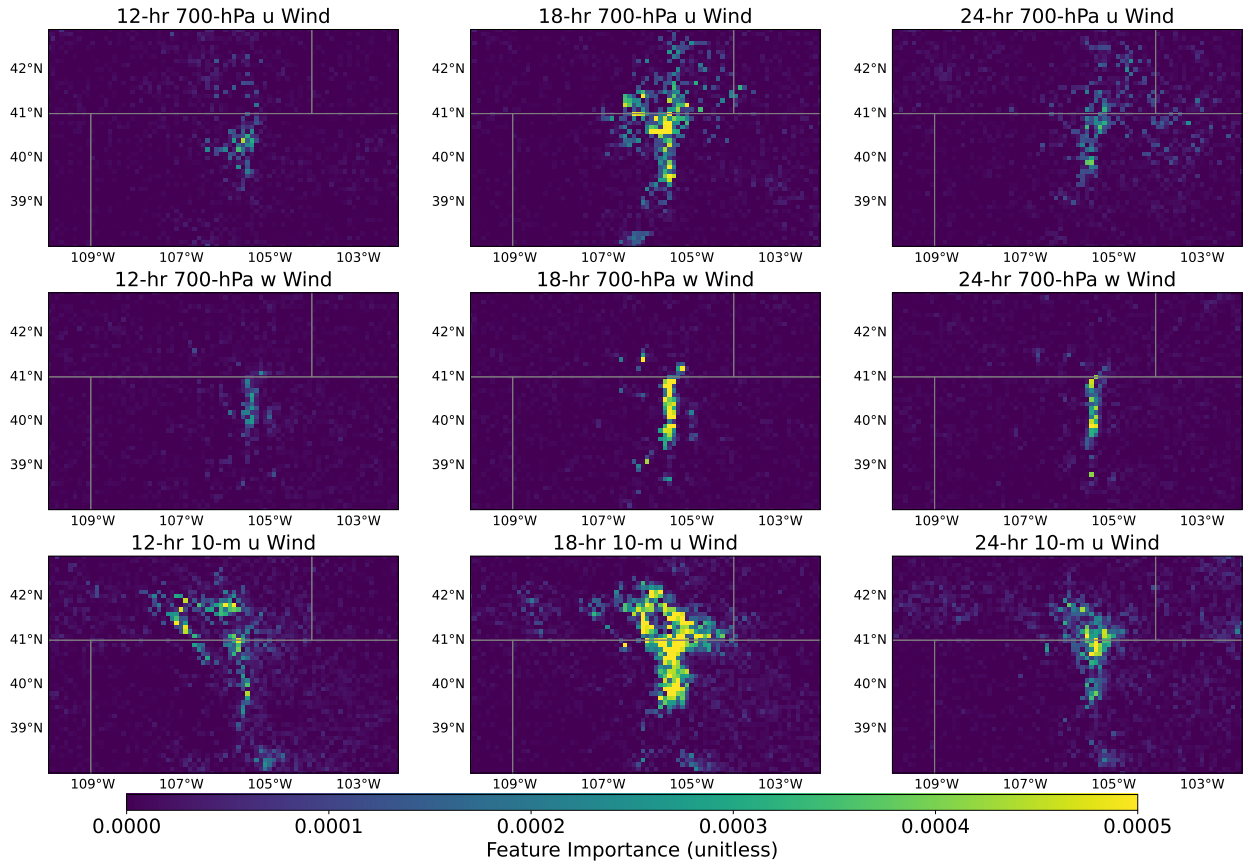

Figure 3.3: Feature importances for the Day 1 Boulder RF. The variables presented by row from the top to the bottom are the 700-hPa u wind, the 700-hPa w wind, and the 10-m u wind. These variables are shown at forecast hours 12, 18, and 24 by column from left to right. The brighter shading indicates greater importance of that variable at that location. Grey lines represent state borders. States within these plots are WY, NE, CO, and UT starting in the top-left corner and moving clockwise around boundaries of the plot.

Figure 3.3 displays the first sample of feature importance plots for the Day 1 Boulder RF for the 700-hPa u and w wind and the 10-m u wind for forecast hours 12, 18, and 24. These variables account for the majority of the feature importance for this RF. Note, the forecast hours depicted represent the middle of the forecast hours provided as input features to the RF. This means the RF relies on the forecasts from the middle of the verification window more heavily as the 18-hr forecast contains the brightest pixels for each of these three variables. As previously discussed, this could point to a diurnal dependence of windstorms as the 18-hr forecast is valid at 11 MST or roughly midday. However, windstorms and especially more temporally confined high wind observations occur throughout all hours of the day. We suspect that the models struggle more with

wind events occurring near the beginning or end of the verification window as the signatures in the predictors are spread across two different days' predictor sets though the RF will only "see" one of those sets. Thus, the RF deems the predictors during the middle of the day as more important as it more successfully classifies wind events occurring in the middle of the verification window.

The second feature of note in Figure 3.3 is the locations of the enhanced pixels. Recall that no direct terrain information is provided as predictors to the ML models. Surface pressure, which can be used as a proxy for terrain elevation, is part of the surface potential temperature calculation used in the stability predictor but is not otherwise directly given to the models. However, we see the feature importances highlighting the terrain of the Front Range. This makes sense as terrain plays a vital role in downslope windstorms so the values of these variables along the terrain are more important than their values at other locations. Furthermore, the RFs ingest these features as 1-D arrays and they are still able to learn the terrain through the predictive power of each predictor point. They have no awareness of where these points lie within the 2-D geographic domain or even the size of the domain.

The final aspect of the Day 1 Boulder RF feature importance plots worth discussing is the variables the model deems importance themselves. These three subset variables represent the zonal pressure gradient at approximately the ridgeline of the Front Range, the vertical motion at this level, and the zonal pressure gradient near the surface. We know that these particular aspects of the atmosphere are important when forecasting downslope windstorms, and the model agrees with this. The pressure gradient at the ridgeline provides the potential for the windstorms and the vertical motion either indicates wave breaking or the possibility of winds aloft being pushed down to the surface. The RF also learned that when the CSU-WRF forecasts stronger winds at the surface, that means higher winds are more likely.

While some of these observations on the RF behavior seem obvious or trivial, keep in mind the only influence this study had on the model's learning is the selection of input features, the architecture, and the hyperparameter tuning. All three of these aspects are optimized and experimented with, but no setting or code instructed the RF to prioritize utilizing these atmospheric variables in

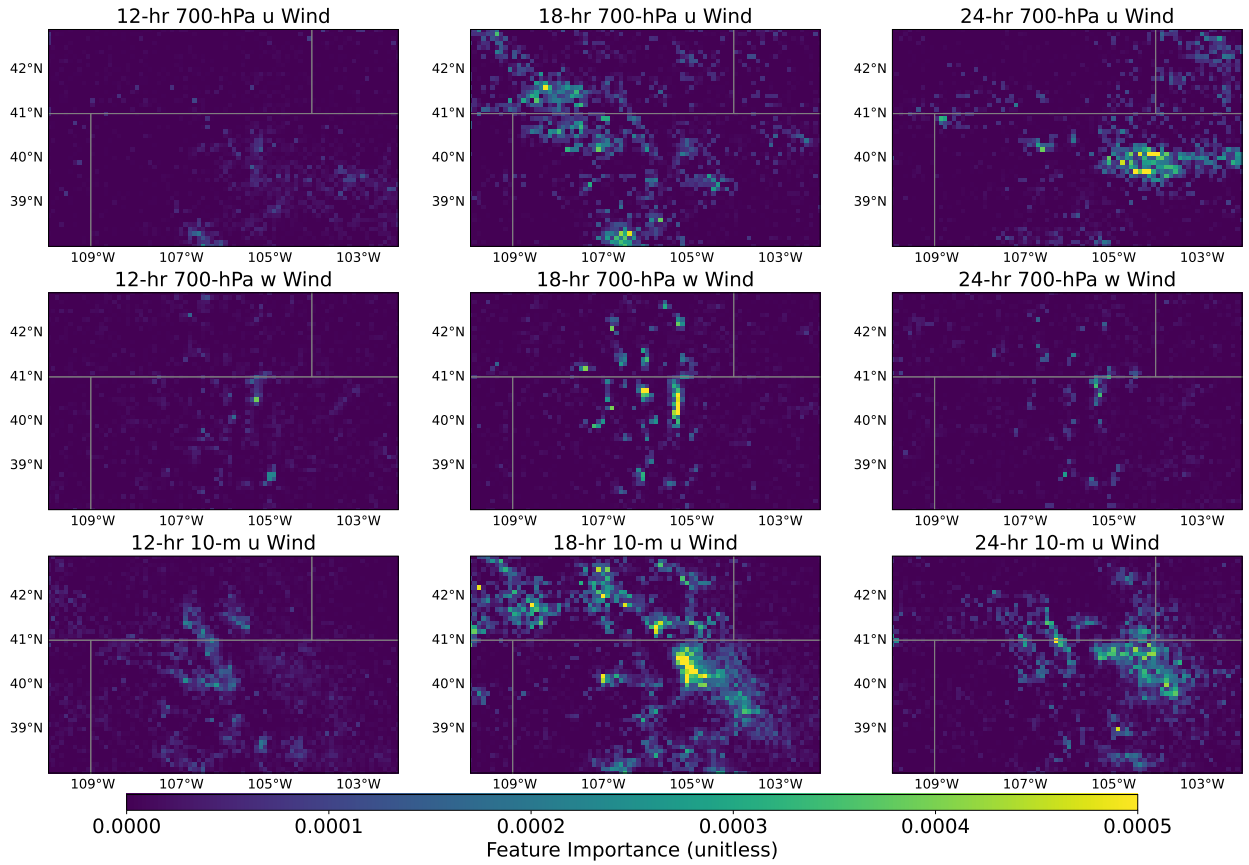

Figure 3.4: Feature importances for the Day 1 Fort Collins RF. Atmospheric variables, forecast hours, and shading same as in Figure 3.3.

or along the high terrain. Also noteworthy are the variables that are not important to the RFs predictions. For example, one might think that the 700-hPa geopotential height would be important, however, the information contained in this field is likely duplicated in the three components of the 700-hPa winds, which have greater influence on the model's predictions. Although these feature importance plots do not tell us anything new about forecasting downslope windstorms based on the discussion in Section 1.2, we gain trust in the model's predictions as these plots indicate behavior that mimics a human forecaster.

The same subset of variables and forecast hours of feature importances for the Day 1 Fort Collins RF are shown in Figure 3.4. Here we note many of the same observations discussed above, but two additional aspects contrast the Day 1 Boulder plots in Figure 3.3. First, on average the feature importance covers a larger area resulting in lower magnitudes of feature importance

at any one gridpoint. We know this RF performs worse than RFs at other locations due to the small sample size of wind events in Fort Collins. This study hypothesizes that the spread out feature importances is the RF attempting to look at more data points as it fits a solution to the training dataset. The greater number of gridpoints with feature importance means more points are contributing to the classification of a greater proportion of the training samples. In the Day 1 Boulder RF, less points contributing to the same proportion of sample classification results in less points with high importance. In Fort Collins, the RF must rely on more points to make its forecast, and while this is analogous to a forecaster analyzing more data, it is also another indicator that the forecast problem in Fort Collins is more difficult.

The second contrast in Figure 3.4 proves more subtle. Especially in the 700-hPa and 10-m u wind fields, a slight pattern of west-to-east motion appears in the feature importance locations. We still note the 18-hr forecast dominating the feature importance magnitude, but the overall shape of elevated feature importance begins upwind of the Front Range summits and transits east to the high plains over time. This suggests the RF tracks an atmospheric feature temporally through the input features and uses that in its classification decision. The wind events in Fort Collins, or at least the events that the RF is more successful, could be driven by transient features such as shortwaves. These plots alone cannot prove this, but this observation could point to an additional ingredient necessary to push the wind magnitudes over the threshold that is not necessary at the other two more frequent wind event locations.

This main purpose of the feature importance analysis is gaining trust in the RFs prediction process by understanding what they learned during model training. As many features the RFs highlighted as important to wind event forecasting are parallel to the atmospheric variables and geographic locations utilized by human forecasters, stronger evidence exists the RFs learned physically relevant information about downslope windstorms and do not identify wind events due to pure chance.

3.2.2 Permutation Importance

Despite the utility of the feature importance analysis analysis above, two weaknesses exist in the feature importance calculation. First, as an inherent property of the RF after training in Scikit-learn, feature importance is computed on the training dataset. An important feature to the model's predictions on the training data, therefore, does not mean this feature holds the same predictive power on a held-out dataset (Scikit-learn Developers 2025). Second, feature importance tends to favor data with high cardinality or more unique values (Scikit-learn Developers 2025). For example, the feature importance may be artificially inflated on a field of wind direction given in degrees from north compared to an 850-hPa temperature field. The former contains values ranging from 0° to 360° while the latter will only vary by a few tens of Kelvin unless something meteorologically significant is occurring. Permutation importance aims to ameliorate both of these concerns.

Permutation importance measures the contribution of each feature on a model's predictions on a given dataset (Scikit-learn Developers 2025). The method involves randomly shuffling input features and reevaluating the model's performance by comparing it to its baseline performance. As this does not involve the presence of any specific property of the model's architecture or prediction process, permutation importance is agnostic to the type of ML model. Thus, this study applies the permutation importance technique to both the RFs and CNNs.

Scikit-learn does include built-in functionality to assess permutation importance, however, this study implements its own methodology. Besides needing to apply this XAI technique to the CNNs built outside of Scikit-learn, correlations in the input features require specific handling. The goal when permuting a feature is to break the relationship between that feature and the target output to assess the model's performance when this relationship no longer exists (Chase et al. 2022). However, correlations within the features allow the model to rely on other non-permuted features that may mask the true importance of the permuted feature. Thus, an analysis must also account for these correlations when implementing this technique (McGovern et al. 2019).

The input features of this study consist of the output of a gridded traditional weather model where any gridpoint correlates with neighboring gridpoints both spatially and temporally. As a result, permuting one gridpoint at a time will not yield meaningful insights. We implement an aggressive input feature permutation scheme that both breaks these aforementioned correlations and also simplifies the number of permutation combinations that need testing. For each model, the values of the chosen variable are randomly shuffled across all samples, gridpoints, and forecast hours. We then recalculate the model's accuracy against the two-year test dataset and compare the accuracy to the baseline accuracy. A drop in accuracy implies this variable is important to the model's predictive power on this test dataset. This process is repeated 100 times for each variable for stability in the permuted accuracy values. All models are evaluated separately against the test dataset. This allows further generalization of the model behavior analysis as the model has not seen these test data during training or hyperparameter tuning.

Box and whisker plots depicting the results of the permutation importance trials by atmospheric variable are shown in Figures 3.5, 3.6, and 3.7 for Cheyenne, Fort Collins, and Boulder, respectively. The decrease in accuracy indicates variables with positive values (to the right of the vertical dashed line) cause the model to perform worse when that variable is randomly shuffled. The box and whiskers allow the display of the stability of the results across the 100 trials per variable; if a variable contains a narrower distribution of accuracy decrease value that indicates more robust results and vice versa for variables with wider distributions. However, we are not as concerned with specific accuracy decrease values, but rather the sign of those values and their magnitude relative to the other variables. For example, if a variable's entire distribution lies greater than zero, the results show this variable contributes to a decrease in model accuracy when permuted even if that variable's distribution spans a wider range of accuracy decrease values compared to other variables.

First, we note the difference in the distribution ranges between the RFs and the CNNs across all three locations on both forecast days. With the exception of the Day 2 Boulder RF, the permutation trials result in narrower distributions for the RFs compared to the CNNs. Also, the

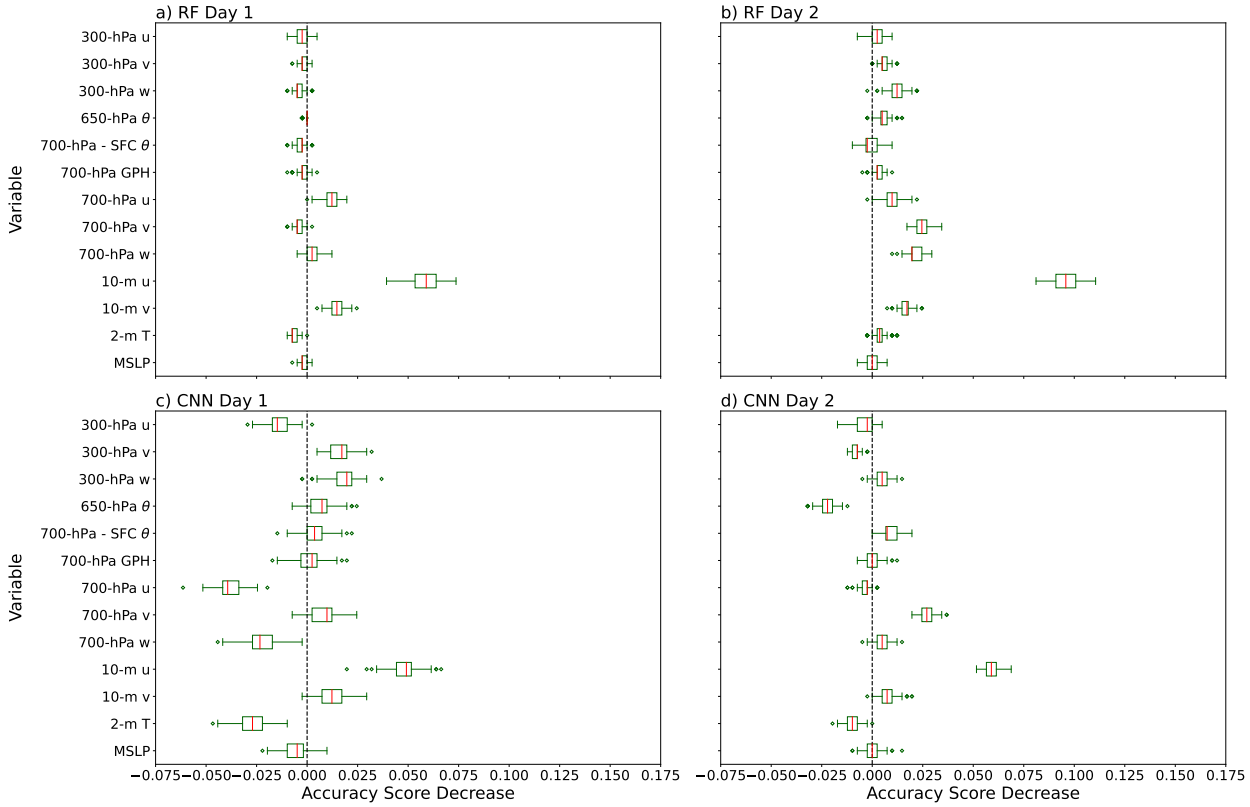

Figure 3.5: Box and whisker plots of permutation importance given by atmospheric variable versus model accuracy decrease for the Day 1 and Day 2 RFs and CNNs in Cheyenne. The vertical dashed line represents no change in accuracy score.

distributions farther away from zero tend to cover wider ranges of accuracy decrease values compared to distributions closer to or spanning zero. Therefore, these results suggest more certainty in the non-importance of these variables. Many variables across the models either have interquartile ranges (IQRs) or whiskers (depicting ± 1.5 * IQR) that intercept zero accuracy decrease meaning their inclusion as predictors may not increase the models' accuracy. This may also be due to correlations with the data despite the study's best efforts to break them. When one variable is shuffled there remains enough correlated information in the unshuffled variables leaving the resulting accuracy unimpacted. Only the Fort Collins Day 1 CNN contains IQRs and whisker ranges above zero for all of the variables. This shows that this network spread out the information during training as it relies on more variables for its predictive power. As noted above in the feature importance results, this speaks to the difficulty of the high wind forecast problem in Fort Collins. This may

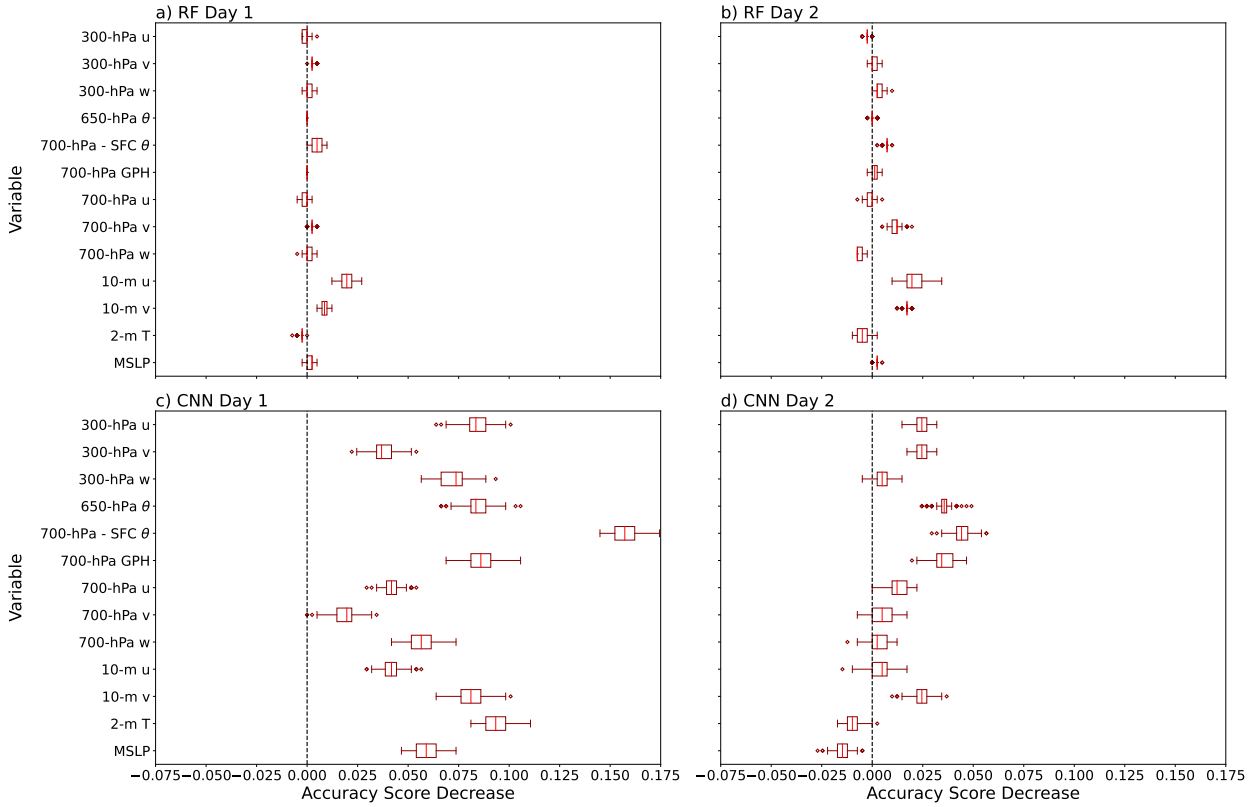

Figure 3.6: Box and whisker plots as in Figure 3.5 for Fort Collins.

also account for the CNN's enhanced detection capability compared to the Day 1 Fort Collins RF that uses far fewer variables in its predictions.

Across many of the models, the 10-m u wind increases the accuracy of the model and in some cases significantly so. This confirms the conclusions above drawn from the subset of feature importances, and we would expect such a variable to be important to an ML model. This also points to the strength of the 10-m u wind forecast of the CSU-WRF itself as it proves to be a valuable predictor to many of the models. In fact, looking back at the high wind contingency metrics in Figure 2.6, the CSIs for the CSU-WRF (derived from the 10-m wind magnitude) compare to the CSIs for the RFs in Cheyenne and Boulder. Correspondingly, the Cheyenne and Boulder RFs permutation importance results show greater accuracy decreases during the shuffling of the 10-m u wind values. Of course, it is possible the ML models learn to recognize bad 10-m u wind forecasts from the CSU-WRF and adjust their predictions accordingly, but given the performance metrics for the CSU-WRF in Section 2.4 this is unlikely.

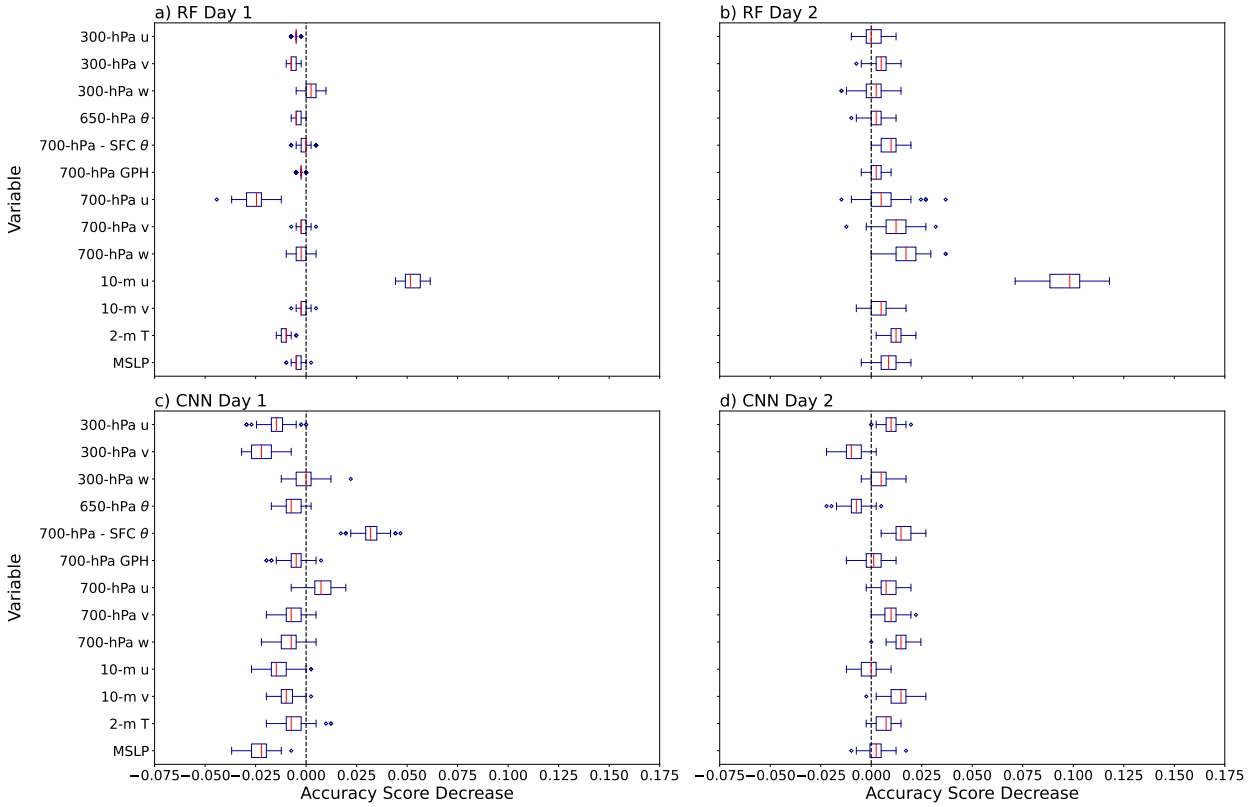

Figure 3.7: Box and whisker plots as in Figure 3.5 for Boulder.

Contrary to the feature importance results, the permutation of the kinematic variables at 700-hPa do not always result in accuracy decreases. Shuffling the 700-hPa zonal and vertical winds for the Day 1 Cheyenne CNN actually improves the model's accuracy. It is impossible to tell whether this is due to a physical reason in the atmosphere, but it could be that winds mix down from a higher level to the surface in Cheyenne and inputting winds at that level may benefit the model. Higher accuracy decreases associated with the vertical motion at the jet stream level compared to 700-hPa hint at this possibility.

The permutation importance technique highlights two other variables compared to the feature importance results. First, the stability variable, the potential temperature difference between 700-hPa and the surface, contains greater magnitudes of accuracy decreases especially for the Fort Collins and Boulder CNNs. Again, this builds trust with these models as a stable layer at the mountain ridge height is known to aid in the development of downslope windstorms. The stability variable is unique in the predictor field as most of the other variables represent kinematic fields and

the 650-hPa potential temperature and 2-m temperature do not themselves replicate any stability information. The higher accuracy decreases for the stability permutation could confirm that no correlation exists in the other data that repeats the stability information. Therefore, these results could speak to the uniqueness of the information contained in these fields and not the importance of the fields to a model's predictive power.

The second variable this technique highlights compared to the feature importance analysis is the meridional wind components at all levels. Initially, this study hypothesized the importance of the u component as this represents the component of the wind perpendicular to the Front Range, which runs approximately north-south. However, these results show that a stronger dependence on precise wind direction exists at each location. The input features contain no direct wind direction forecasts, so the models are able to use the relative magnitudes of the u and v components to learn this information. This boosts the importance of the v components that dictate the perpendicularity to the ridgeline of the winds.

Finally, as discussed above many variables seem to either not matter to the models' predictive power or serve to worsen their performances. It is possible there are too many variables in the input features that hamper the models' training processes leading to overfitting on signatures that do not predict high wind events accurately. Permutation importance can be used for feature selection where models are retrained using only variables important to the previous iteration's accuracy on the out-of-sample set. Also, the accuracy decreases depicted in Figure **??** include all non-events, moderate, and high events. This study heavily weighted maximizing high wind event CSI during hyperparameter tuning, not overall accuracy. Thus, while some variables' presence may decrease the overall accuracy of a model or look unimportant, they may still be important for the model's high wind event forecasts.

3.2.3 Saliency Maps

Saliency or gradient sensitivity examines how the output of a neural network changes given a particular input (Mamalakis et al. 2022a). Mathematically, this sensitivity, $S_{i,n}$, is simply the partial derivative of the output with respect to each of the input variables, X_i, at a given gridpoint

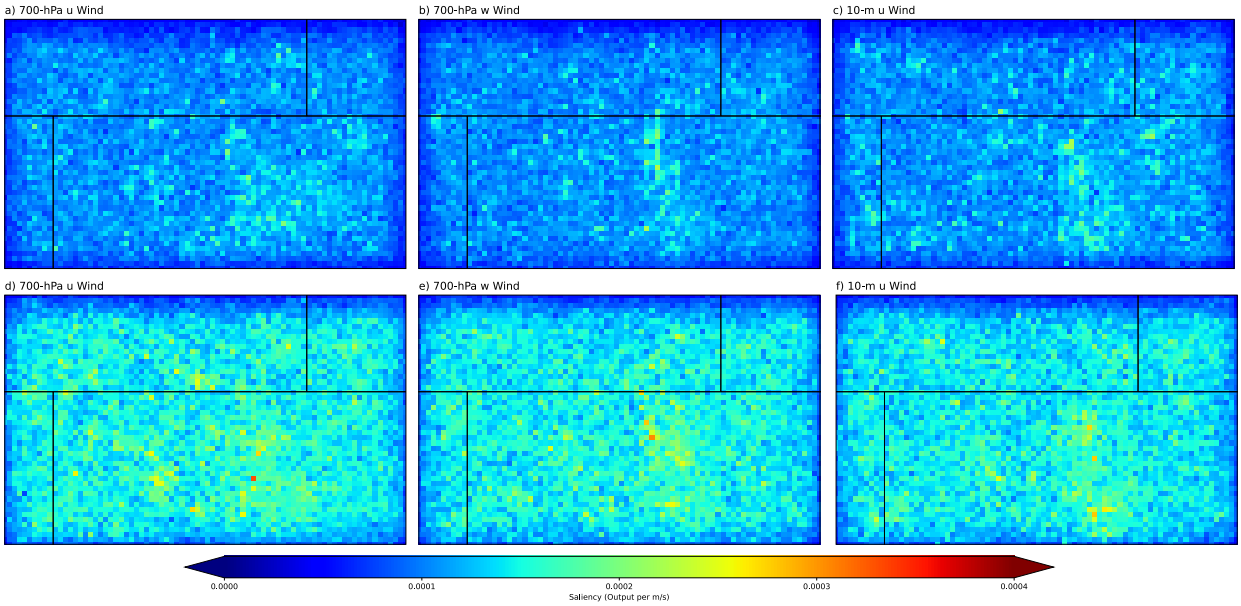

Figure 3.8: Saliency maps depicting the sensitivity of the high wind classification output on high wind event samples for 18-hr 700-hPa u and w wind and 10-m u wind for the Boulder Day 1 CNN without channel dropout layers shown in the top row (a-c) and with channel dropout layers shown in the bottom row (d-f).

for sample n:

$$S_{i,n} = \frac{\partial \hat{F}}{\partial X_i}\bigg|_{X_i = x_{i,n}}, \tag{3.1}$$

where $\hat{F}$ is the function learned by the neural network (Samek et al. 2017; Mamalakis et al. 2022b). The degree to which perturbing $x_{i,n}$ at a gridpoint results in a change to the network's output dictates the magnitude of this sensitivity. This method disentangles the sign of the sensitivity so positive (negative) saliency means changing the variable results in an increase (decrease) to the model's output. Additionally, saliency maps are not ignorant to zero input so if a model's output is sensitive to the absence of a particular variable this method still highlights these regions (McGovern et al. 2019; Mamalakis et al. 2022a). The output dimensions of saliency remain the same as the input predictors. This is advantageous as the results can be visualized as maps for meteorological predictors in a way familiar to forecasters (McGovern et al. 2019). In this way, we create saliency maps similar to the feature importance maps constructed in Section 3.2.1.

It is important to note this method estimates the sensitivity of the output to a given input variable but not the attribution or relative contribution to the output of that variable. The output may be sensitive to a given input, but the magnitude of the change in the output of this variable may be insignificant. The sensitivity represents the gradient of the output for that variable at a gridpoint while other XAI attribution methods aim to quantify how much the actual output changes (Mamalakis et al. 2022a). To illustrate the difference, imagine the output high wind probability neuron from one of this study's CNNs is very sensitive to the 700-hPa u wind over a particular gridpoint. This indicates a steep gradient in the model's learned predictive function and the accompanying saliency map would highlight this gridpoint. However, the possibility remains that this sensitivity only results in a small increase or decrease in the actual high wind probability. Thus, because the model output is sensitive to a particular input feature does not necessarily mean this feature is important.

The main drawback when creating saliency maps occurs when applying this method to deeper and more complex neural networks. In these situations, the gradient tends to converge to white noise as the spatial autocorrelation vanishes in a process termed "gradient shattering" (Mamalakis et al. 2022a). Indeed, this study encountered this problem while developing the CNNs. Figure 3.8 illustrates this problem as it pertains to this study through saliency maps for the 18-hr 700-hPa u and w winds and 10-m u winds, which are the variables that displayed stronger importance for the RFs above. Specifically, these maps depict the high wind classification output sensitivity to these variables on high wind event days. Initial CNNs whose saliency maps are shown in the top row of the figure did not contain channel dropout layers. Despite some noise in these maps, the saliency does increase in magnitude and cluster around the geographic areas of higher terrain similar to how we observed with the RFs previously. As discussed in Section 2.2.4, channel dropout layers added to the CNNs to decrease overfitting by increasing the strength of the regularization results in better performance. However, this added complexity increases the white noise signal in the saliency maps for these models shown on the bottom row of Figure 3.8. We could argue the features noted in the models without channel dropout are still present in the models with channel dropout, but

this would be difficult to identify without prior knowledge. Furthermore, due to the magnitude of the background noise it is difficult to determine whether other areas of clustered saliency values are actually significant. The result is models that perform better at their prediction task are less explainable with this XAI method. This study does not investigate saliency further nor ascribes any preliminary results to any physically relevant model behavior. The breakdown of this method provided motivation to move onto other techniques that do not directly involve the forecast ML models themselves as described in the next section.

This section focused on three techniques applied directly to the ML models, and the results remain specific to the model being tested and the dataset being used to conduct the testing. This next section presents a methodology that exists outside of the forecast ML models and utilizes the input features. This provides new insights into the ML models' performances and the nature of downslope windstorms at these three locations.

3.3 Indirect Method: Dimensionality Reduction and Clustering of Input Features

3.3.1 Background

Other methods exist outside of analyzing the forecast ML models' behaviors directly as understanding features of the input predictors and even the CSU-WRF forecasts outside of our predictor domain also provides insights useful to the final forecasting process. Stepping away from the forecast ML models is not traditionally considered XAI, but previous studies have applied these indirect techniques to understand atmospheric processes and traditional weather model behavior when forecasting them. The field of atmospheric science benefits from a wealth of both observational and forecast data (whether through reanalyses, reforecasts, or climate projections), and specifically these studies attempt to synthesize these data for a given forecast problem through automated weather typing.

Jiang et al. (2015) implemented an approach utilizing self-organizing maps (SOMs) on reanalysis data to identify synoptic weather types over Australia. The study optimized the number of maps (weather types) for cluster analysis and data projections in addition to finding that data standardization improved the representation of large-scale and small-scale synoptic features and

pattern recognization in the SOMs. They extended this framework to understanding air quality over Sydney, Australia. Due to Sydney's subtropical coastal-basin environment, the spatial variability in air quality is dictated by both synoptic and mesoscale features. The SOM methodology provided greater linkages between the observed mesoscale interactions and the identified synoptic features in the SOMs rather than relying on the known synoptic patterns alone. This led to higher fidelity in the air quality forecasts for the city (Jiang et al. 2017).

Two studies focused on identifying synoptic patterns over North America that drive extreme precipitation events over the continent. The first study used data from PRISM and the Livneh group in addition to ERA5 reanalysis to identify one to four weather types that preceded the heaviest precipitation events across the continental United States including events considered unprecedented at the time. They noted that in 16 of the 18 watersheds they examined the frequency of these weather types is increasing (Prein and Mearns 2021). The second study focused on the predictability of precipitation from the North American Monsoon (NAM). The results showed skill in forecasting NAM precipitation at the sub-seasonal timescale through typing long-range weather model data from the Integrative Forecasting System (IFS) from the ECMWF as wet, normal, or dry, which aids regional planners' water management strategies. Furthermore, they attributed 500-hPa height anomalies resembling a Rossby wave train as the source of predictability for the top 25% of years with the highest frequency of monsoonal flow (Prein et al. 2022).

Moving into the deep learning sphere, Chattopadhyay et al. (2020) demonstrated the capability of coupling a CNN to clustered 500-hPa height fields from the Large-ensemble (LENS) Community Project to identify cold spells and heat waves in North America solely. The CNN predicted whether an event would transpire at 1-5 days lead time with at least 25% higher accuracy than random chance. The addition of 2-m temperature contributed to a reduction in false alarms that plagued the initial results that only utilized 500-hPa heights. The study demonstrated the predictive power of a combined deep learning and clustering approach on limited fields from existing model data.

The final study we discuss also describes a fusion of deep learning with cluster analysis, and provides the main motivation for the methodology in this section. Li et al. (2022) identified "Dunkelflaute" events over Europe with a framework that applied k-means clustering to patterns encoded by a convolutional autoencoder (AE). Dunkelflaute events are weather patterns producing simultaneous reduced low-level winds and solar insolation due to cloud cover. While these events do not feel extreme to the casual weather observer, they greatly inhibit solar and wind power production causing challenges for energy system operators balancing supply and demand especially as these events are more common during the cold season. The k-means algorithm clustered 25 anomalous patterns of which five were identified as producing a higher incidence of Dunkelflaute events. Commercial power production data confirmed that these clusters exhibited reduced renewable energy production on days belonging to those clusters. By identifying these days with publicly available model forecast data, these Dunkelflaute producing clusters can be forecast in advance without relying on proprietary commercial power production trends.

3.3.2 Methodology

This section presents the methodology employed by the Dimension-Reducing Autoencoder Gaussian Mixture Model (DRAGMM) framework. As previously discussed, the DRAGMM is inspired from the Chattopadhyay et al. (2020) and Li et al. (2022) studies by combining deep learning with a cluster analysis. Specifically, we use a convolution AE to create encoded images of the input predictors that are then clustered by a Gaussian mixture model. We can then study the incidence of high wind events and the performance of the forecast ML models from Chapter 2 in each cluster. We hypothesize that we can successfully create clustered data that allows for actionable insights in an operational forecast setting.

Recall that the input predictors for this study comprise a 50x80x65 cube of gridded forecast data from the CSU-WRF. As a result, these data are highly dimensional, which would be challenging to cluster directly. Thus, we need a way to distill the information contained in these highly dimensional predictors into an image with lower dimensionality. AEs are neural networks that detect features within the input data and learn latent representations of these data with lower

dimensions without supervision (Géron 2019). This study trains convolutional AEs for the same reason CNNs are chosen as forecast models above, they excel at image recognition tasks because the include convolutional and pooling layers similar to CNNs.

Specifically, the AEs consists of two parts that can be run in inference mode separately after training. The encoder comprises the first part, which consists of convolutional, max pooling, and channel dropout layers to reduce the dimensions of the 50x80 feature maps into a single 3x5 pixel image. The output of the encoder is called the latent dimensions or encoded image. The job of the decoder is to reconstruct the original input feature grids from this encoded image through deconvolutional layers and channel dropout layers. The output image from the decoder is compared to the original input image to quantify the success of the decoder. If the decoder successfully reconstructs the original input features consistently, then the encoder is also successfully capturing the variance in the original input features in the encoded images. Therefore, the encoder reduces the dimensions of the input features while preserving the variance in the data. The advantage of the AE approach is as an unsupervised ML method, no labels are required to train an AE. Essentially, the input images themselves are the labels as the training task is to reproduce the input images while squeezing the data through the latent dimensions.

We train separate AEs for the Day 1 and Day 2 predictors. Training is accomplished with the same training period as the forecast CNNs with hyperparameter tuning occurring over the same validation dataset. The test period consisting of the 2021 and 2022 wind season remains held out for future case studies. Both AEs capture a high level of variance of the validation input features with R^2 values of 0.723 and 0.740, respectively, when comparing the difference between the input images and the reconstructed images. We now have 15-pixel encoded images of our originally 260,000-pixel input feature images representing four orders of magnitude dimensionality reduction.

With these encoded images, we introduce an algorithm to separate the images into clusters. As each image represents a day in our dataset, the goal is to create clusters with meteorologically relevant features that lead to insight generation. Unlike the Li et al. (2022) study that implemented

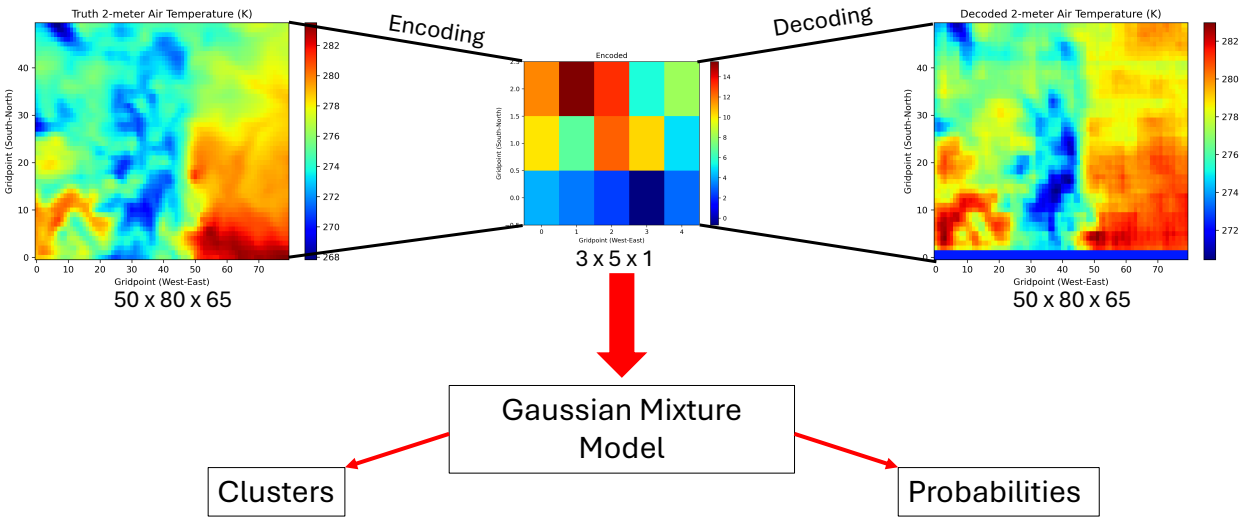

Figure 3.9: Schematic of the DRAGMM framework. This example shows the reconstruction of a 2-m temperature field. The encoded image in the center contains the latent information from all 65 channels, not just the 2-m temperature channel shown. The GMM receives this encoded image and determines the probabilities of it belonging to each cluster.

the common k-means clustering algorithm, the DRAGMM clusters the encoded images with a Gaussian mixture model (GMM). A GMM assumes the data being clustered are generated from a finite number Gaussian distributions with unknown parameters. This essentially generalizes the k-means technique by incorporating information about the covariance structure of the data (Scikit-learn Developers 2025). The first attempt at clustering did utilize k-means within Scikit-learn, but found poor results in metrics quantifying model quality and model complexity. The k-means models became overly complex without creating enough separation between the clusters of encoded images. Training a GMM resulted in better clustering with simpler models. An secondary advantage of GMMs is they not only predict to which cluster an encoded image belongs, but also the probability the image belongs to that cluster along with the corresponding probabilities from the other clusters. Therefore, we gain information on the confidence of the DRAGMM framework in addition to the clusters themselves.

Figure 3.9 summarizes the DRAGMM framework. After training the AE, the decoder is no longer involved in the inference stage as it only serves to ensure the encoder captures as much of the initial variance as possible. The DRAGMM can operate in forecast mode after the CSU-WRF data

are produced. The same daily predictor fields supplied to the forecast ML models are encoded with the AE followed by cluster identification by the GMM. This allows the insights associated with this framework to be actionable as they are available to the forecaster in real-time. A demonstration of this follows the results in the form of a case study.

3.3.3 Results

Identifying the clusters comprises the first task concerning the results. Clustering the training and validation data into four clusters proved sufficient both from a model performance standpoint and for physical interpretation. We tested higher numbers of clusters, but found that the additional clusters were combinations of the original four clusters adding needless complexity. The remainder of the study focuses on the four identified clusters that result from the DRAGMM framework.

In order to extract meteorologically relevant features from each cluster, we composite the atmospheric variables at each geopotential height level included in the input predictors across all days belonging to each cluster. All five of six hourly forecast hours that make up the day's input predictors are included in the composite. This allows comparison of the means of these fields between the clusters, and also aids in naming the clusters. We refer to these as feature domain composites because the composites are created with the actual features input into both DRAGMM and our forecast ML models. The feature composites for Day 1 are presented in Figures 3.10-3.12.

We first analyze jet stream characteristics of each cluster by examining Figures 3.10a-d. As the strength and orientation of the jet stream is the clearest meteorological difference between the clusters, we name the clusters “bora," “benign," “strong jet," and “weak jet" respective of subfigure order in Figure 3.10 and refer to the clusters by this naming convention henceforth. The strong jet labeled cluster certainly has the strongest winds at 300-hPa in addition to the highest magnitude of downward vertical motion along the Front Range. We already expect to see favorable dynamics for downslope windstorms in this cluster. The weak jet cluster also appears to be named appropriately as we note the weakest 300-hPa winds in this cluster. However, downward vertical motion still exists along the terrain between Fort Collins and Boulder. The benign cluster depicts the jet core to the southwest of the forecast locations implying troughing over the area with neutral vertical

motion. Finally, the jet placement in the bora cluster is difficult zoomed into the feature domain, but we observe neutral vertical motion at this level as well.

The 700-hPa geopotential height and vertical motion fields in Figure 3.10e-h provide additional context to the jet features noted in above. All four clusters display some magnitude of leeside troughing that we would expect to see during synoptic flow over terrain. We also note the characteristic coupling of upward vertical motion windward of the Front Range and downward vertical motion to lee where our forecast location exist. However, the features at this level in the benign cluster are very weak so less wind events are to be expected in this cluster. On the opposite end of the spectrum, the signatures are most prevalent in the strong jet cluster as anticipated from the 300-hPa conclusions. The bora cluster exhibits strong northwest flow hinting at the presence of cold air-driven events that aligns with the definition of a bora wind (Markowski and Richardson 2010). The weak jet cluster 700-hPa pattern is similar to the strong jet cluster though appropriately weaker.

Figure 3.11 remains in the middle levels with the thermodynamic feature composites. The first four maps show the 650-hPa potential temperature composites, which continue to support the naming convention. The bora cluster confirms cold air pushing over the Front Range even despite minor adiabatic warming in the lee. The benign cluster is also cold owing to the trough noted aloft, but without the downslope adiabatic signature, we continue to expect wind events to be rare in this cluster. Both jet clusters indicate strong warming downwind of the terrain. The tighter temperature gradient in the strong jet cluster indicates the stronger winds relative to the weak jet cluster.

The bottom half of Figure 3.11 provides the stability proxies for each cluster. In these images, negative values in the 700-hPa to surface potential temperature difference are masked as zero due to model terrain intersecting the 700-hPa pressure level so we only draw conclusions regarding the positive stability between the clusters. Stability exists in every cluster downwind of the terrain, which is expected for mean conditions over the nonconvective wind season occurring mainly during the winter months. Looking upwind of the terrain provides more details into the separation of the clusters. The strongest upwind stability exists in the bora cluster once again pointing to a cold

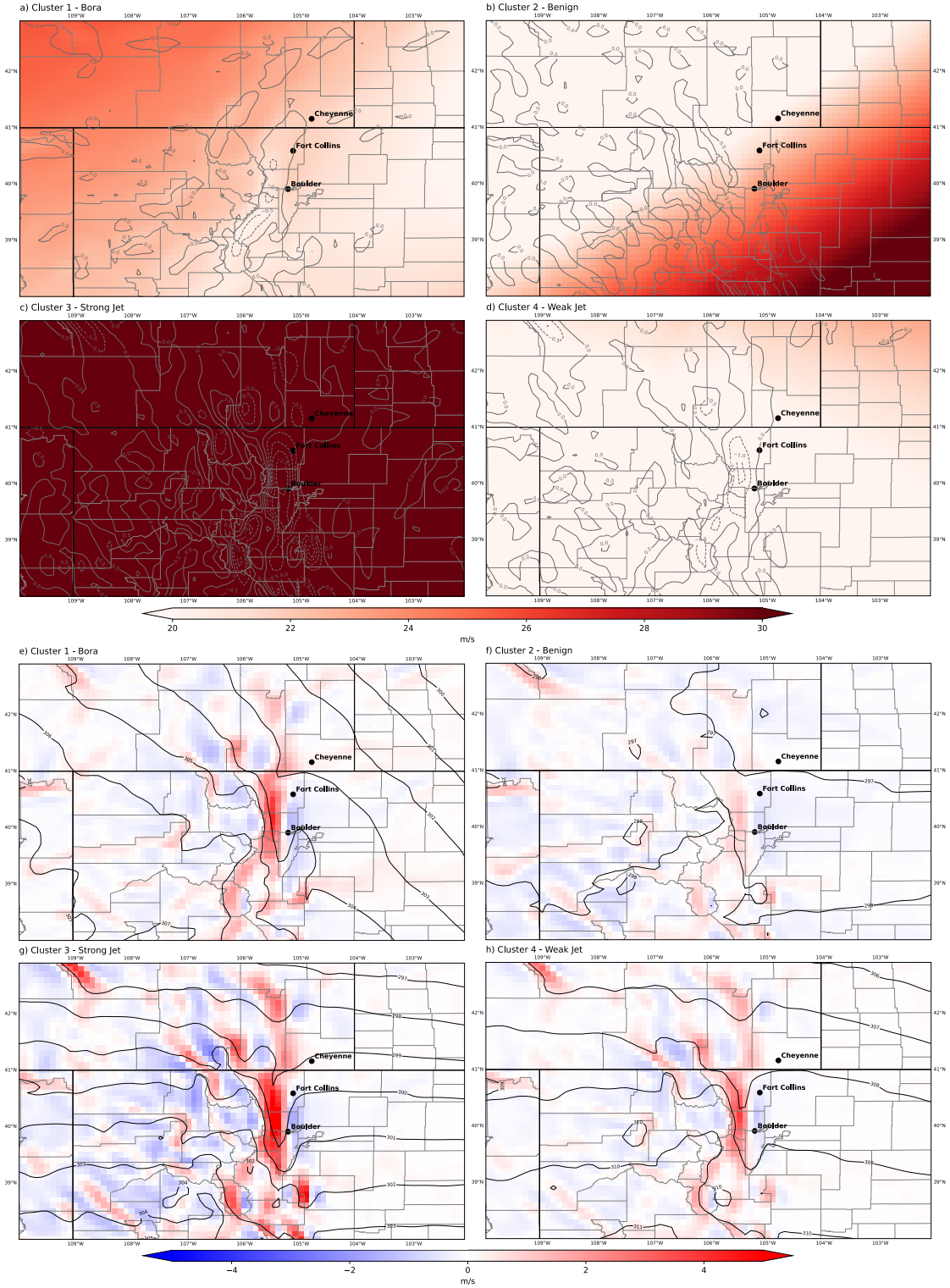

Figure 3.10: Day 1 feature domain composites for each cluster of (a-d) 300-hPa horizontal wind magnitude (m s^{-1}, red shading) and vertical motion (m s^{-1}, grey contours) and (e-h) 700-hPa geopotential height (dam, black contours) and vertical motion (m s^{-1}, red and blue shading) over the input feature domain. Cheyenne, Fort Collins, and Boulder are labeled for geographical reference in this figure and subsequent feature composite figures.

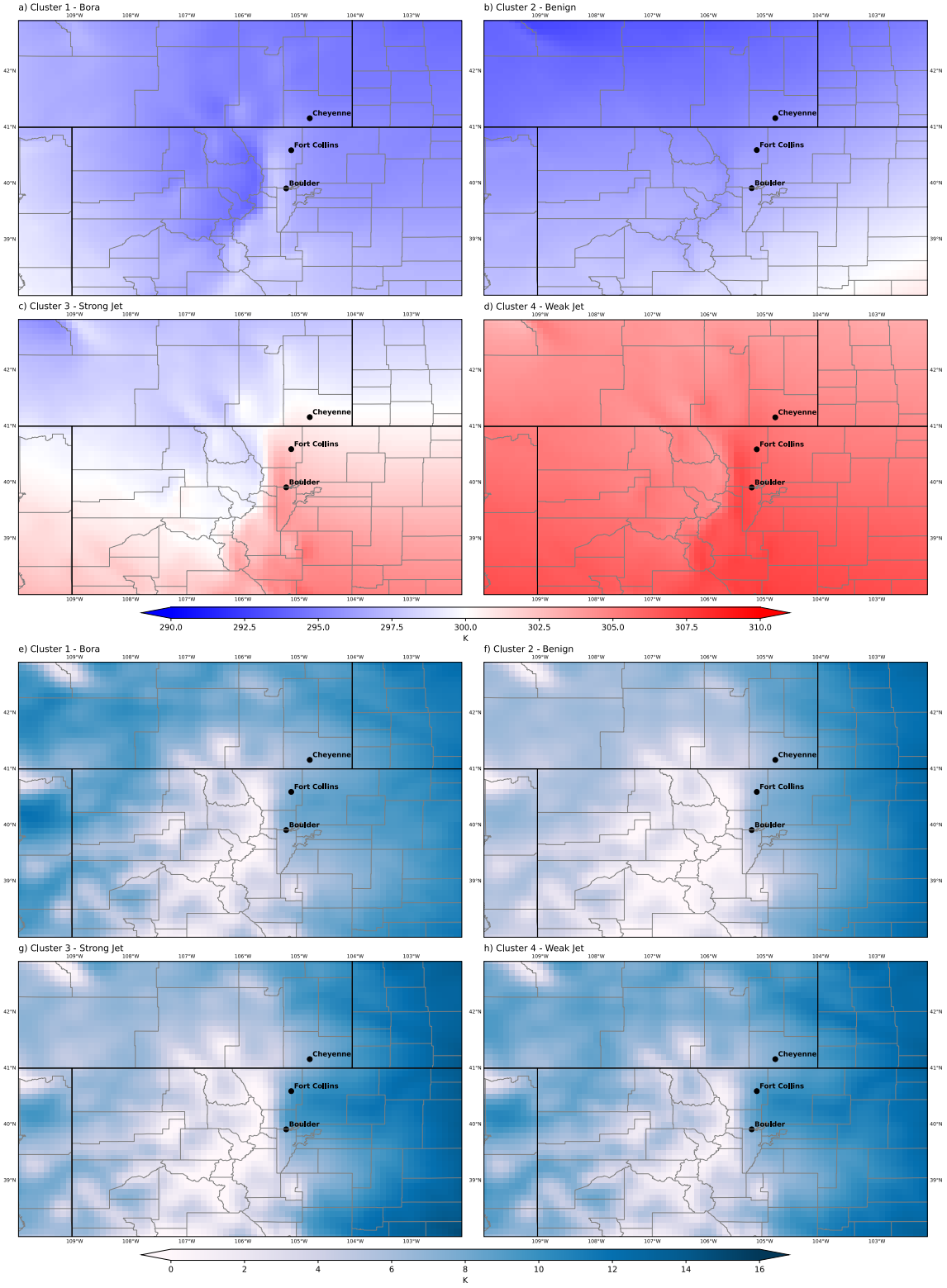

Figure 3.11: Day 1 feature domain composites for each cluster of (a-d) 650-hPa potential temperature (K, red and blue shading) and (e-h) the potential temperature difference between 700-hPa and the surface (K, turquoise shading). Negative values in the 700-hPa to surface potential temperature difference are masked as zero as much of these negative values result from the model terrain intersecting the 700-hPa level.

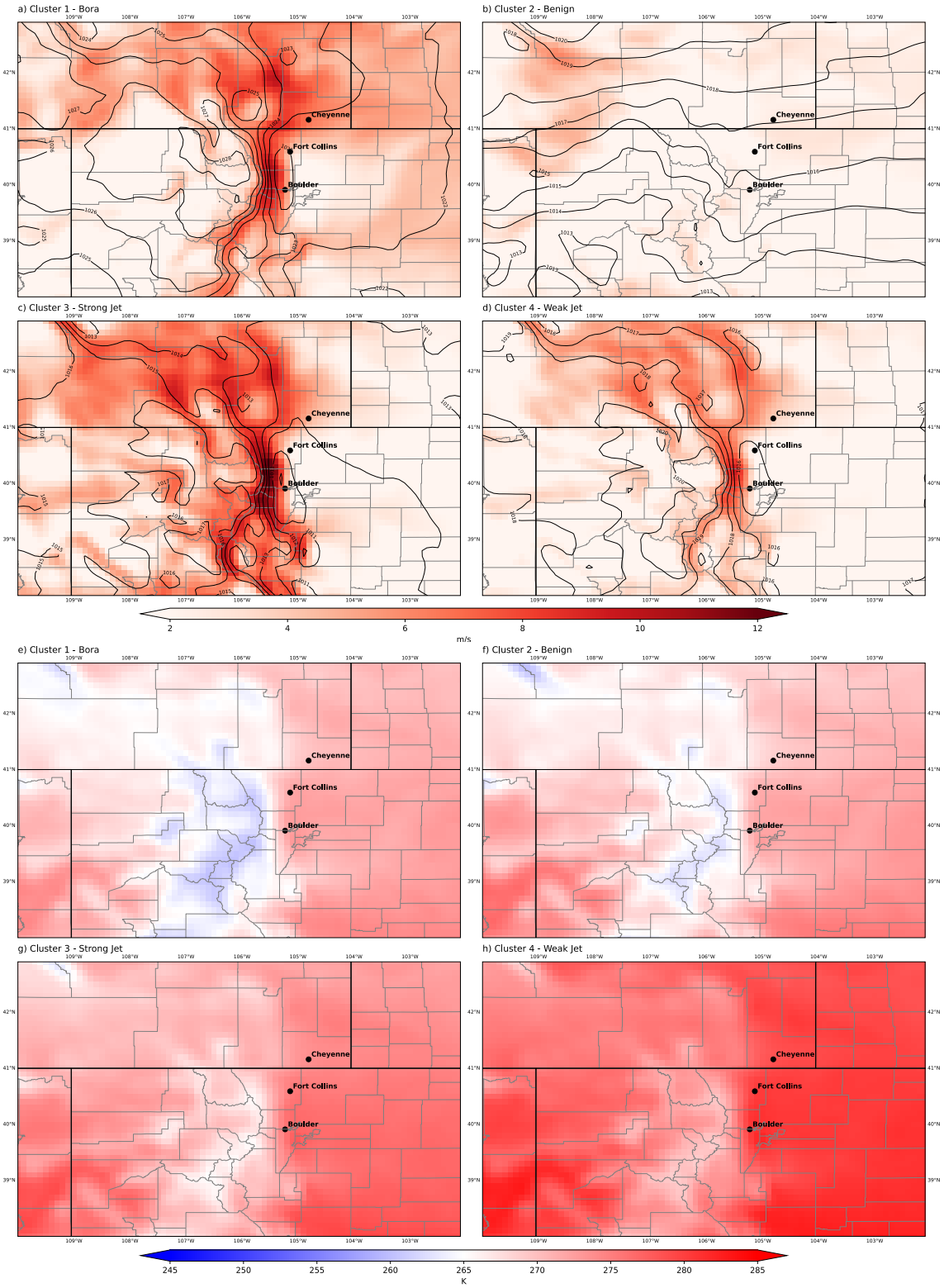

Figure 3.12: Day 1 feature domain composites for each cluster of (a-d) MSLP (hPa, black contours) and 10-m wind magnitude (m s^{-1}, red shading) and (e-h) 2-m temperature (K, red and blue shading).

air outbreak driving any wind events occurring in that cluster. The benign and strong jet clusters contain weaker upwind stability, but the known stronger kinematics present in the strong jet cluster likely overcome the weakened upwind wave trapping. The weak jet cluster has stronger upwind stability than the strong jet cluster potentially indicating better wave trapping dynamics before the flow encounters the Front Range ridgeline. This may compensate for the weaker jet dynamics characteristic of this cluster meaning wind events in this cluster may be more thermodynamically driven than kinematically driven.

The final feature domain composites for Day 1 comprised of near-surface atmospheric variables are shown in Figure 3.12. The strongest mean surface pressure gradient and 10-m wind speeds belong to the strong jet cluster followed by the bora and weak jet clusters. These three clusters also depict the north-south orientation of the pressure gradient required for westerly downslope winds. This pressure gradient is tightest between Fort Collins and Boulder, though the north-south extent is greater in the strong jet cluster. The higher mean 10-m wind magnitudes extend east the Cheyenne in the bora and strong jet clusters indicating high wind events may be less favorable at this location during a day classified as weak jet. This north-south pressure gradient does not exist in the benign cluster with the exception of minor leeside troughing. The 2-m temperature composites mainly mirror the previous discussion of the 650-hPa potential temperature composites.

As we know which days in the training and validation set belong to which clusters, we expand our composite analysis to the entire domain of the CSU-WRF forecast. These data are not seen by DRAGMM or the forecast RFs and CNNs, but we can gain further insight to the nature of these clusters at the synoptic scale since the clusters are driven by the input features derived from the CSU-WRF forecast. We also extend the analysis to atmospheric fields not included in the input features. Figures 3.13 and 3.14 display these maps referred to as full domain composites of each cluster.

Starting at 300-hPa in Figure 3.13a-d, we now observe the whole picture of the mean synoptic pattern in each cluster. The bora and benign clusters appear similar but with an important difference in the location of the trough axis. The trough in the bora cluster appears more transient allowing

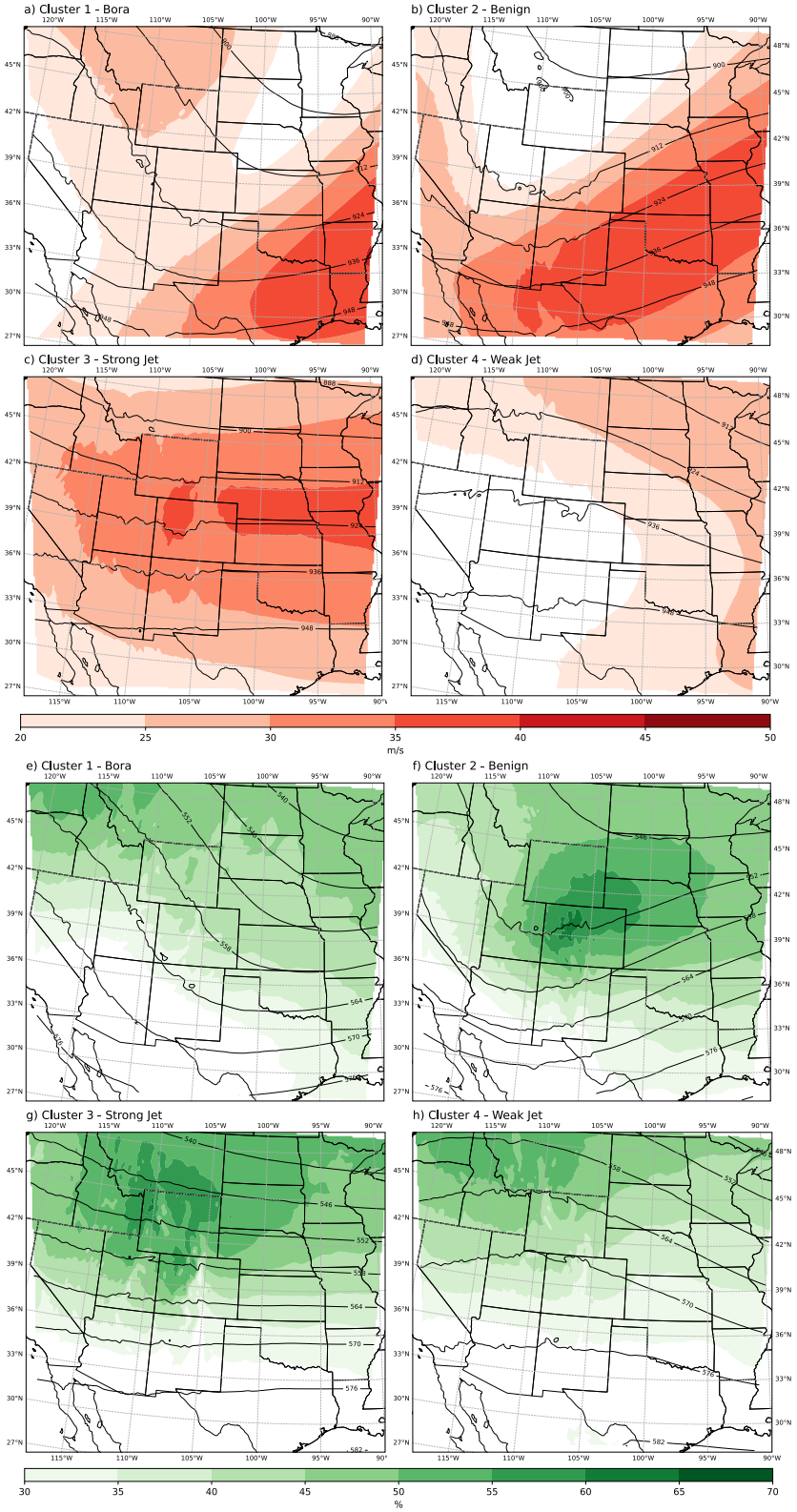

Figure 3.13: Day 1 composites of (a-d) 300-hPa geopotential height (dam, black contours) and wind speeds (m s^{-1}, red shading) and (e-h) 500-hPa geopotential height (dam, black contours) and relative humidity (%, green shading) for each cluster across the full CSU-WRF domain.

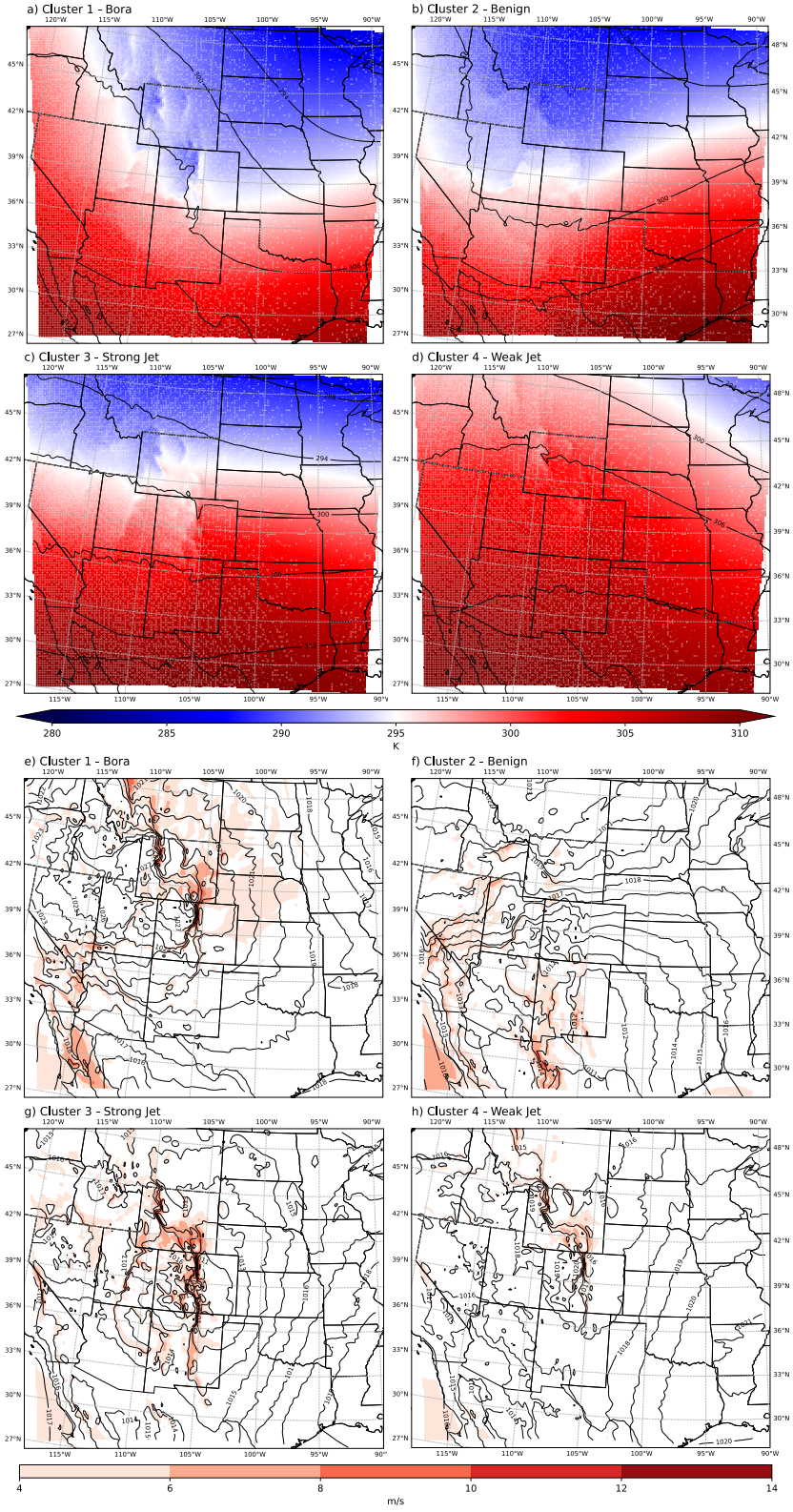

Figure 3.14: Day 1 composites as in Figure 3.13 but for (a-d) 700-hPa geopotential height (dam, black contours) and potential temperature (K, blue and red shading) and (e-h) MSLP (hPa, black contours) and 10-m wind speed (m s^{-1}, red shading).

the nose of an upwind jet streak to impinge the forecast area providing the upper level dynamics necessary for wind events. The benign cluster, however, positions the trough slightly upstream and more positively tilted resulting in southwest flow over the Front Range that is not conducive to downslope windstorms for the three forecast locations. This explains why benign aptly describes this cluster despite the nearby presence of the jet stream. The strong jet and weak jet clusters depict 300-hPa wind speeds consistent with their nomenclature. The strong jet is characterized by zonal flow, and the weak jet cluster shows broad ridging upstream. Both hint at the possibility of embedded shortwaves possibly important to the onset of higher low level winds. Moving down to 500-hPa in the bottom half of the figure, we mainly note the 300-hPa features reflecting down as expected with no other notable observations.

Similar patterns exist in the potential temperature fields at 700-hPa in Figure 3.14 as in the potential temperature predictors at 650-hPa. The trough in the mean geopotential height contours at this level is well downstream of Colorado and Wyoming in the bora cluster, which reinforces the northwest flow attributable to a cold air outbreak. The shortwave trough over the Front Range is more noticeable in the strong jet cluster at 700-hPa, but it is difficult to known whether this is a feature antecedent to wind events or a reflection of the resultant leeside pressure falls.

So far we have discussed bora events at length as that pattern is clearly captured by DRAGMM, however, we have not mentioned chinook events characterized by warm air accelerating down the terrain and that famously frequent Boulder (Markowski and Richardson 2010). Looking at the geopotential height patterns and thermodynamic fields in the feature and full domain composites to this point, it does not appear the chinook events belong to any one cluster. They likely occur in both the strong and weak jet clusters. The warmer air upstream in the weak jet cluster hints the stronger likelihood of chinook events in this cluster, but the strength of the adiabatic warming in the strong jet cluster could also be characterized as a chinook. With the weaker synoptic kinematics in the weak jet cluster, mesoscale features not captured by composites created from 12-km model data become more important to the onset of high wind storms in this cluster. Thus, the DRAGMM

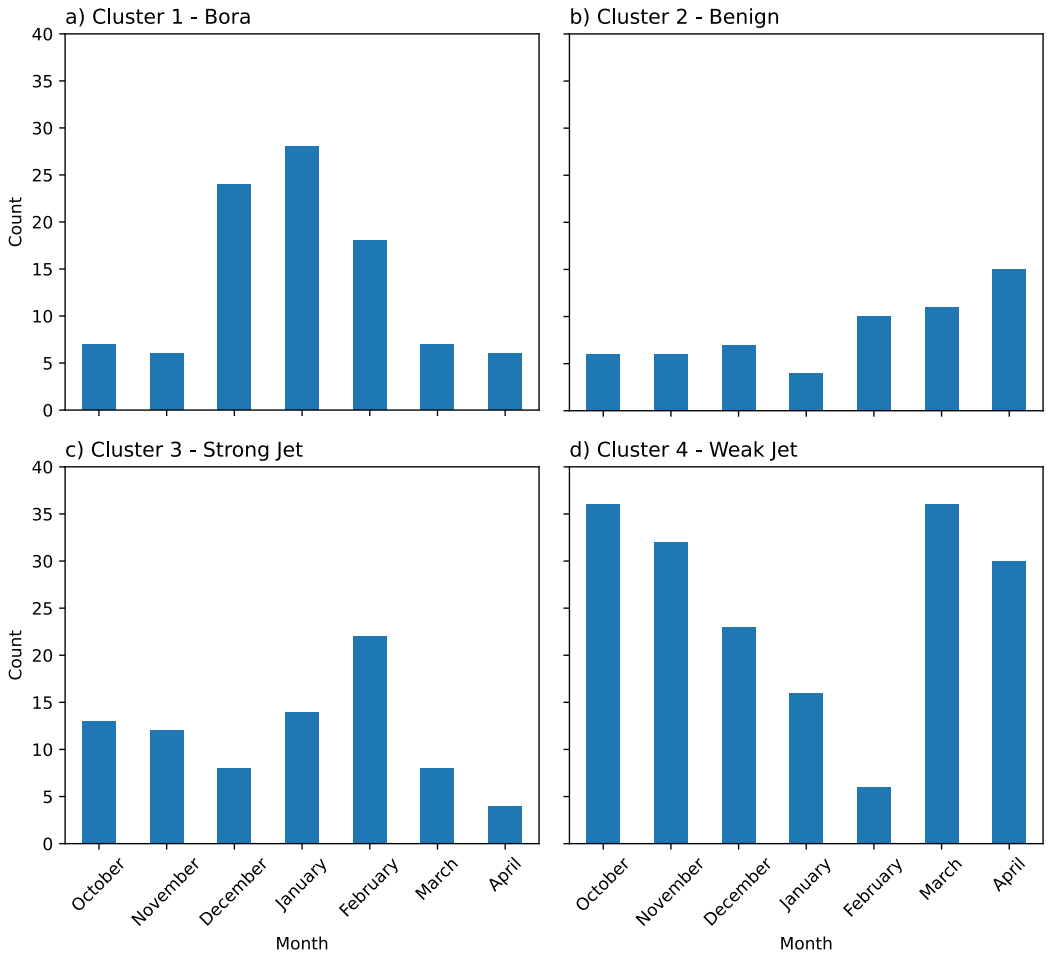

Figure 3.15: Histograms of number of days by month belonging to each cluster for Day 1 in the validation dataset.

would not distill these features to be able to cluster them. This might also indicate that the forecast RFs and CNNs may struggle more in the weak jet cluster.

The final composites of MSLP and 10-m wind speed bolster our previous observations above. In the three active clusters higher pressures are seen upstream of the ridgeline especially in the bora cluster. The averaging in these composites is not as clean as the other composites, and many of the salient MSLP and 10-m wind speed features are easier to observe in the smaller feature domain composites.

Having justified the cluster names through synoptic feature analysis of the composites, Figure 3.15 displays the distribution of days classified as each cluster by month. This provides

information on the seasonal variation of the occurrence of each cluster in the validation dataset. The higher distribution of days in the bora cluster aligns with the cold winter months as expected. The benign cluster remains low in frequency throughout the wind season as it accounts for the patterns not associated with higher winds along the Front Range. The frequency increases as the season concludes, which is evidence of the transition away from the favorable high wind synoptic patterns. We observe the opposite pattern with the strong jet frequency for the same reason. The jet overhead of the Front Range becomes less common during the spring transition season, which decreases the frequency of the strong jet cluster. The weak jet cluster is the cluster of the transition season, however. We noted broad upstream ridging in this cluster indicative of transient synoptic flow that is observed more in the fall and spring. This contrasts with the events in the bora cluster that require higher amplitude synoptic wave patterns to retrieve colder air farther northwest of the Front Range.

The Day 2 results from DRAGMM mirror the cluster characteristics from Day 1 so we adopt the same names and consider them an extension of the Day 1 clusters though they are encoded and clustered independently. The corresponding Day 2 feature and full composites are available in Appendix B. The benign cluster's overall characterization shift on Day 2 is worth noting. The baroclinic zone appears closer to Colorado in the Day 2 benign composites throughout the atmospheric layers as shown in Figure 3.16. The 300K potential temperature values (white shading in Figure 3.16c) are farther northwest on the Day 2 composite. While this specific potential temperature value is arbitrary, it does indicate that the Front Range is located in warmer air on average in the Day 2 benign cluster. Furthermore, despite the weaker mean wind speeds at 300-hPa compared to Day 1, the mean MSLP plot shows closed low pressure contours possible representing leeside cyclogenesis due to the supportive negatively tilted longwave trough with height. The main conclusion is the Day 2 benign cluster is not as “benign" as the corresponding Day 1 cluster. During the two year validation period, Boulder observed one high wind event on days classified as benign on Day 1 compared to six on days with the same classification on Day 2. Boulder is the southernmost forecast location and located closest to the tighter mean pressure gradient associated with the

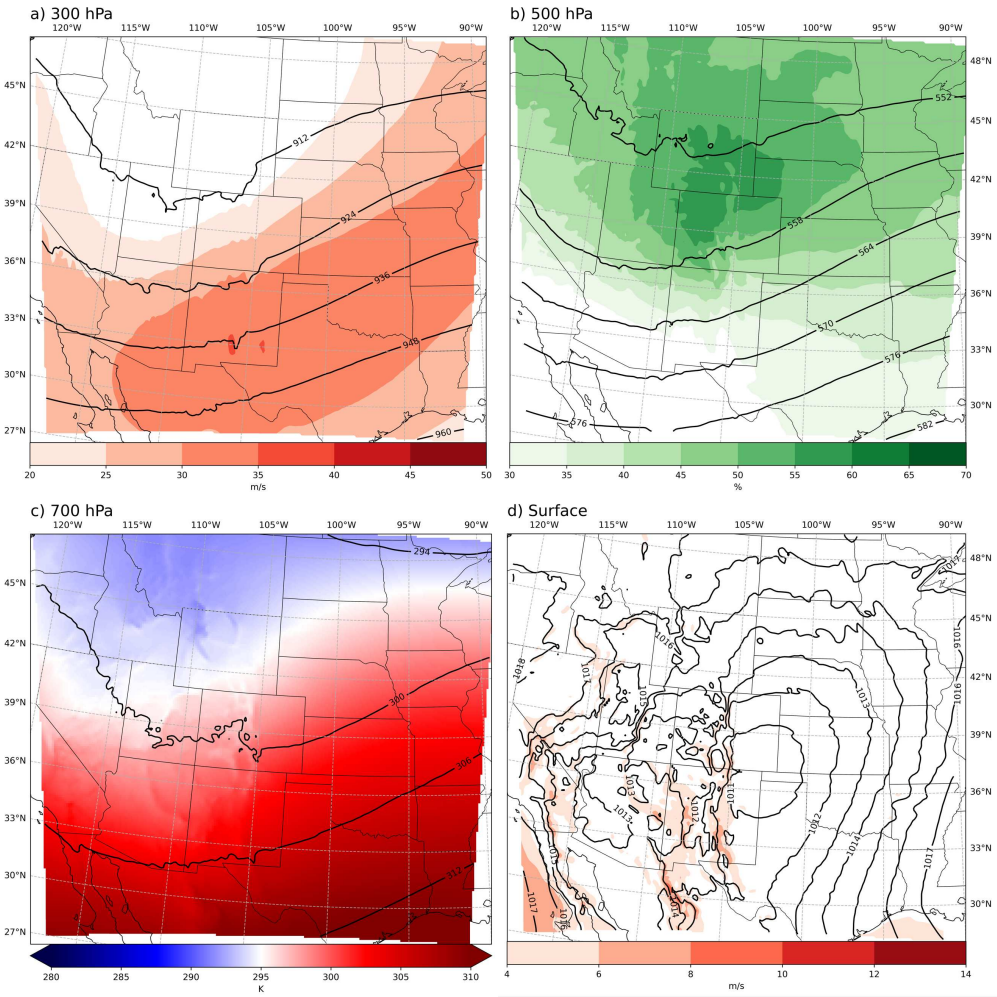

Figure 3.16: Day 2 full domain composites of the benign cluster of the (a) 300-hPa geopotential height and wind speeds, (b) 500-hPa geopotential height and relative humidity, (c) 700-hPa geopotential height and potential temperature, and (d) MSLP and 10-m wind speed. Contours and shading as in Figures 3.13 and 3.14.

low pressure depicted in the Day 2 MSLP composite. While not a significant increase in events compared to the number of events observed in the other clusters, it provides an additional insight into the benign cluster at longer lead times.

Figure 3.17 contains the histograms of number of days classified in each cluster by month for the Day 2 lead time. Once again, the overall results support the conclusions noted in the Day 1 histograms. More days in the winter months are classified as weak jet on Day 2 smoothing out the tendency for a higher number of weak jet days during the transition months noted on Day 1. The 650-hPa potential temperature fields in the Day 2 composites are cooler than Day 1 likely due to

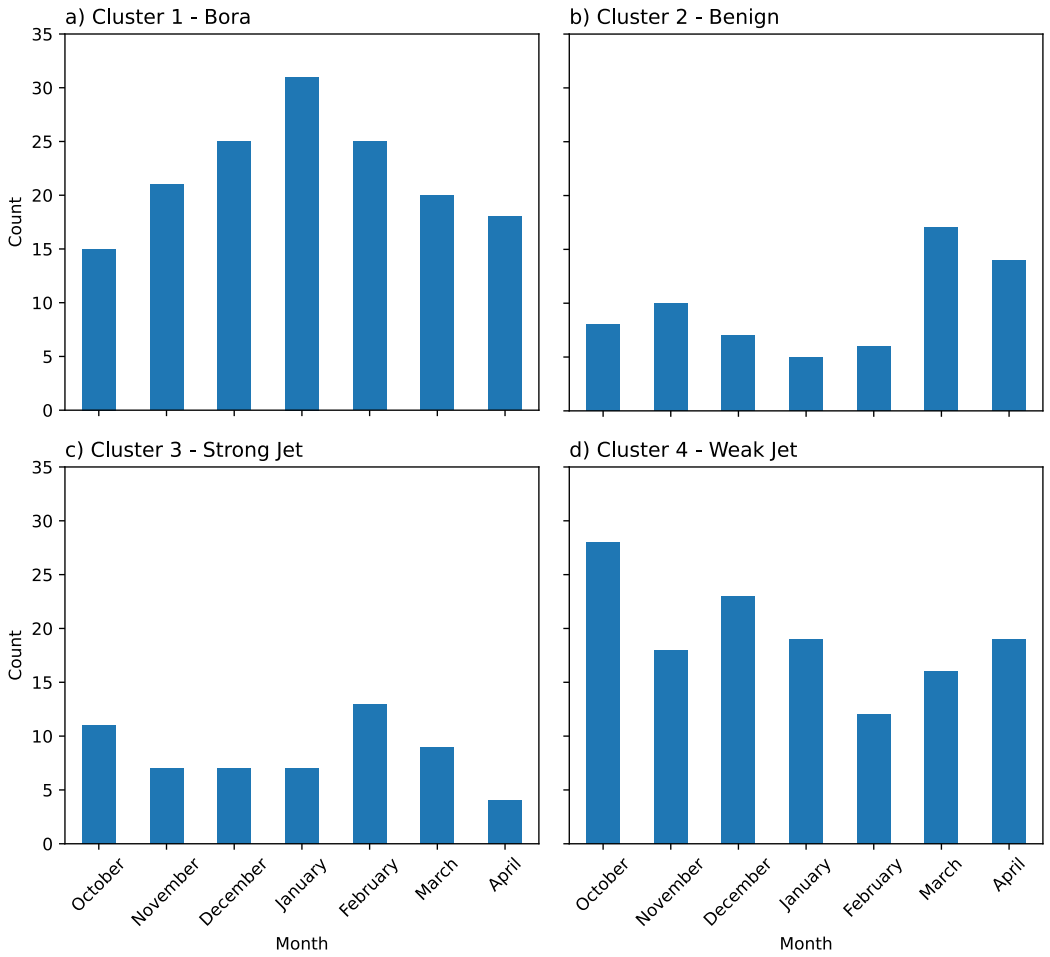

Figure 3.17: Histograms of number of days by month belonging to each cluster for Day 2 in the validation dataset.

the increase in these winter days classified as weak jet. This furthers the ambiguity of this cluster that we already know likely relies on mesoscale dynamics to initiate high wind events.

The final insights we discuss tie the DRAGMM clusters to the performance of the forecast ML models. By assessing each model's performance in each cluster, we refine the expectations of the model's behavior on a given day. A model may perform well in one cluster and poorly in another cluster. Knowing this allows a forecaster to assess confidence in a given model's forecast. It also helps the forecaster choose which models to trust for a particular forecast.

Each model's metrics for each cluster are depicted on performance diagrams in Figure 3.18. These figures plot the POD versus the success ratio defined as $1 - \text{FAR}$. Models with higher values

on both axes indicate better performance. Because the CSI is derived from the POD and FAR, every point on the diagram has a corresponding CSI value as shown in the blue shading. Additionally, these diagrams display the frequency bias, which is the ratio of positive forecasts (hits and false alarms) to verified events (hits and misses) (Lagerquist et al. 2017). Plotting the models on the same performance diagram allows comparison between the model types and changes between the Day 1 and Day 2 lead times across all locations in each cluster.

Overall, the models perform better in the strong jet and bora clusters. These clusters feature prominent synoptic signals likely represented better by the CSU-WRF and easier to distill and cluster by DRAGMM. This means a forecaster can trust the models more on average on days classified as bora or strong jet.

The metrics degrade from Day 1 to Day 2 in the bora cluster, especially the Cheyenne RFs and CNNs. Interestingly, the Cheyenne models and Fort Collins CNN increase in detection on Day 2 on strong jet days. The strong jet features on Day 2 assist these models in increasing forecast lead time although at the cost of more false alarms. The weak jet cluster as anticipated above is the difficult cluster for the models. The Day 1 Cheyenne models have the best performance on Day 1 but become unusable on Day 2 in this cluster. Contrastingly, the Boulder RF and CNN show improvements in POD and FAR, respectively, for Day 2 while their Day 1 counterparts have the worst metrics in this cluster compared to their metrics in the other clusters.

The benign cluster is an outlier due to the small number of events occurring in this cluster. For example, the Boulder models miss the one event that verified on Day 1 in this cluster, but then forecast all but one of the Day 2 events (and similarly in Cheyenne). This small sample size causes abrupt changes between the two forecast days' metrics. Finally, as with our previous analysis of the metrics in Fort Collins, the small climatology of high wind events hampers the establishment of robust trends in the metrics. Recall, the RF never forecasts high winds so those points exist at 1.0 on the x-axis in these diagrams (zero POD and zero FAR). On Day 1, five of the eight high wind events in Fort Collins occurred on strong jet days, and the CNN performs relatively well. However, four of the eight Day 2 events occurred on bora days, and the Fort Collins CNN detected none of

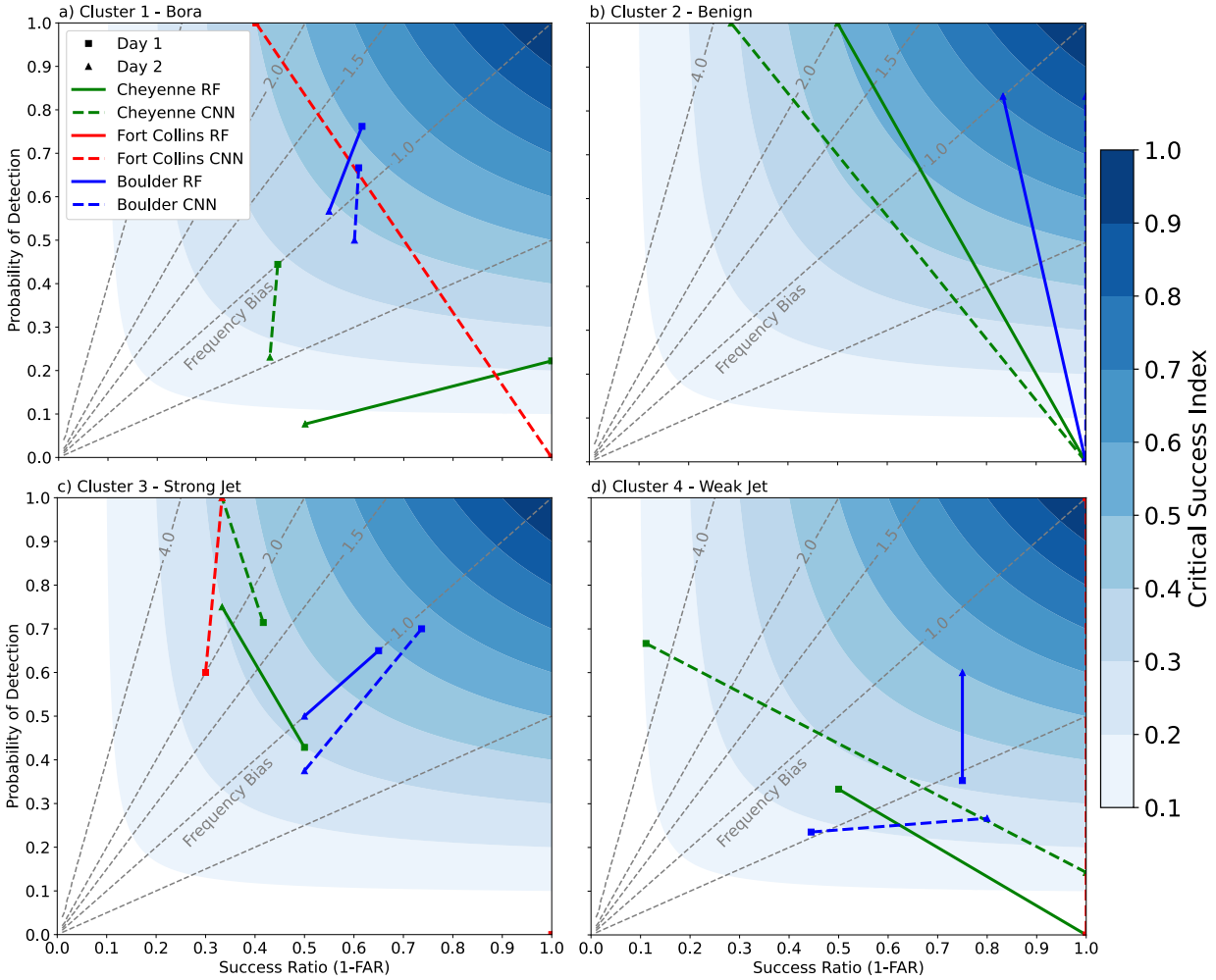

Figure 3.18: Performance diagrams for high wind events in each cluster in the validation dataset. Day 1 models are represented by a square and Day 2 models are shown with a triangle. Solid lines link a Day 1 RF to its corresponding Day 2 RF at the same location and similarly for CNNs with a dashed lines. Cheyenne, Fort Collins, and Boulder models have green, red, and blue colors, respectively. Any coordinate pair of the performance diagrams gives the CSI values (blue shading), and the gray dashed lines represent the frequency bias.

those explaining the red dashed line cutting across Figure 3.18a. The Fort Collins CNN performs better in the strong jet cluster, and two Day 2 bora high wind events that elude the detection of the Day 2 CNN become reclassified as strong jet on Day 1. This highlights the importance of noting which cluster each forecast day is classified when pairing the model performance.

In this section, we presented the DRAGMM framework to distill features in the input predictors for clustering by a GMM. We then analyzed the characteristics of these clusters, and compared how the forecast ML models from Chapter 2 performed in each cluster. The ability to identify four distinct meteorological regimes from 15 pixels of data points to the strength of this methodology. The next section presents a case study that demonstrates how to use the insights gained from this chapter in an operational forecast setting

3.4 RF vs. CNN Case Study - Boulder on 4-5 March 2020

On 5 March 2020, Boulder observed a wind gust of 26 m s^{-1} at 0538 UTC, just enough magnitude to verify high winds for the 4 March 2020 Day 1 forecast period as defined by this study. In fact, another gust of 29 m s^{-1} was observed shortly after 0600 UTC in the next day's verification window. While this event is considered one event by any realistic definition, due to the timing it counts as two high wind days in this study. This may create difficulty for the Boulder ML models as the high wind signatures are split across two forecast days' predictors. After this peak around 0600 UTC, no other high wind observations were recorded through 5 March. Figure 3.19 shows the wind observation trace for Boulder from midnight MST on 4 March to midnight MST on 5 March.

In this section, we apply the forecast ML models and the DRAGMM framework to the Day 1 forecast as if we had to issue a forecast at the beginning of this verification day. This illustrates how to combine the predictions from the ML models with the insights presented in the previous sections of this chapter. We present a simple mental model in Figure 3.20 summarizing the forecast process. First, we obtain the predictions from the ML models for the location and day of the forecast. Next, DRAGMM classifies the day into one of the clusters, and we utilize this to assess the forecast models' performances in that cluster. Finally, we leverage domain knowledge, which in this case

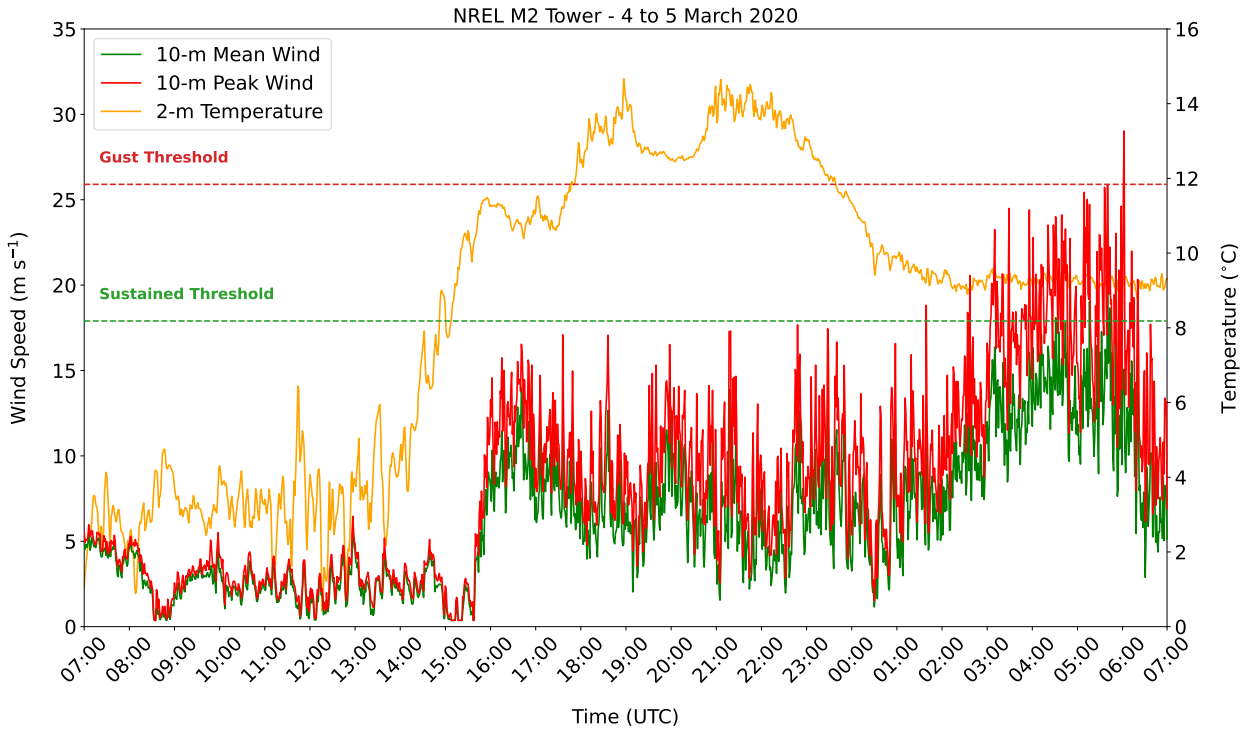

Figure 3.19: 10-m mean wind speed (m s^{-1}, red), 10-m peak wind gust (m s^{-1}, green), and 2-m temperature (°C, orange) observed at the NREL M2 Tower from 0700 UTC on 4 March 2020 through 0700 UTC on 5 March 2020. The red and green dashed horizontal lines denote the high wind gust threshold (25.9 m s^{-1}) and high wind sustained wind threshold (17.9 m s^{-1}), respectively. High winds verified just before and after 0600 UTC (2300 MST). Data from Jager and Andreas (1996).

is meteorological expertise. By combining the known synoptic features of the clusters with the real-time data available, a forecaster can assess the applicability of the insights associated with the cluster analysis. Ultimately, the forecaster decides whether to discard the forecast ML models' predictions, or choose one model over another based on this analysis.

For the Boulder Day 1 forecast on 4 March 2020, we first check the predictions from the RF and CNN models. For this day, the CNN predicted a moderate wind event while the RF forecast a high wind event. As previously discussed, the CNN does not output calibrated probabilities unlike the RF, which in this case assigned probabilities of 9.8%, 42.9%, and 47.3% for non-event, moderate winds, and high winds, respectively. The high wind class edged out the moderate class by only 4.4% in the RF, nor did the high wind class capture the majority of the decision trees. The

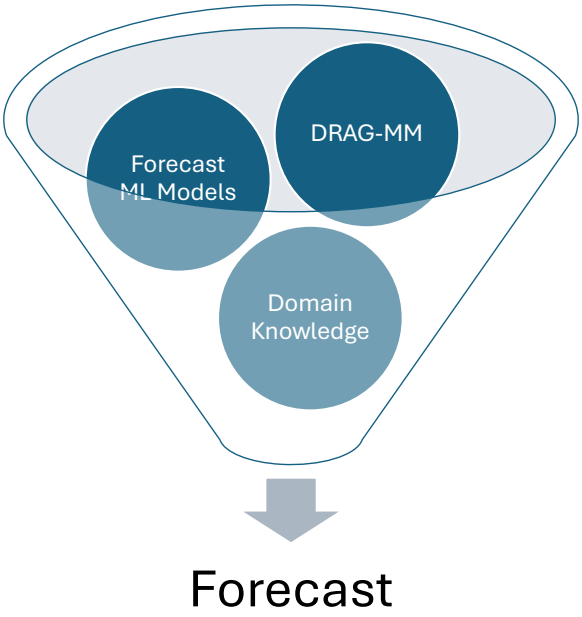

Figure 3.20: Mental model combining the forecast ML models, the DRAGG-MM framework, and meteorological knowledge and experience (domain knowledge) into the forecast process.

forecast ML models leave us with different answers and no compelling information to alleviate the uncertainty.

Next, we check the DRAGMM output for this day. 4 March was classified as a weak jet day, which are more common during the winter to spring transition. Right away, we know the ML models struggle in this cluster, and events tend to occur with weaker kinematics at the synoptic level. This furthers the uncertainty in the ML models' forecasts. The performance diagram in Figure 3.18d indicates that the RF outperforms the CNN on Day 1 in the weak jet cluster. The CNN detects less than a quarter of high wind events while also issuing false alarms more than half of the time. Though we do not need to consider the FAR in this scenario as the CNN did not forecast high winds, we should be concerned about the low POD. The RF has a better POD at 35%, which is still poor. However, the RF's FAR is 25% in this cluster, and as the RF is forecasting high winds, there is a much stronger likelihood that this high wind forecast will verify compared to if the CNN had forecast high winds. Despite the initial uncertainty, the metrics in the weak jet cluster push the forecast towards the RF.

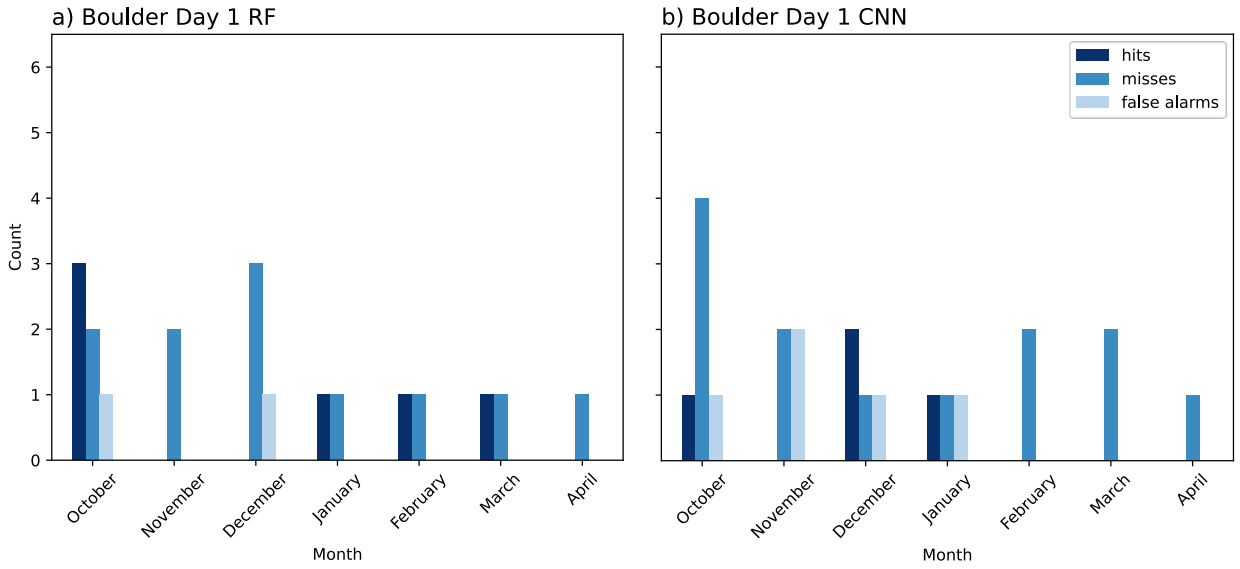

Figure 3.21: High wind event metrics by month for the Boulder Day 1 (a) RF and (b) CNN in the weak jet cluster over the validation dataset. The number of hits, misses, and false alarms are each displayed rather than as a rate.

We subdivide the performance metrics in this cluster further by month in Figure 3.21. Now we compare how the RF performs relative to the CNN in the month of March in this cluster. The RF did detect one event in March during this two-year period, but missed the other event. No false alarms were issued bolstering the confidence in the RF's high wind prediction on the case study day. The CNN, however, not only does not forecast high winds in March for this cluster (no hits or false alarms), but also does not predict high winds for February or April either. This suggests the CNN consistently underforecasts in the weak jet cluster during the late winter and spring months. This blind spot is something a forecaster must keep in mind when weighing the output of these two models.

Lastly, we look at the meteorological factors that have not been analyzed to this point. The prior paragraphs suggest the RF high wind forecast to be more trustworthy especially when weighing the impact of missing a high wind event. The DRAGMM output weak jet as the cluster for 4 March, but the probabilities of this day belonging to the various cluster direct our attention to specific synoptic and mesoscale features. The results from DRAGMM indicate a 77% match for the weak jet cluster with the other 23% showing a possible match to the strong jet cluster. Thus, we further

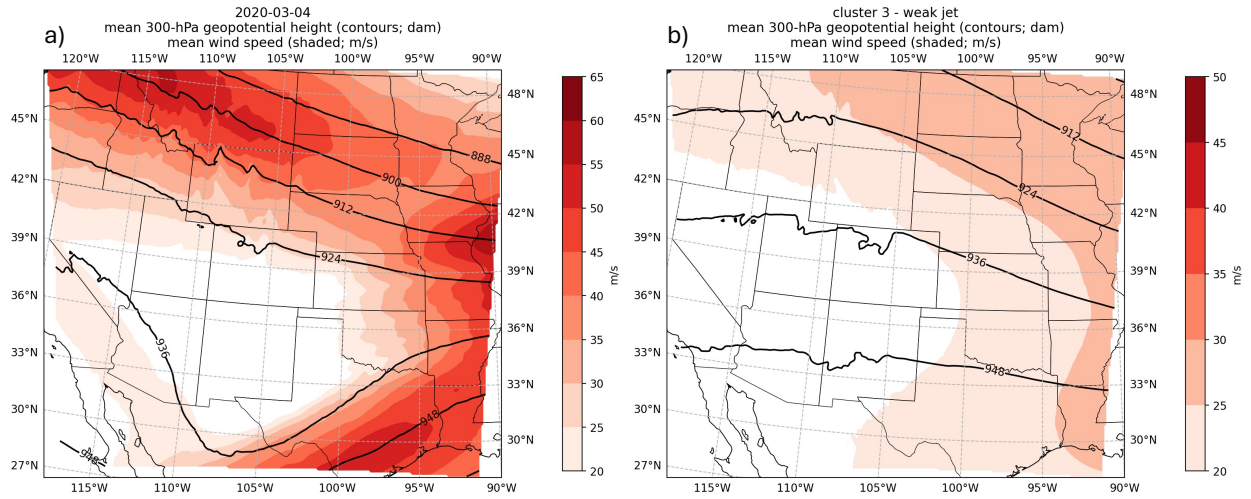

Figure 3.22: Mean 300-hPa geopotential height (dam, black contours) and wind speeds (m s^{-1}, red shading) composites for (a) the 4 March CSU-WRF across the 06, 12, 18, 24, 30-hr forecasts and (b) the weak jet cluster as in Figure 3.13d.

investigate the features of these two clusters as they compare to the CSU-WRF forecasts for 4 March.

We composite the 4 March CSU-WRF across the 6, 12, 18, 24, and 30-hr forecasts as these are the forecast hours encompassing the Day 1 predictors. Figure 3.22 displays this composite of the 300-hPa geopotential height and wind speed and compares it to the weak jet composite shown in the previous section. This validates the strong match to the weak jet cluster as the jet is not directly over the Front Range but rather two branches split to the north and south before combining over the Midwest. The case study composite has higher wind speeds, but this is expected as it in averaged over far fewer forecasts. We do not observe the transient ridging upstream in the weak jet composite. In fact, synoptic troughing is present over Colorado.

As previously discussed, smaller-scale features likely determine whether a high wind event occurs in the weak jet cluster due to the lack of strong synoptic forcing. We look near the surface at MSLP and 10-m wind speed on the feature domain scale and compare the composites seen in Figure 3.23. Additionally, we compare the 4 March and weak jet composite to the strong jet composite due to the partial match to this cluster. The forecast pressure gradient on 4 March resembles the strong jet composite more than the weak jet composite. This indicates the potential

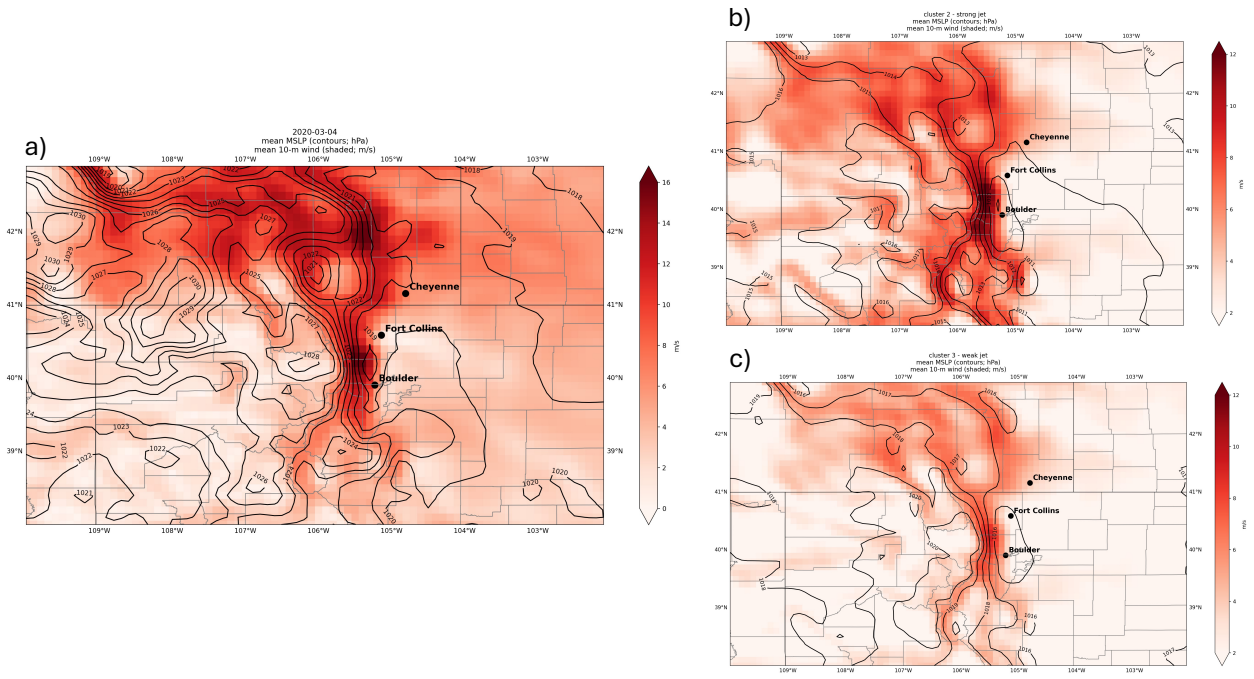

Figure 3.23: Mean MSLP (hPa, black contours) and wind speeds (m s^{-1}, red shading) composites for (a) the 4 March CSU-WRF across the 06, 12, 18, 24, 30-hr forecasts, (b) the strong jet cluster, and (c) the weak jet cluster as in Figures 3.12c-d.

existence of sufficient mesoscale forcing at least as forecast directly by the CSU-WRF. This adds confidence to the high wind forecast.

Given the ML model forecasts and performances in the weak jet cluster in March as previously discussed in conjunction with the comparison of meteorological features in the composite analysis, it is reasonable to assume this would convince a forecaster to predict high winds in line with the RF, which is the correct forecast. Of course, much more meteorological data would be available to the forecaster through observations and other traditional forecast models. We limit these data here to illustrate the amount of information that can be gleaned from this fusion of ML forecast models, the DRAGMM, and domain expertise.

3.5 Conclusion

In this chapter, we generated insights about the behavior of the Chapter 2 ML models and the meteorological regimes in which high wind events occur through direct and indirect methods. With methods applied directly to the ML models, we observed that added complexity to improve

model performance can result in models that are less explainable. Even so, feature importance analysis on the RFs showed the 700-hPa u and w winds and the 10-m u winds account for the most feature importance especially during the middle input forecast hours for each forecast day. When we aggregated the feature importances by atmospheric variable, the vertical motion at 300-hPa is highlighted in addition to the proxy for stability. The permutation importance technique also emphasizes the stability proxy as containing unique information relative to the other variables. We presented the DRAGMM as an indirect method of insight generation by distilling features within the input predictors into smaller images that are clustered. This resulted in four distinct synoptic regimes in which we noted the forecast ML models tend to perform better in clusters with stronger synoptic forcing.

By applying these insights to a real forecasting case study, we showed the utility of a forecast process that combines the ML models with the DRAGMM and the meteorological knowledge of the forecaster. In this case, these insights could steer a forecast away from an incorrect CNN forecast while increasing confidence in a correct RF forecast. DRAGMM can be run in real time, meaning that this process could be expanded to an operational forecast setting.

So far, we have tied our data-driven forecasts and insights to the 12-km CSU-WRF output. As discussed, this is chosen to test these capabilities at an input resolution on the scale of global models, which increases their geographic flexibility. In the next chapter, we explore whether CNNs can be improved by increasing the resolution of the input data from the traditional weather model.

CHAPTER 4: DOES RESOLUTION MATTER? IMPROVING FORECAST ML MODELS WITH HRRR-DERIVED PREDICTORS

4.1 Introduction

The previous chapters introduced RFs and CNNs that classify forecasts from the 12-km CSU-WRF into three wind event categories. We also used the same predictors to identify four different meteorological regimes that occur during the wind season, which aids in understanding the various physical drivers of downslope windstorms as well as the ML models' behaviors. Numerical weather prediction data with 12-km grid spacing proves coarse when forecasting these downslope windstorms, as the wave breaking features and terrain-influenced impacts occur well below that scale. Therefore, the ML models are downscaling the information from a 12-km grid to the point location of the forecast. This method takes the place of running a higher resolution model that would attempt to directly simulate the windstorm over a location, which requires additional computing resources and delays the output information to the forecaster. Our results showed some success at increasing the lead time beyond 24 hours, though plenty of room for improvement remains. Specifically, the CNNs showed the ability to detect these events better than the RFs, but at the expense of increased false alarms. Another way of stating this is the CNNs have a higher POD but are less precise. This can still be of use to forecasters by signally which days high wind events should be considered, but if the forecaster is already considering the possibility of a high wind event, a positive CNN prediction with a high FAR will not assist the decision process.

Since the dawn of numerical weather prediction, efforts have been directed at increasing model resolution in order to resolve smaller-scale features and improve the overall accuracy of the model. The initial motivation for this study included using input model data with resolution similar to global weather models. The creates flexibility as this methodology could be implemented in a new region without needing to create reforecasts of a higher resolution traditional weather model on which to train the ML models. However, as higher resolution input model data already exists for this study's forecast reason, we investigate whether increasing the resolution of the input predictors positively impacts the CNNs forecast. This potentially allows the CNNs to learn features not

resolved by the 12-km CSU-WRF forecasts, and these features may prove as better discriminators in the occurrence of high wind events. On the other hand, more data points within the same area may not provide any additional information to the CNNs resulting in similar metrics as discussed in Chapter 2. We hypothesize that better resolving mesoscale features will further enhance the detection capabilities of the CNNs especially in the weak jet cluster, but overall the extra predictor points will not provide enough information to dramatically improve the CSIs of the CNNs.

The next section provides background information on the model that supplies these new input predictors, the HRRR. Section 4.3 presents the methodology to incorporate these predictors into an ML pipeline similar to that of the CSU-WRF-driven ML models. Sections 4.4 and 4.5 contain the model performance results and a case study using the HRRR-driven CNNs before the chapter conclusion.

4.2 HRRR Background

The following summary on the capabilities and configuration of the HRRR is drawn from an article by Dowell et al. (2022) (unless otherwise cited), which provides the authoritative in-depth information about the history, development, and operations of the model. The HRRR is a convection-allowing version of the Advanced Research version of the WRF that combines higher resolution, rapid update cycles, and data assimilation to enhance the prediction of rapidly-evolving mesoscale phenomena such as severe convective threats, snow bands, downslope windstorms, and low visibility and ceiling impacts to aviation. Specifically, its 3-km grid spacing and hourly updates allow the assimilation of the most current observational data critical to capturing the evolution of these phenomena on their smaller spatial and temporal scales. The HRRR ingests weather observations types standard to most operation modeling systems in addition to three-dimensional radar data from the Multi-Radar Multi-Sensor (MRMS) project, of which the latter is key to the HRRR's success.

Relevant to the use case of predicting downslope windstorms, the vertical coordinate system, was changed in 2018 with HRRRv3 from a terrain-following sigma coordinate to a hybrid pressure-sigma vertical coordinate. The hybrid approach transitions from a terrain-following sigma

coordinate near the surface to a pure-pressure coordinate at the mid-to-upper levels. The vertical transition is weighted by a third-order polynomial that is a function of a user-specified level in the atmosphere at which the vertical coordinate becomes fully isobaric (Park et al. 2019). This inhibits small-scale numerical noise aloft when simulating flow over a mountain barrier and improves mountain-wave turbulence forecasts (Kim et al. 2019). The HRRR uses 51 vertical levels compared to the 36 levels in the CSU-WRF.

HRRRv4 became operational in late 2020 with the most significant upgrade attributed to the data assimilation system. This version implements a 36-member 3-km ensemble Kalman filter allowing for explicit depiction of convection systems within data assimilation itself. Further developments on this ensemble data assimilation system will align with efforts to develop a rapid refresh framework within the Unified Forecast System (UFS) that the National Oceanic and Atmospheric Administration (NOAA) is moving toward to enhance model development collaboration (James et al. 2022).

4.3 Methodology

The overall goal of this chapter is to replicate the development of the CNNs in Chapter 2 as much as possible so valid performance comparisons between the two sets of CNNs are achieved. The methods utilized to construct predictors, design network architecture, and verify performance mirror those described in Section 2.2 for the CSU-WRF-driven CNNs. Thus, this section focuses on specifics required for the HRRR-driven CNNs that differ from the CSU-WRF-driven CNN development.

Though the HRRR updates every hour, the input features are constructed with the HRRR initialized at 00 UTC to align with the 00Z CSU-WRF and verification window utilized previously. The same input predictors listed in Table 2.2 are extracted or derived from the 06, 12, 18, 24, and 30-hr HRRR forecasts at the resulting in the same 65 channels in the predictor cube. The 00 UTC HRRR does extend out to 48 hours (as do the 06, 12, and 18 UTC initializations), however, due to the six hour offset between the model run and the beginning of the verification window, a complete

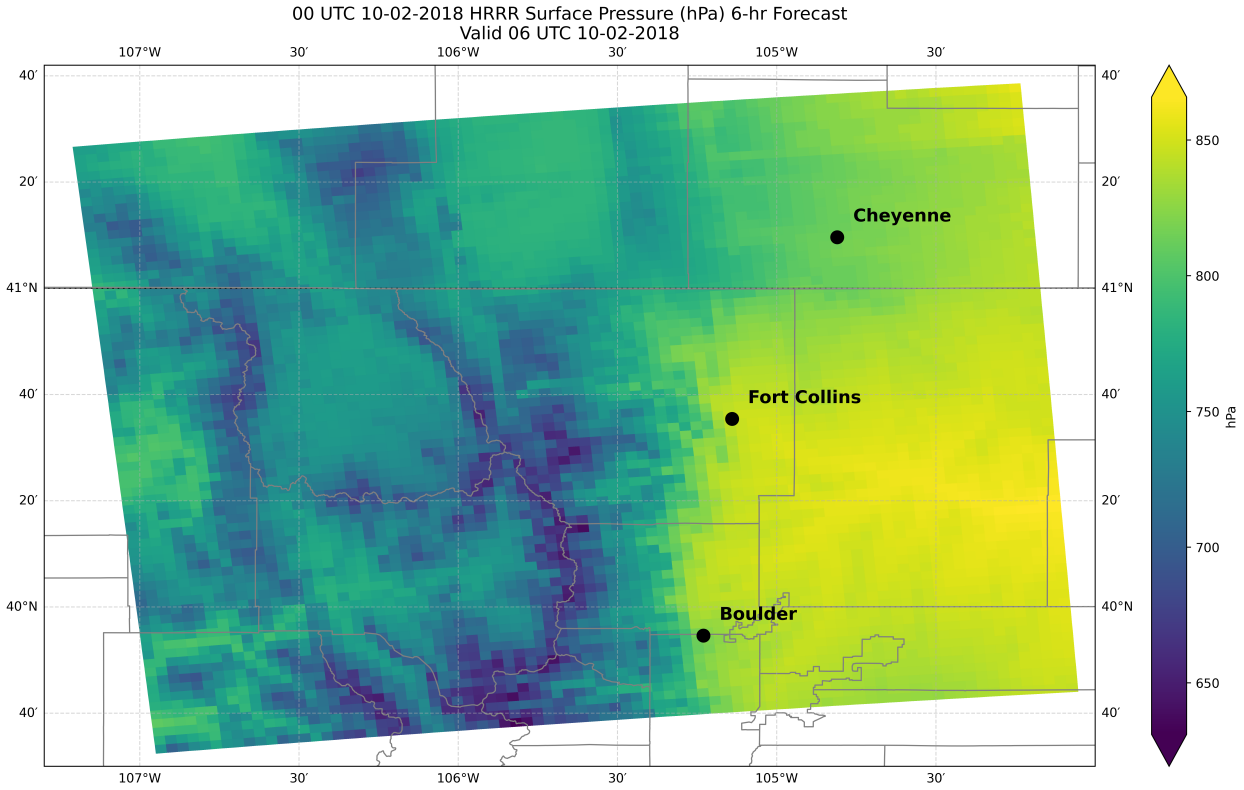

Figure 4.1: Example of a HRRR surface pressure (hPa) forecast over the predictor extraction domain. As a proxy for terrain height, surface pressure depicts the terrain within the domain.

Day 2 predictor cube requires a 54-hr forecast. To maintain integrity of the analysis of the two types of CNNs, no Day 2 HRRR-driven CNNs are developed or discussed in this study.

The domain of the input predictors differs significantly from the CSU-WRF predictors. Due to the decrease in grid spacing, using the same CSU-WRF predictor domain would increase the amount of data points in one two-dimensional feature map sixteen-fold. This would greatly increase the computing resources required and lengthen the time necessary to train the models. Therefore, we extract the predictors from the domain shown in Figure 4.1. The size of the domain is 71 by 83 grid points and focuses on areas upwind of the forecast locations inclusive of the higher terrain. The grid size is chosen in an effort to provide the new CNNs with a similar amount of data points as the previous CNNs as we cannot compare the same geographic input areas. However, to capture the north-south extend of the forecast locations and an adequate amount of terrain,

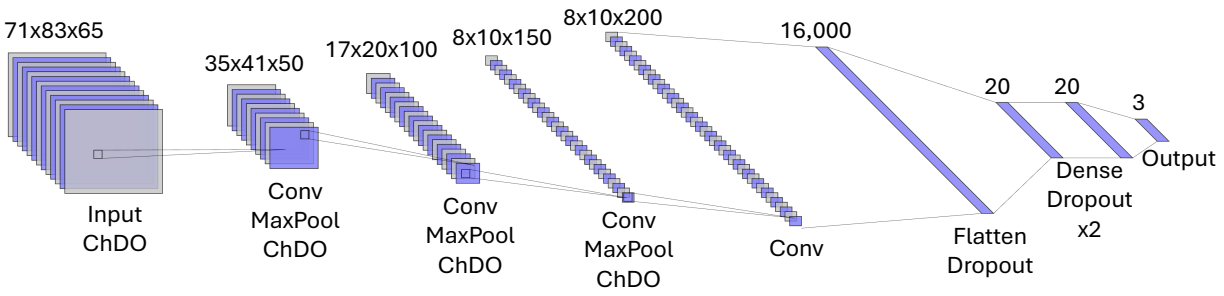

Figure 4.2: The architecture of the HRRR Boulder CNN. Notation as in Figure 2.5.

the resulting predictor cube dimensions of 71x83x65 represents a 47.3% increase in data points compared to the CSU-WRF-derived input features.

The Herbie software package is used to automatically download the HRRR data (Blaylock 2024). As depicted by Figure 4.1, the data are extracted directly from the model grid in xarray in contrast to the CSU-WRF predictors that are regridded to a latitude-longitude grid before input array construction. This simplifies the data preparation and removes an interpolation step during regridding. This preserves the resolution of the gradients and patterns contained within the 3-km data.

At least HRRRv3 is required to extract these features so the earliest wind season that can be used begins in 2018. This yields five wind seasons at the time of model training, which we utilize the seasons beginning in 2018-2020 for training and 2021-2022 for validation and hyperparameter tuning. The latter period aligns with the two-year test period for the CSU-WRF CNNs and provides the basis for subsequent comparisons. This training period is shorter than that of the CSU-WRF CNNs, which potentially offsets the advantage in the increase in predictor data.

Tuning the models indicated additional layers enhanced the forecast metrics. The overall architecture still follows the channel dropout, convolutional layer, maximum pooling layer pattern within the previous CNNs. The final HRRR CNNs required an additional iteration of the previous pattern in addition to an extra dense layer before the model output. The resulting models are deeper meaning this added complexity is necessary to fit these input data to the correct wind event classification. Figure 4.2 displays the architecture of the HRRR Boulder CNN. As before,

the other CNNs follow this same architecture with differences in hyperparameters and number of dense nodes.

The label data containing the wind classifications are recreated for each location for the training and validation periods. As an operational product from NOAA, no gaps in the model run archives exist so the previous label indices would not match, and this also maximizes the shorter training period. As before, wind events are labeled against the criteria in Table 2.1. With the models trained, we present the results and compare to the CSU-WRF CNNs and the direct HRRR forecasts.

4.4 Results

The standard contingency table metrics are calculated as before with respect to a specific wind event classification with the remaining wind event classes counting equally as non-events. Similar to the CSU-WRF direct forecast verification, we classify the HRRR output maximum 10-m wind forecast in a day against the sustained wind criteria to generate the corresponding metrics. The HRRR additionally outputs 10-m wind gust potential fields that are verified separately against the wind gust criteria creating two separate HRRR forecasts in the following metric figures. The wind gust potential is designed to capture the highest possible gust by weighting wind speeds within the planetary boundary layer and the depth of the planetary boundary layer. Thus, this value is higher than the mean observed wind gust at that time (Benjamin et al. 2023). Both the 10-m wind and 10-m wind gust potential are output for every forecast hour. The HRRR does provide sub-hourly forecasts of 5-minute average 10-m wind and 10-m wind gust potential every 15 minutes but only out to 18 hours, which only covers half of the verification window. This is unfortunate as the 5-minute average 10-m wind speed closely mimics a sustained wind observation. Instead, due to the increased resolution and higher sample frequency, the hourly output 10-m wind values used for verification likely overestimate the HRRR's daily maximum sustained wind forecast compared to the CSU-WRF 10-m wind forecast. In other words, the HRRR's hourly 10-m wind values are more representative of wind gusts as opposed to sustained winds in the model output. Nevertheless, we verify these against the sustained criteria.

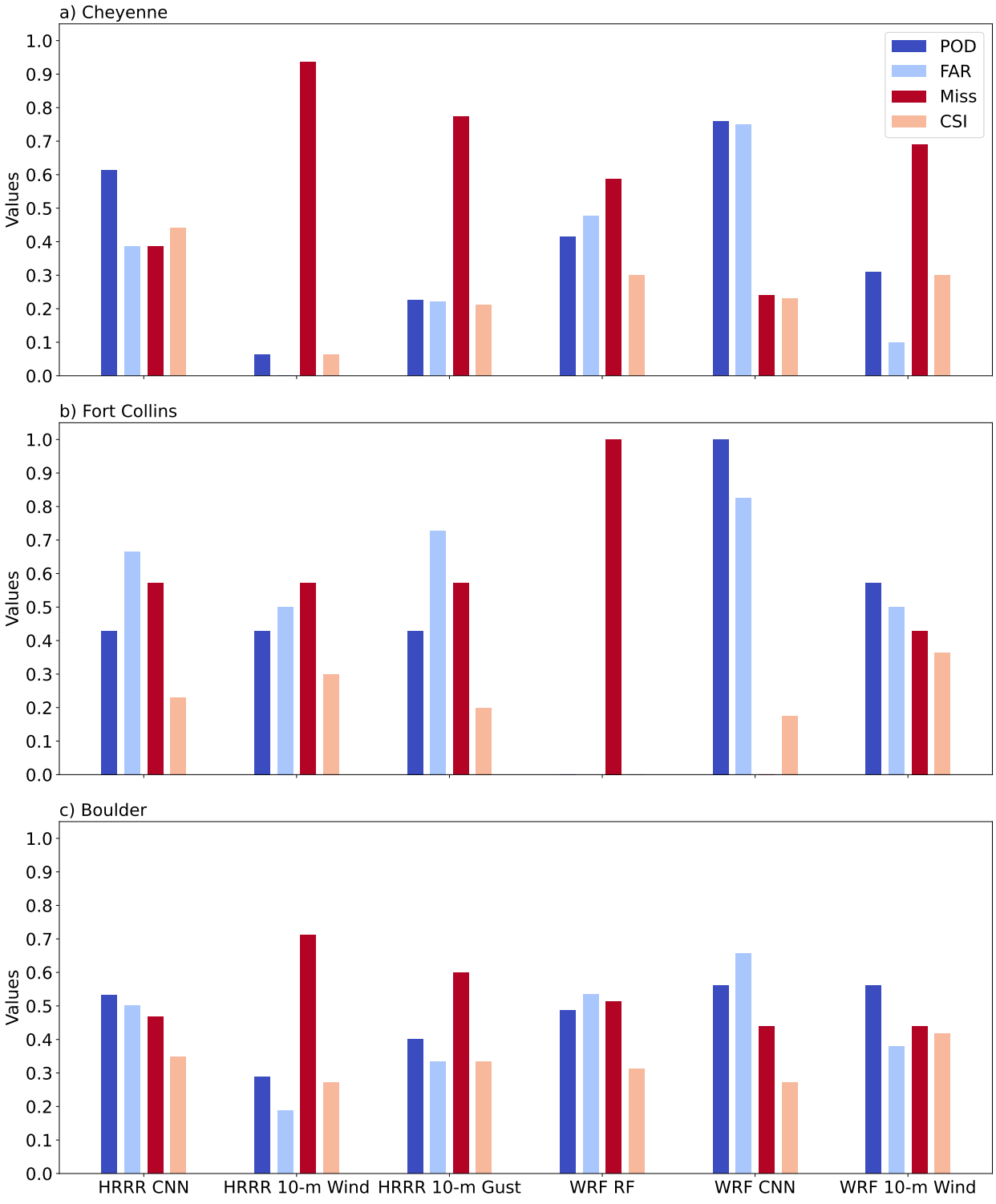

Figure 4.3: High wind event contingency table metrics for the HRRR CNNs, the HRRR direct 10-m wind forecast, the HRRR direct 10-m wind gust potential forecast, the CSU-WRF RFs, the CSU-WRF CNNs, and the direct CSU-WRF 10-m wind forecast, from left to right. Colors and forecast locations same as in Figure 2.6

Figure 4.3 presents the dichotomous high wind event metrics for the HRRR CNNs and direct forecasts in addition to the Day 1 high wind metrics from the CSU-WRF RFs, CNNs, and direct forecasts from Figure 2.6 for comparison purposes. In Cheyenne, the HRRR CNN outperforms all other models on the basis of CSI. Notably, it improves either direct HRRR forecast significantly. The Cheyenne CSU-WRF CNN does detect more events than the HRRR-driven CNN but with a much higher FAR as previously discussed. The HRRR CNN still detects over 60% of high wind events and with a 39% FAR. In most forecasting applications, FARs below 40% are considered successful especially for high impact phenomena such as downslope windstorms. While we do realize an improvement in CSI by input HRRR predictors to the CNN architecture, the greater success is the decrease in the FAR as this increases the trustworthiness of the output to the human forecaster. As a tangential observation, both direct HRRR forecasts perform worse than the direct CSU-WRF forecast, which is counterintuitive given the HRRR's higher resolution and better data assimilation capabilities.

Moving south to Fort Collins, we also note an increase in CSI compared to the CSU-WRF ML models. Once again, this is primarily due to the decrease in FAR even with the decrease in POD compared to the CSU-WRF CNN. The small sample size in Fort Collins factors in again, but these results suggest there might be middle ground between the CSU-WRF CNN high detection, high false alarm forecasts and the HRRR CNNs lower detection, lower false alarm predictions. However, the HRRR CNN does not improve the direct HRRR forecasts at least as this study verifies them. All three forecasts correctly identify three of the seven Day 1 high wind events in Fort Collins, and the 10-m wind forecast did so with a low FAR. The HRRR CNN results might be tied more to the performance of the HRRR itself in Fort Collins as opposed to the ML application itself. This contrasts with the Cheyenne results above where the CNN architecture clearly offers improvement over the direct HRRR output.

Finally, the high wind results in Boulder in Figure 4.3c show mixed signals when comparing the HRRR CNN to the direct HRRR output and the CSU-WRF metrics. The HRRR CNN outperforms the direct HRRR 10-m wind forecast and detects more events that the 10-m gust forecast but with a

higher FAR. This yields similar CSI values, but if detection remains important to this high impact forecasting problem, the HRRR CNN offers improvement over the direct HRRR output. However, the HRRR CNN performs similarly to the CSU-WRF ML models and direct forecasts. The HRRR CNN does have a lower FAR than its CSU-WRF counterpart with only a 3% decrease in POD. Additionally, the HRRR CNN does detect 5% more events than the CSU-WRF RF with 3% lower FAR. The direct CSU-WRF 10-m wind forecast outperforms the HRRR CNN in POD and more substantially in FAR with a decrease of 12%. This aligns with the fact that the direct CSU-WRF forecast has a higher CSI than both direct HRRR forecasts. Boulder has the highest climatology of high wind events so the higher sample size boosts the performance of the CSU-WRF-driven ML models, which closes the performance gap with the HRRR CNNs. Therefore, the HRRR CNN application likely is more appropriate at locations with small event sample sizes or when training data are only available over a smaller time period.

Next, we discuss the results with respect to moderate wind events at the three forecast locations. Figure 4.4 displays the same metrics as above for moderate wind events. In Cheyenne, the HRRR CNN again outperforms the other models, though, the CSU-WRF RF CSI is only 0.03 lower. With both FARs well below 40%, the higher POD means the HRRR CNN is providing a more useful forecast. As above, the HRRR CNN improves the direct HRRR forecasts at this location for the moderate wind class.

In Fort Collins, the HRRR CNN has the highest POD, but also among the highest FAR so both the direct HRRR forecasts and CSU-WRF RF perform better on the basis of CSI. The sample size in Fort Collins increases significantly for moderate wind events so again we observe that the utility of the HRRR CNN is diminished compared to ML models driven by lower resolution predictors. Less fidelity in the atmospheric fields is likely necessary for forecasting moderate winds that do not require as specific mesoscale triggers. The HRRR CNN performs better in all four metrics compared to the direct HRRR forecasts in Boulder, but only on par with the CSU-WRF ML models and direct output.

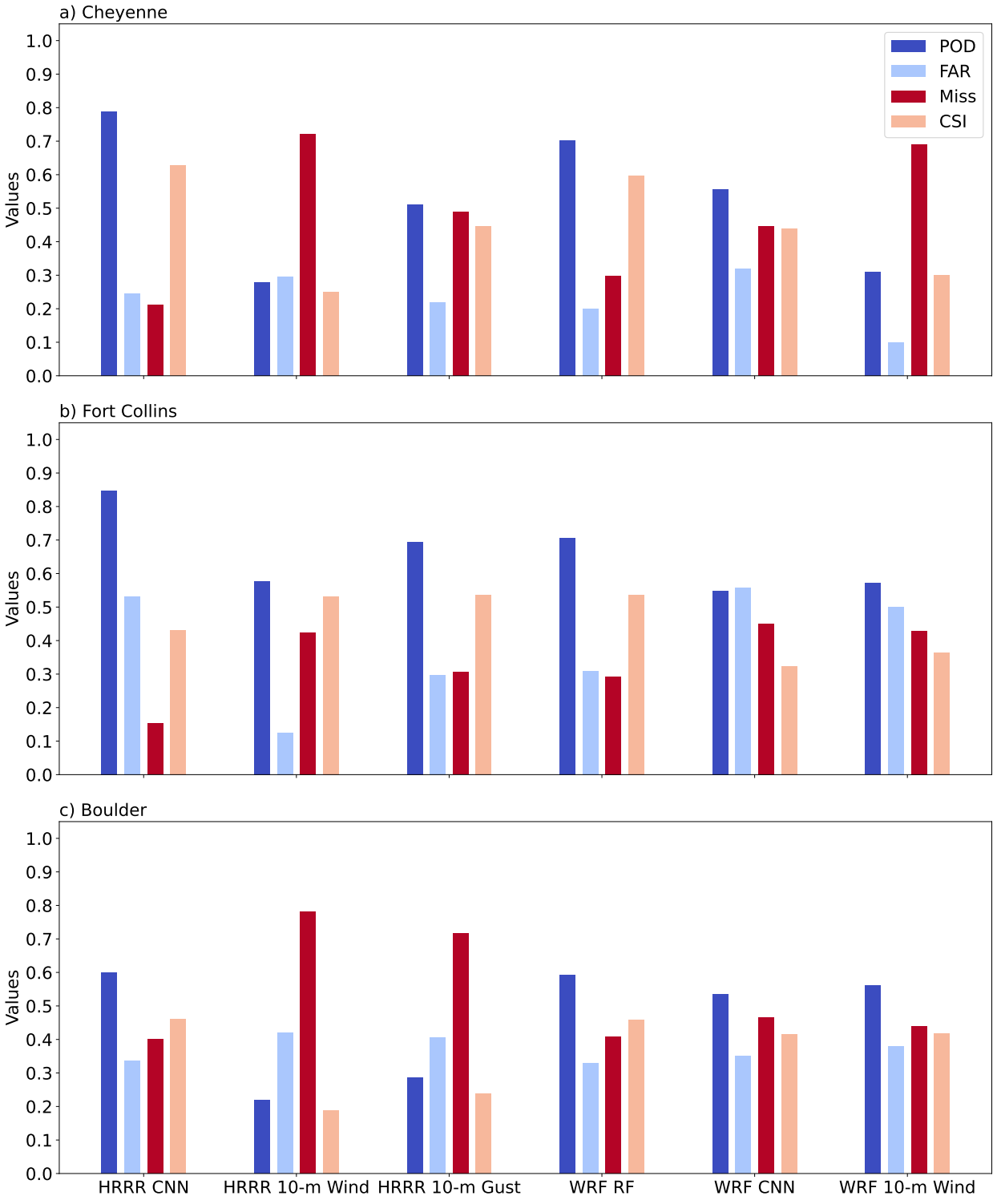

Figure 4.4: As in Figure 4.3 but for moderate wind event metrics.

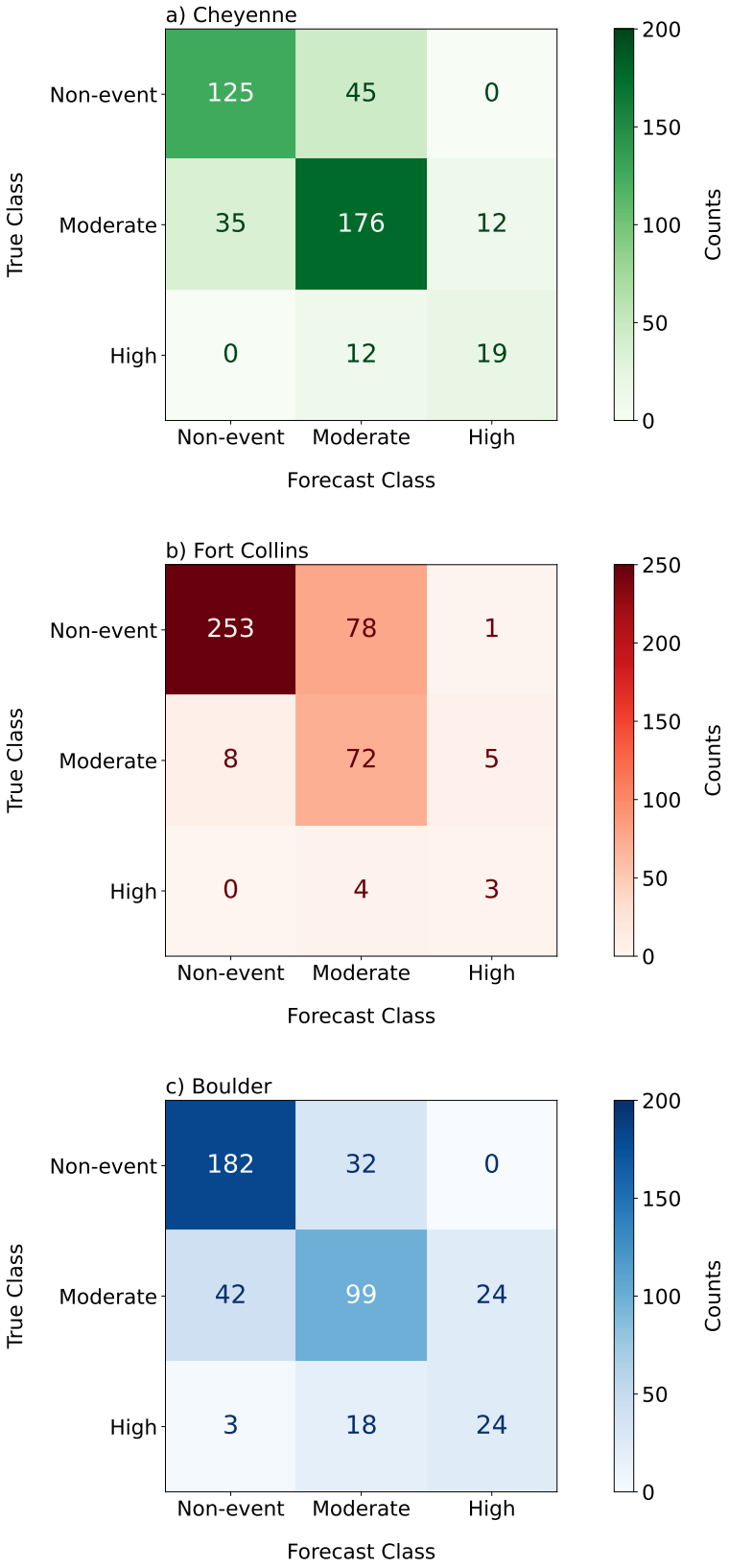

Figure 4.5: Confusion matrices as in Figure 2.10 for the HRRR CNNs.

To quantify the multicategorical performance of the HRRR CNNs, we present the confusion matrices in Figure 4.5. As previously discussed, the confusion matrices allow us to assess when the CNN is incorrect by two categories especially the impactful case when the CNN forecast no wind event but a high wind event occurs. Looking at the lower-left corner of the confusion matrices, we note that only three such cases occur all in Boulder. This only makes up 6% of the high wind events in Boulder during this two-year period. Furthermore, only one two-class false alarm occurred across the three locations in Fort Collins. This again underscores the reduction in FAR achieved by the HRRR CNNs compared to the CSU-WRF CNNs. Due to the criticisms of the FIRM metric discussed in Section 2.5, these metrics are not shown. Additionally, the dichotomous metrics clearly indicate the HRRR CNN in Cheyenne provides the most improvement in both wind event classes over the direct HRRR forecasts and the CSU-WRF-driven forecasts.

Overall, the results indicate the higher resolution predictors reduce the FAR even with a smaller geographic extraction domain and shorter training period. We do not observe an increased detection capability compared to the CSU-WRF CNNs as hypothesized. Thus, the higher resolution appears to positively impact the precision of the CNNs in this study by helping them discriminate between signatures in the input features that resemble high wind event signatures at lower resolution. Even though the detection hypothesis is incorrect, the HRRR better representing mesoscale features spatially and temporally prevents the CNNs from overforecasting more than it increases the POD. This still is valuable to the overall forecaster process especially as the FARs for the CSU-WRF CNNs are detrimental to the trustworthiness of the models.

4.5 Case Study: 2021 Marshall Fire

4.5.1 Background

On 30 December 2021, the Marshall Fire ignited in the foothills west of Boulder in dry grass and was rapidly pushed into urban areas by a powerful downslope windstorm (Fovell et al. 2022). The fire killed two people while destroying more than 1,000 structures causing more than $500 million damages (Juliano et al. 2023b). In the end, 30,000 people evacuated from an area not particularly known for wildfire risk of this magnitude, and thus, the affected population had low

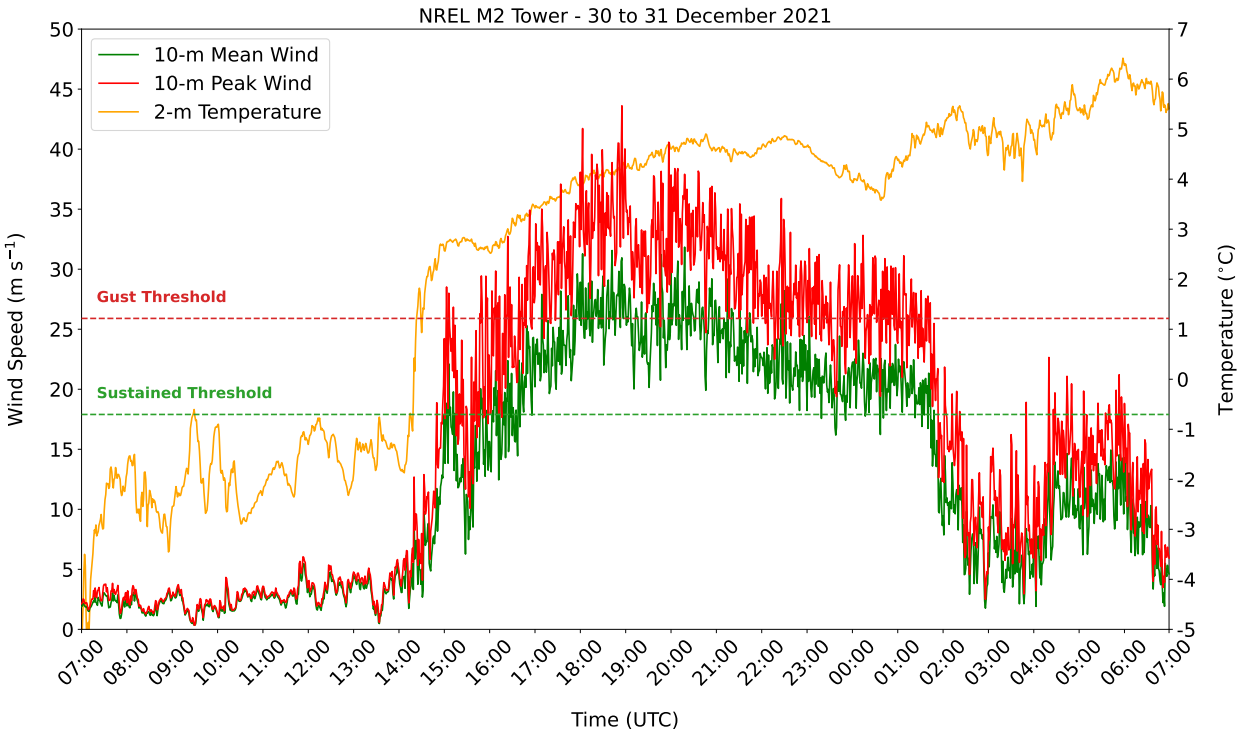

Figure 4.6: 10-m mean wind speed (m s^{-1}, red), 10-m peak wind speed (m s^{-1}, green), and 2-m temperature (°C, orange) observed at the NREL M2 Tower on the day of the Marshall Fire (0700 UTC 30 December to 0700 UTC 31 December). The red and green dashed horizontal lines denote the high wind gust threshold (25.9 m s^{-1}) and the high wind sustained wind threshold (17.9 m s^{-1}), respectively. The first, more severe phase of the windstorm began at 1500 UTC on 30 December and concluded at 0200 UTC 31 December. Data from Jager and Andreas (1996).

prior fire evacuation experience (Forrister et al. 2024). Although the Boulder foothills are famous for the occurrence of downslope windstorms, the synergy between the ambiguous meteorological situation, the dry conditions, and ignition sources made the catastrophic fire difficult to anticipate in real time.

Figure 4.6 contains the wind observations from the NREL M2 Tower, which is located about 5 km from the ignition location. Wind gusts first crossed the 25 m s^{-1} threshold at 15 UTC (08 MST) on 30 December 2021 before subsiding by 02 UTC (19 MST) on 31 December 2021. The fire ignited from two locations around 2.5 to 3 hours after the onset of the winds and just before they peaked at 19 UTC (Juliano et al. 2023b; Forrister et al. 2024). In fact, the highest wind gust observed at 51.4 m s^{-1} (115 mph) at 1906 UTC and again at 1911 UTC occurred at a station just

11 km south of the ignition source (Fovell et al. 2022). The fire spread rapidly prompting the first dissemination of an evacuation order at 1847 UTC (Forrister et al. 2024).

Despite the severity of the wind speeds observed throughout this event, poor representation of the windstorm in the operational forecast models in the hours leading up to the onset resulted in short lead time on its forecast. While the weather pattern supported the possibility of elevated winds along the Front Range, uncertainty concerning the severity and timing meant that the NWS first issued a high wind warning at 1036 UTC based on the progression of HRRR models runs initialized between 00 and 09 UTC, which was less than five hours before the beginning of the windstorm (Fovell et al. 2022). In the area forecast discussions preceding the windstorm, forecasters correctly identified the favorable high wind conditions on the morning and afternoon of 30 December with models at the time indicating gusts up to 35 m s^{-1} at the higher peaks of the Front Range. They noted the lack of a stable layer at the mountain top, which tempered expectations of a mountain wave event. The first major snow of the season was also on the horizon so this required much of the forecasters' attention as well. As the event approached, the HRRR began depicting much higher surface wind speeds to the intrigue of the forecasters as the classic mountain top stable layer was still absent. Enough subsidence aloft and a well-mixed boundary layer shown in the model allowed enough of the elevated mid-level winds to reach the surface prompting the high wind warning issuance. These forecast discussions are obtained from an retrieval service maintained by the Iowa Environmental Mesonet (2025).

Modeling studies and doppler-on-wheels observations confirm the presence of a hydraulic jump, in addition to surface stations reporting weak easterly winds suggesting the presence of a rotor (Fovell et al. 2022; Juliano et al. 2023b). Unfortunately, observed skew-T data do not exist for the Front Range or within a reasonable distance during this time period so we cannot compare the actual stratification of the atmosphere to the model predictions. However, data derived from ascending and descending aircraft into Denver International Airport did show the development of a subsidence inversion near the ridgeline as the windstorm onset approached (Benjamin et al. 2023). As pointed out by Fovell et al. (2022), a change in the placement of a mid-level horizontal

shear zone by a shift in the upstream trough weakened the mid-tropospheric winds over northern Colorado that coincided with the onset of the downslope windstorm. Subsequently, as these mid-tropospheric winds restrengthened, the windstorm abated. The misplacement of this shear zone in earlier forecast models greatly reduced the predictability of this wind event. While the focus of this study is not on specifically quantifying or improving the errors in these models, we present the progression of select aspects of the HRRR forecast as it drives our ML models.

4.5.2 HRRR Forecasts

First, we investigate a series of 0-hr HRRR analyses at 300 hPa from 18 UTC on 29 December to 00 UTC on 31 December, which are shown in Figure 4.7. The first two panels show a positively tilted longwave trough upstream of Colorado keeping the jet stream south of the Front Range. This combination of the jet location with southwest flow aloft is not conducive to downslope windstorms. However, on 06 UTC on 30 December, the jet migrates north due to the trough filling and by 12 UTC the winds become more westerly. The last two panels show the 300-hPa reflection of the mid-tropospheric horizontal shear zone over northern Colorado identified by Fovell et al. (2022). The left-entrance region of the jet also likely provide subsidence to aid the stronger midlevel winds to mix down the lee slopes of the mountains.

These analyses can be considered as observational at least in terms of the data assimilation system of the HRRR. This does not show how the HRRR forecast changed through time as the forecasters would have analyzed at every hourly cycle. Forecasters rely on run-to-run consistency to evaluate the confidence in a given model solution or the establishment of a new solution, and this was the case for NWS forecasters on this day (Benjamin et al. 2023). In fact, they cite the consistency of the HRRR cycles from 00 UTC onward as key to the issuance of the high wind warning. Thus, the following figures depict HRRR forecasts for the same point in time at different initialization times. Specifically, each figure shows the 00 UTC through the 15 UTC HRRR cycle forecasts valid at 15 UTC on 30 December at the onset of the windstorm. This includes the HRRR data fed into the HRRR CNN as well as the cycles cited by forecasters as driving the issuance of the high wind warning.

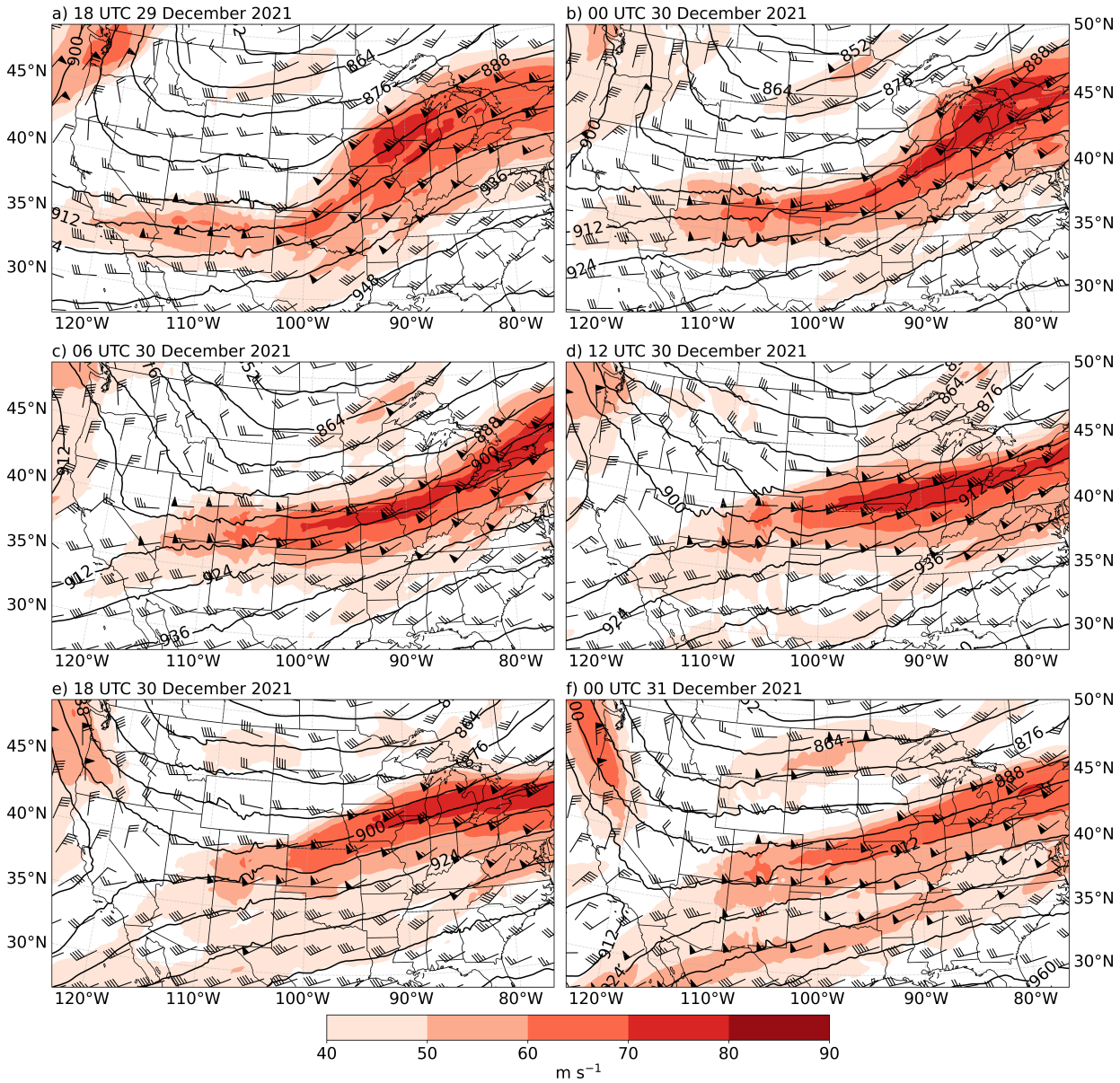

Figure 4.7: Time series of HRRR analyses of 300-hPa geopotential height (dam, contours) and winds (m s^{-1}, wind barbs and red shading) every six hours spanning 18 UTC 29 December to 00 UTC 31 December.

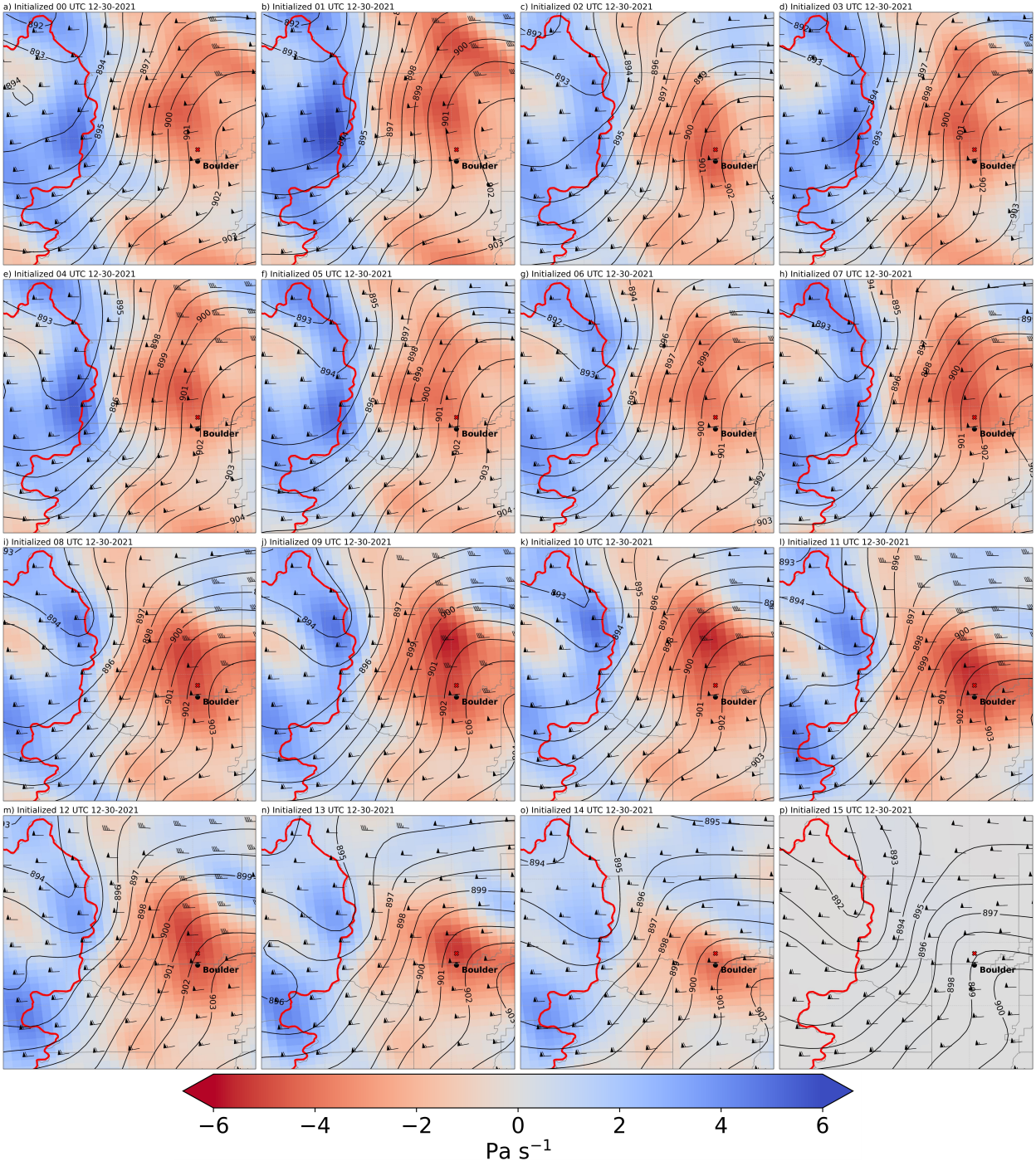

Figure 4.8: Forecasts of 300-hPa geopotential heights (dam, contours), winds (m s^{-1}, barbs), and vertical motion (Pa s^{-1}, red and blue shading) valid at 15 UTC 30 December by the 00-15 UTC HRRR cycles. The red cross denotes the ignition location of the Marshall Fire and the Boulder label represents the location of the NREL M2 Tower. The red line traces the continental divide. Vertical motion not shown in the 15 UTC cycle due to model spin-up.

Figure 4.8 displays the 300-hPa geopotential height, winds, and vertical motion forecasts for the 00-15 UTC HRRR cycles. Across all forecasts, the horizontal wind speeds are about the same and the coupling between downward vertical motion on the windward side of the mountain peaks (represented by the continental divide traced in red) and upward vertical motion over foothills. We note that the winds become more westerly in initializations approaching the beginning of the windstorm. Additionally, the upward vertical motion strengthens over Boulder in the 07 UTC cycle onward that might be a reflection of top of the hydraulic jump forming in the later forecast cycles. These two shifts indicate more favorable downslope windstorm conditions.

At 700-hPa in Figure 4.9, the horizontal winds increase in the later forecast cycles indicating the HRRR converging on the windstorm solution. At this level, the vertical motion couplet is mirrored across the continental divide compared to 300 hPa. All forecast cycles show strong areas of downward vertical motion representing the winds aloft downsloping through the 700 hPa level. The later cycles consistently show stronger magnitudes of descent. The geographic location of the downward vertical motion maximum also correctly shifts south in the later cycles closer to where the highest wind gusts were recorded near the Marshall Fire ignition point.

Near the surface, Figure 4.10 presents the MSLP, 10-m wind, and 10-m gust forecasts for the same HRRR cycles as the previous figures. The later HRRR runs clearly favor the windstorm solution with the tightening pressure gradient and wind gusts exceeding 40 m s^{-1}. These solutions also show the low pressure and easterly flow associated with the possible rotor just downstream of the highest wind speeds. The location of the wind maxima are too far north. The HRRR favors the area where the continental divide jogs west more parallel to the flow perhaps due to local terrain effects within the model. Despite the geographic displacement, comparing the 14 UTC cycle to the 00 UTC cycle shows a dramatic change in the windstorm severity at 15 UTC. After watching the 08 UTC and 09 UTC cycles complete, it is understandable why the NWS forecasters issued the high wind warning at 1036 UTC.

The final HRRR cycle time lags we present are vertical profiles of potential temperature and winds over the NREL M2 Tower for the 00-15 UTC HRRR cycle forecasts valid at 15 UTC as

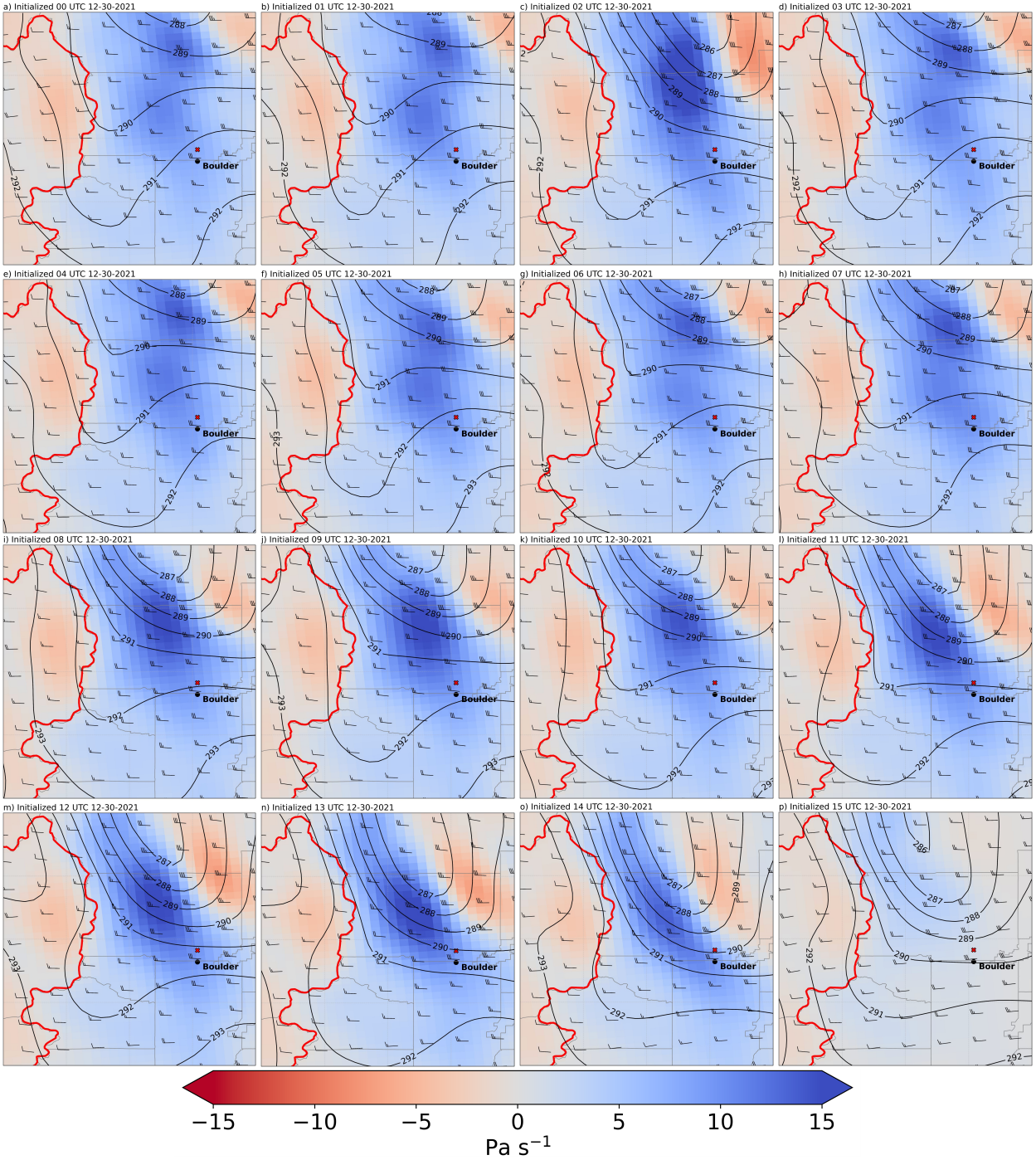

Figure 4.9: Valid time, initialization times, colors, and symbols as in Figure 4.8 for the 700-hPa HRRR forecasts.

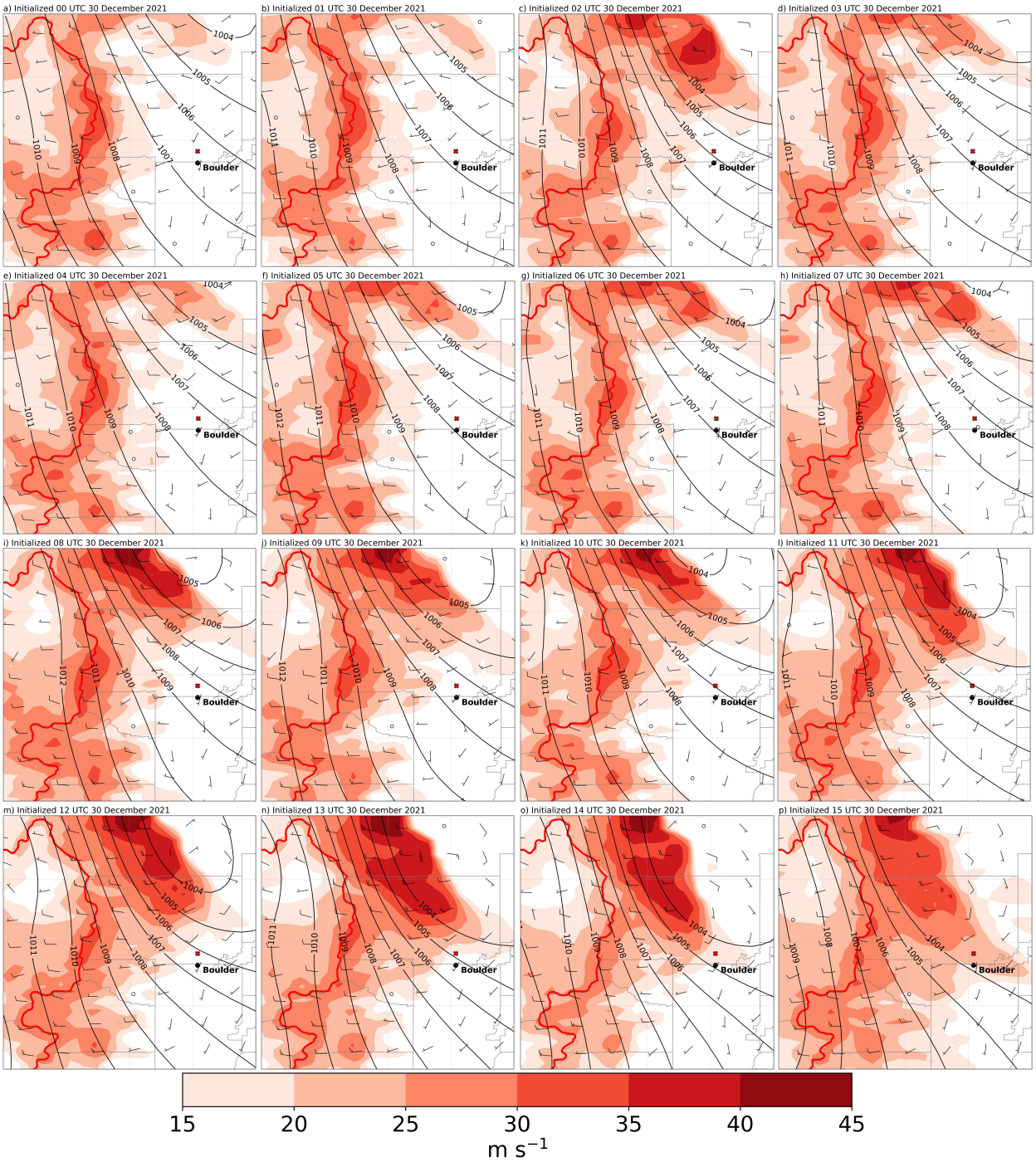

Figure 4.10: Valid time and initialization times as in Figure 4.8. Contours represent MSLP (hPa) and wind barbs depict the 10-m winds (m s^{-1}). Red shading shows the 10-m wind gust (m s^{-1}) forecasts.

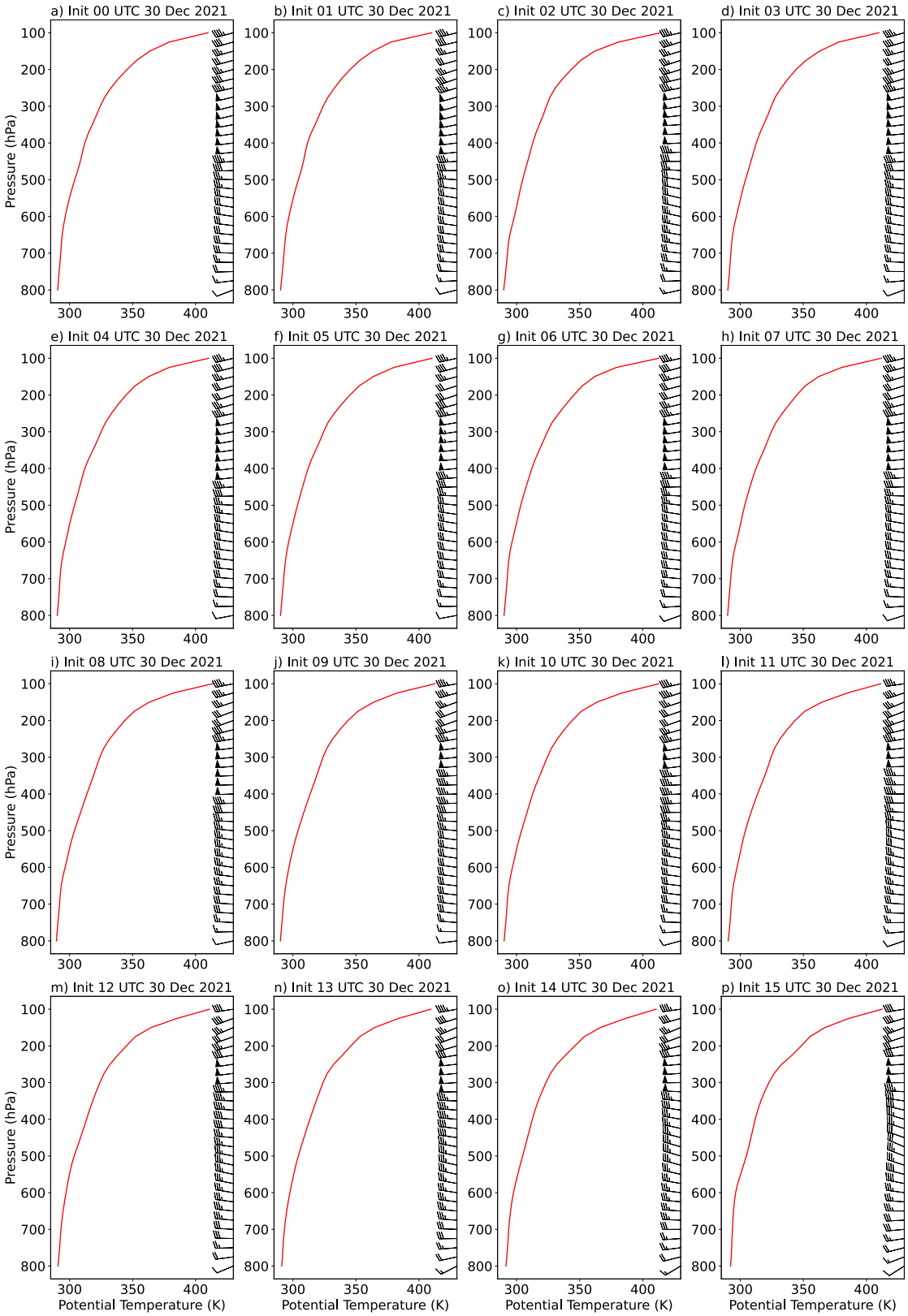

Figure 4.11: Potential temperature (K) and wind (m s^{-1}) forecasts with height over the NREL M2 Tower location for the same valid time and HRRR cycles as Figure 4.8.

before. These profiles do not depict a dramatic increase in stability in the midlevels, though, slight warming between 500 and 600 hPa is seen in the 15 UTC analysis compared to the 15-hr 00 UTC HRRR forecast. This might indicate the model is increasing stability at and above the ridgeline. We observe the weakening in the winds at 400 hPa in the later cycles that played a role in the onset of the first phase of the windstorms as previously discussed. This is the correction of the location of the horizontal shear zone in the midlevels appearing in these vertical profiles.

Throughout these forecast cycle time lags, the HRRR consistently converges on the windstorm solution as shown by the increasing low-level horizontal pressure gradient, gust forecasts, depiction of the hydraulic jump, and weakening midlevel winds. The later these features appear in the forecast solutions, however, the less lead time a forecaster has on their wind warning. The goal with the ML application is to extend the lead time of the forecast by identifying these windstorms before their explicit depiction in a physics-informed model. A survey of the evacuees from the Marshall Fire concluded that one of the factors that decreased delay time in evacuated was pre-fire risk perception (Forrister et al. 2024). Increasing the lead time of the wind warning, one of the main factors of fire risk, may have increased the fire risk assessment of the affected individuals before the fire started. It is unrealistic to expect that residents in the lee of the Front Range prepare to evacuate during every high wind warning. However, knowing the winds could potentially be dangerous would increase the perceived risk when given the evacuation order and reduce the delay in actually leaving. Counterintuitively, prior awareness of the wildfire threat in general contributed to a reduction in the likelihood of evacuation (Forrister et al. 2024).

4.5.3 ML Models Forecasts

After establishing background on the meteorological conditions and progression of the HRRR forecasts on the day of the Marshall Fire, we next focus on the ML model forecasts. The emphasis is not as much on the forecasts themselves as much as when the forecasts are available before the event onset. From the HRRR CNN methodology, there is no Day 2 HRRR CNN so only a Day 1 forecast is available. Unfortunately, the CSU-WRF ensemble did not fully complete on 30 December, so we only have forecasts from the Day 2 models on 29 December.

Both the Day 2 CSU-WRF RF and CNN forecast moderate winds for 30 December. The RF probabilities were 9%, 50%, and 41% for non-event, moderate, and high winds, respectively, indicating the reasonable possibility of high winds. In fact, within the FIRM framework with $\alpha = 0.3$, the forecast directive states that high winds should be forecast based on the 41% probability. The DRAGMM classification for Day 2 was strong jet, which is a cluster characterized by favorable dynamics for downslope windstorms for Boulder. However, forecasters certainly were aware of the possibility of high winds for 30 December on 29 December as evidenced by the NWS area forecast discussions. The problem was that traditional weather models on that day were not depicting a convincing windstorm in their outputs.

Moving forward temporally to 30 December, the Day 1 HRRR CNN correctly forecast high winds. At face value, this seems insignificant as the direct HRRR output did create a windstorm on that day. The importance of the Day 1 HRRR CNN's forecast is that it was derived from the 00 UTC HRRR cycle, and if implemented operationally, the forecast would have been available by at least 02 UTC. This is well before the more convincing HRRR cycles became available and the issuance of the high wind warning. Of course, it is impossible to know whether an additional data point from a CNN would have convinced forecasters to issue the warning sooner. The success lies in that the CNN correctly identified a high wind event in model data before the event was explicitly depicted.

The winds that drove the rapid intensification and spread of the Marshall Fire ended up being a classic downslope windstorm for the Boulder area. Slight errors in the synoptic features in the days leading up to the event resulted in model forecasts that were not convincing and decreased the lead time of the high wind forecast. The ML models in this case study show capability at converging on the high wind solution before their physics-informed counterparts and despite their predictors being derived from them.

4.6 Conclusion

In this chapter, we presented three additional CNNs trained from predictors derived from the 00 UTC HRRR cycle and compared performance across these CNNs, the CSU-WRF Day 1 ML

models, and the direct forecasts from the CSU-WRF and the HRRR. In general, the HRRR CNN outperforms the other models especially in the locations with limited event climatologies. It achieves this mainly through the reduction of false alarms despite deriving the predictors from a smaller geographic area. This suggests that input predictor resolution does matter at least within this study's framework because resolving meteorological features at the mesoscale level enables the ML to discriminate between more high versus moderate wind cases than the ML models driven by 12-km data.

We also showed that the HRRR CNN is able to recognize a high wind event before its explicit representation in the physical model in the case of the Marshall Fire. This may have enabled the increase of lead time of the high wind forecast before the ignition of the fire. The methodology chose the 00 UTC HRRR cycle to align with the CSU-WRF ML models that are also initialized at 00 UTC. Additional ML models could be trained on the 06, 12, and 18 UTC HRRR cycles that also extend to 48 hours, which would decrease the amount of time between the ML forecasts in an operational setting.

CHAPTER 5: CONCLUSION

5.1 Summary

This study describes and tests methodology to increase the predictability of downslope windstorms to mitigate their impacts with respect to direct wind damage, transportation disruptions, and wildfire intensification and spread. Four motivations drive the goals of the study: the application of ML to physics-informed weather models to improve the windstorm forecasts, the adaptability and flexibility of these techniques, their trustworthiness, and the generation of actionable insights for the forecaster. In this brief concluding chapter, we align the results from the study back to these four objectives.

The 12 forecast ML models described in Chapter 2 meet the first two objectives due to the experimental design. We utilize the 12-km CSU-WRF as the traditional weather model to derive predictors that the RFs and CNNs use to for training and inferencing. By using 12-km model data with successful forecast outcomes from the ML models, this architecture could be adapted to any geographic location as global weather model data of similar resolution are available. Furthermore, we did not specify specific geographic points to train the models, but rather supplied a larger domain of data. This allows the ML to optimize the predictor locations and would adapt to new areas with different terrain. We found that the CNNs have a better detection capability, but this comes with increased false alarms to the point where the model becomes unusable in some cases. RFs outperform CNNs at locations with sufficient event climatology by optimizing detection and false alarms. This indicates that the CNN architecture is more appropriate for locations with highly class imbalanced data or limited availability of training data. However, due to the small sample size of high wind events in Fort Collins, those ML models proved unreliable, though some use could be made of the RF probabilistic forecasts.

Transparent verification of these ML models aligns with the third goal of trustworthiness. Verification is important to recognize the strengths and weaknesses of each model and not inadvertently miscategorize one model as superior to another. Due to the different climatologies across the forecast locations, comparing models across locations is difficult. Overall, the conclusions drawn from

the dichotomous metrics are validated by the multicategorical approaches including confusion matrix and FIRM analyses. Care must be taken when choosing the risk parameters for the FIRM scoring system as changing the relative penalties associated with misses and false alarms skews the perceived model performances when no clear forecast directive exists.

The XAI techniques discussed in Chapter 3 also contribute to the trustworthiness of the models. When investigating feature and permutation importances, we noted that many of the atmospheric variables important to the ML models' predictions matched with the overall understanding of downslope windstorm development. The lone proxy predictor for atmospheric stability degraded the models' performance when its features were randomly shuffled. The RFs also prioritized geographic predictor locations near the Front Range terrain despite not receiving this information directly. Finally, the u wind predictors proved valuable to the model's forecasts as this wind component is perpendicular to the spine of the north-south Front Range.

As discussed in the motivation in Section 1.3, a component of the actionable insights goal is the real time availability of the insights to the forecast process. As a proof of concept, the forecast ML models run in real time and the predictions are available on a website (Zoellick 2025). The ML models run automatically as soon as the CSU-WRF forecast cycle is complete, and the output is displayed as a text message. Figure 5.1 shows an example of a forecast from 25 February 2025. The forecasts are archived so the user can compare the previous Day 2 forecast to the current day's Day 1 forecast for trend analysis. Although no formal solicitation of the use of these forecasts have been made, the NWS Boulder has mentioned them at least once in an area forecast discussion on 5 May 2024 as displayed in Figure 5.2 (Iowa Environmental Mesonet 2025). While these models do not run in an operational setting, this example further illustrates the availability of the forecasts produced by this framework in real time.

The DRAGMM framework was also presented in Chapter 3 that mainly contributes to the final goal of generating actionable insights. We used a convolutional AE to reduce the dimensions of our predictors into a 15-pixel encoded images, which were then clustered by a GMM. This resulted in four distinct synoptic weather regimes contained within the dataset. Three regimes, the bora,

```
Forecasts derived from CSU 12-km WRF ensemble mean initialized at 00Z on 20250225
Days are defined as 06Z - 06Z (23 MST - 23 MST)
**UPDATE** 12/9/24: Days 1 & 2 Cheyenne RF model output added!
-------------------------------------------------------------------------
DAY 1 FORECAST - Tuesday 02/25          DAY 2 FORECAST - Wednesday 02/26

** CHEYENNE (KCYS) **                  ** CHEYENNE (KCYS) **

Random Forest: MODERATE wind event      Random Forest: MODERATE wind event

Random Forest Probabilities:            Random Forest Probabilities:
Non-event: 6.88%                        Non-event: 31.71%
Moderate: 50.65%                        Moderate: 63.64%
High: 42.47%                            High: 4.65%

CNN Forecast: HIGH wind event           CNN Forecast: MODERATE wind event


** FORT COLLINS (CHRISTMAN FIELD) **    ** FORT COLLINS (CHRISTMAN FIELD) **

Random Forest: MODERATE wind event      Random Forest: NO wind event

Random Forest Probabilities:            Random Forest Probabilities:
Non-event: 20.21%                       Non-event: 54.37%
Moderate: 67.84%                        Moderate: 45.25%
High: 11.95%                            High: 0.38%

CNN Forecast: HIGH wind event           CNN Forecast: NO wind event


** BOULDER (NREL) **                    ** BOULDER (NREL) **

Random Forest: HIGH wind event          Random Forest: MODERATE wind event

Random Forest Probabilities:            Random Forest Probabilities:
Non-event: 5.49%                        Non-event: 45.02%
Moderate: 36.46%                        Moderate: 45.66%
High: 58.05%                            High: 9.32%

CNN Forecast: HIGH wind event           CNN Forecast: NO wind event
-------------------------------------------------------------------------
** WIND CATEGORIES (MPH) **

Wind <= 25 or Gust <= 35 ---------> NON-EVENT
25 < Wind <= 40 or 35 < Gust <= 58 ---------> MODERATE
Wind > 40 or Gust > 58 ---------> HIGH
-------------------------------------------------------------------------
```

Figure 5.1: Example of the text output for the ML forecast models' predictions on 25 February 2025. Text positioning is modified from the website to accommodate optimal placement within the figure.

Monday morning will also feature an increase in winds as a bora event develops. Models show good cold air advection/pressure rises behind the passing trough axis with mid-level/ridgetop flow near 50kt. Winds will quickly increase across the foothills and adjacent plains early in the day and continue through at least the evening hours. There is still some question as to the spatial extent and duration of the strongest winds, with the HRRR notably stronger across the lower foothills and the typical very windy spots near Boulder/Highway 93. Machine-learning guidance from CSU is also split close to 50/50 on if we see higher wind gusts this far east (though their criteria is slightly different than ours). Have opted for a High Wind Warning in the foothills and will message the potential for stronger gusts in the immediately adjacent plains. Further east, wind gusts will still be fairly strong (30-50 mph) but not quite at the threshold for any highlights.

Figure 5.2: The forecast ML models mentioned in a NWS Boulder area forecast discussion on 5 May 2024 highlighted in yellow.

strong jet, and weak jet account for most instances of high wind events across the three locations. The models performed better in the bora and strong jet cluster as both feature prominent synoptic features easily identified in the CSU-WRF forecasts. Forecasters should remain cautious of ML model forecasts in the weak cluster as thermodynamic and mesoscale forcing appears to dominate the onset of high wind events that are not captured well in the training process. In the case study, we demonstrated how to use the cluster identification from DRAGMM in real time along model performance within the cluster by month to avoid trusting an underforecast from a CNN and focus on a correct RF forecast. We also utilized the cluster probabilities output from DRAGMM to show that a day could potentially match different aspects of more than one cluster, which can also assist in the forecast process.

Returning to the first goal, Chapter 4 experimented with increasing the resolution of the physics-informed model to derive predictors for CNNs. The HRRR was chosen as a readily available model with 3-km grid spacing. The hypothesis of increasing detection as a result of the increased predictor resolution was incorrect, but a noticeable decrease in false alarms was observed as well. This

increases the trustworthiness of the HRRR-based models compared to some of the CSU-WRF CNNs that had false alarms of at least 80%. The HRRR CNNs achieved this with predictors derived from a smaller geographic area due to computational constraints and a shorter training period. The high resolution and improved vertical coordinate in the HRRR likely produces less mountain wave signatures from vertically propagating numeric noise off the peaks. A decrease in these false signatures may have assisted the HRRR CNNs in discriminating the true high wind events.

The Marshall Fire case study also presented in Chapter 4 illustrated the difficulty in achieving adequate lead time on a downslope windstorm when known favorable conditions exist, but errors in the traditional weather models prevent an explicit depiction of the event until hours before the onset. The HRRR did eventually forecast the high wind event, but the delay cost the forecasters lead time. The HRRR CNN correctly forecast the event from the 00 UTC HRRR cycle on 30 December, well before the windstorms explicit 15 UTC onset was shown in the model output and 10 hours before the issuance of the high wind warning. Furthermore, CSU-WRF RF probabilities from the day prior indicated the strong possibility of high winds occurring.

5.2 Future Work

Future work remains grounded in these four motivations and continues their respective lines of effort. The bulk of this study focused on deterministic forecasts despite the prevalence of stochastic tools utilized by forecasters. The main advantage of the RF architecture is probabilistic forecast out-of-the-box, but the same is not true for CNNs. This study attempted to create uncertainty quantification (UQ) for the CNNs by rerunning their predictions many times with the dropout layers active, known as Monte Carlo (MC) dropout. This did not produce enough dispersion in the predictions as the probabilities remained uncalibrated and unreliable. This is not the fault of the CNN architecture itself, but rather a limitation of MC dropout. Previous research has shown that MC dropout creates distributions that are too narrow as also demonstrated by this study's UQ efforts (Barnes et al. 2023). MC dropout addresses out-of-regime uncertainty (an example being uncertainty stemming from a limited amount of high wind cases), but not the underlying uncertainty within the data or the chaotic nature of the atmosphere. Thus, MC dropout is not an effective UQ

approach for most atmospheric science applications (Haynes et al. 2023). Future work could focus on obtaining probabilistic forecasts from CNNs, which show better detection capability, through learning and predicting a wind distribution rather than a straightforward classification task. The area forecast discussion in Figure 5.2 above illustrates the use of probabilistic information from the current ML models. Additionally, informal feedback was received on an underforecast event that was still reflected in the high wind probability. We mention this to highlight the significance of stochastic forecast information in today's operations that validates this proposed future work.

The next potential line of future work involves transfer learning where this framework is applied and evaluated over a new geographic area. This could also encompass a "lift and shift" approach that moves one of the existing ML models and its pre-trained weights to a new area and continues training it on new data. This potentially alleviates small sample size concerns with data sparse regions or relying on observations with a short historical record. The Zonda wind on the eastern slopes of the Andes mountains in Argentina were previously identified as a good candidate for this effort as previous research has attempted to classify wind events with machine learning in this location (Otero and Araneo 2021). Time constraints and emerging research opportunities presented above prevent this study from being able to test this "lift and shift" approach. This would thoroughly evaluate the goals of adaptability and flexibility.

The final area of future work encompasses on further developing the HRRR-driven ML models. The CSU-WRF RFs showed utility especially with the probabilistic information. Similar RFs could easily be trained using the HRRR predictors. A Day 2 HRRR CNN could also be trained with the missing 54-hr forecast data. This study decided not to pursue this for comparison purposes with the CSU-WRF ML models, but the potential certainly exists that a CNN could perform better than the HRRR on Day 2 with less predictors. As mentioned above, additional models could be trained on the other six-hourly HRRR cycles increasing the tempo of these ML-based forecasts. This would assist in forecasting the timing of these windstorms as well. More case studies can be looked at especially where the HRRR either correctly forecast a false alarm by a CSU-WRF ML model or vice versa to elucidate the full effects higher resolution-based predictors have on the

ML output. Training ML with higher resolution data is more expensive so the benefits must be documented before proceeding.

Overall, this study addressed its main motivations by developing new ML models based on existing numerical weather models and learned more about downslope windstorms along the Front Range by studying the predictors themselves and the performance of the ML models in different synoptic regimes. These results indicate extending the predictability of downslope windstorms up to 48 hours lead time with ML especially compared to the direct forecasts from the underlying weather models. While this technique is essentially ML-based model post-processing and was evaluated for wind event classification along the Front Range, it shows that leveraging ML can yield forecast improvements for impactful phenomena without refining equations within a weather model or inventing a sophisticated parameterization for phenomena without incomplete physical understandings.

REFERENCES

Abadi, M., and Coauthors, 2022: TensorFlow (Versions 2.9.2 and 2.16.1). Zenodo, https://doi.org/10.5281/zenodo.7604241.

American Meteorological Society, 2025: Downslope windstorm. Glossary of Meteorology. https://glossary.ametsoc.org/wiki/Downslope_windstorm.

Barnes, E. A., R. J. Barnes, and M. DeMaria, 2023: Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts. *Environ. Data Sci.*, **2**, https://doi.org/10.1017/eds.2023.7.

Benjamin, S. G., E. P. James, E. J. Szoke, P. T. Schlatter, and J. M. Brown, 2023: The 30 December 2021 Colorado Front Range windstorm and Marshall Fire: Evolution of surface and 3D structure, NWP guidance, NWS forecasts, and decision support. *Wea. Forecasting*, **38**, 2551–2573, https://doi.org/10.1175/WAF-D-23-0086.1.

Blaylock, B. K., 2024: Herbie: Retrieve numerical weather prediction model data (Version 2024.5.0). Zenodo, https://doi.org/10.5281/zenodo.11111866.

Bowman, D. M. J. S., C. A. Kolden, J. T. Abatzoglou, F. H. Johnston, G. R. van der Werf, and M. Flannigan, 2020: Vegetation fires in the anthropocene. *Nat. Rev. Earth Environ.*, **1**, 500–515, https://doi.org/10.1038/s43017-020-0085-3.

Brewer, M. J., and C. B. Clements, 2020: The 2018 Camp Fire: Meteorological analysis using in situ observations and numerical simulations. *Atmosphere*, **11**, https://doi.org/10.3390/ATMOS11010047.

Brothers, M. D., and C. L. Hammer, 2023: Random forest approach for improving nonconvective high wind forecasting across southeast Wyoming. *Wea. Forecasting*, **38**, 47–67, https://doi.org/10.1175/WAF-D-21-0215.1.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Wea. Forecasting*, **37**, 1509–1529, https://doi.org/10.1175/WAF-D-22-0070.1.

Chase, R. J., D. R. Harrison, G. M. Lackmann, and A. McGovern, 2023: A machine learning tutorial for operational meteorology. Part II: Neural networks and deep learning. *Wea. Forecasting*, **38**, 1271–1293, https://doi.org/10.1175/WAF-D-22-0187.1.

Chattopadhyay, A., E. Nabizadeh, and P. Hassanzadeh, 2020: Analog forecasting of extreme-causing weather patterns using deep learning. *J. Adv. Model. Earth Syst.*, **12**, https://doi.org/10.1029/2019MS001958.

Chollet, F., and Coauthors, 2015: Keras (Versions 2.9.0 and 3.3.3). https://keras.io.

Clark, A. J., and E. D. Loken, 2022: Machine learning–derived severe weather probabilities from a warn-on-forecast system. *Wea. Forecasting*, **37**, 1721–1740, https://doi.org/10.1175/WAF-D-22-0056.1.

Coburn, J., and S. C. Pryor, 2022: Do machine learning approaches offer skill improvement for short-term forecasting of wind gust occurrence and magnitude? *Wea. Forecasting*, **37**, 525–543, https://doi.org/10.1175/WAF-D-21.

Coen, J. L., and W. Schroeder, 2015: The High Park Fire: Coupled weather-wildland fire model simulation of a windstorm-driven wildfire in Colorado's Front Range. *J. Geophys. Res.:Atmos.*, **120**, 131–146, https://doi.org/10.1002/2014JD021993.

Czernecki, B., M. Taszarek, M. Marosz, M. Półrolniczak, L. Kolendowicz, A. Wyszogrodzki, and J. Szturc, 2019: Application of machine learning to large hail prediction - the importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmos. Res.*, **227**, 249–262, https://doi.org/10.1016/j.atmosres.2019.05.010.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Durran, D. R., 1990: Mountain waves and downslope winds. *Atmospheric Processes over Complex Terrain*, No. 45, Meteor. Monogr., Amer. Meteor. Soc., 59–81.

Flora, M. L., B. Gallo, C. K. Potvin, A. J. Clark, and K. Wilson, 2024: Exploring the usefulness of machine learning severe weather guidance in the Warn-on-Forecast System: Results from the 2022 NOAA Hazardous Weather Testbed spring forecasting experiment. *Wea. Forecasting*, **39**, 1023–1044, https://doi.org/10.1175/waf-d-24-0038.1.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the Warn-on-Forecast System. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Forrister, A., E. D. Kuligowski, Y. Sun, X. Yan, R. Lovreglio, T. J. Cova, and X. Zhao, 2024: Analyzing risk perception, evacuation decision and delay time: A case study of the 2021 Marshall Fire in Colorado. *Travel Behav. Soc.*, **35**, 100 729, https://doi.org/10.1016/J.TBS.2023.100729.

Fovell, R. G., M. J. Brewer, and R. J. Garmong, 2022: The December 2021 Marshall Fire: Predictability and gust forecasts from operational models. *Atmosphere*, **13**, https://doi.org/10.3390/atmos13050765.

Fovell, R. G., and A. Gallagher, 2018: Winds and gusts during the Thomas Fire. *Fire*, **1**, 1–22, https://doi.org/10.3390/fire1030047.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Géron, A., 2019: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd ed., O'Reilly Media, Inc.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artif. Intell. Earth Syst.*, **2**, https://doi.org/10.1175/aies-d-22-0061.1.

Herman, G. R., and R. S. Schumacher, 2016: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea. Forecasting*, **31**, 467–482, https://doi.org/10.1175/WAF-D-15-0108.1.

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

Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Iowa Environmental Mesonet, 2022: IEM computed daily summary of observations. Accessed 13 December 2022, https://mesonet.agron.iastate.edu/request/daily.phtml.

Iowa Environmental Mesonet, 2025: NWS text products by date and issuance center or PIL. Accessed 17 February 2025, https://mesonet.agron.iastate.edu/wx/afos/old.phtml.

Jager, D., and A. Andreas, 1996: M2 Tower; Boulder, Colorado (data); NREL report no. DA-5500-56489. NREL National Wind Technology Center (NWTC), accessed 17 August 2022, http://dx.doi.org/10.5439/1052222.

James, E. P., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Wea. Forecasting*, **37**, 1397–1417, https://doi.org/10.1175/WAF-D-21-0130.1.

Jiang, N., K. Luo, P. J. Beggs, K. Cheung, and Y. Scorgie, 2015: Insights into the implementation of synoptic weather-type classification using self-organizing maps: An Australian case study. *Int. J. Climatol.*, **35**, 3471–3485, https://doi.org/10.1002/JOC.4221.

Jiang, N., and Coauthors, 2017: Visualising the relationships between synoptic circulation type and air quality in Sydney, a subtropical coastal-basin environment. *Int. J. Climatol.*, **37**, 1211–1228, https://doi.org/10.1002/JOC.4770.

Juliano, T. W., F. Szasdi-Bardales, N. P. Lareau, K. Shamsaei, B. Kosovic, N. Elhami-Khorasani, E. P. James, and H. Ebrahimian, 2023a: Brief communication: The Lahaina Fire disaster: How models can be used to understand and predict wildfires. *Natural Hazards and Earth Systems Sciences Discussions*, 1–7, https://doi.org/10.5194/nhess-2023-164.

Juliano, T. W., and Coauthors, 2023b: Toward a better understanding of wildfire behavior in the wildland-urban interface: A case study of the 2021 Marshall Fire. *Geophys. Res. Lett.*, **50**, e2022GL101 557, https://doi.org/10.1029/2022GL101557.

Kim, J. H., R. D. Sharman, S. G. Benjamin, J. M. Brown, S. H. Park, and J. B. Klemp, 2019: Improvement of mountain-wave turbulence forecasts in NOAA's Rapid Refresh (RAP) model with the hybrid vertical coordinate system. *Wea. Forecasting*, **34**, 773–780, https://doi.org/10.1175/WAF-D-18-0187.1.

Klemp, J., and D. Lilly, 1975: The dynamics of wave-induced downslope winds. *J. Atmos. Sci.*, **32**, 320–339, https://doi.org/10.1175/1520-0469(1975)032<3C0320:TDOWID>3E2.0.CO;2.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444, https://doi.org/10.1038/nature14539.

Leeuwenburg, T., and Coauthors, 2024: scores: A python package for verifying and evaluating models and predictions with xarray. *J. Open Source Software*, **9**, 6889, https://doi.org/10.21105/joss.06889.

Li, B., S. Basu, and S. J. Watson, 2022: Automated identification of "Dunkelflaute" events: A convolutional neural network–based autoencoder approach. *Artif. Intell. Earth Syst.*, **1**, https://doi.org/10.1175/aies-d-22-0015.1.

Lilly, D. K., 1978: A severe downslope windstorm and aircraft turbulence event induced by a mountain wave. *J. Atmos. Sci.*, **35**, 59–77, https://doi.org/10.1175/1520-0469(1978)035%3C0059:ASDWAA%3E2.0.CO;2.

Lilly, D. K., and P. J. Kennedy, 1973: Observations of a stationary mountain wave and its associated momentum flux and energy dissipation. *J. Atmos. Sci.*, 1135–1152, https://doi.org/10.1175/1520-0469(1973)030%3C1135:OOASMW%3E2.0.CO;2.

Lindsey, D. T., and Coauthors, 2011: A high wind statistical prediction model for the northern Front Range of Colorado. *National Weather Association, Electronic Journal of Operational Meteorology*, 2011–2014.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022a: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artif. Intell. Earth Syst.*, **1**, https://doi.org/10.1175/AIES-D-22.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022b: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environ. Data Sci.*, **1**, https://doi.org/10.1017/eds.2022.7.

Markowski, P., and Y. Richardson, 2010: Mountain waves and downslope windstorms. *Mesoscale Meteorology in Midlatitudes*, John Wiley Sons, Ltd, 327–342.

Mass, C., and D. Ovens, 2024: The meteorology of the August 2023 Maui Wildfire. *Wea. Forecasting*, https://doi.org/10.1175/waf-d-23-0210.1.

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

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. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Mercer, A. E., M. B. Richman, H. B. Bluestein, and J. M. Brown, 2008: Statistical modeling of downslope windstorms in Boulder, Colorado. *Wea. Forecasting*, **23**, 1176–1194, https://doi.org/10.1175/2008WAF2007067.1.

Metz, J. J., and D. R. Durran, 2021: Are finite-amplitude effects important in non-breaking mountain waves? *Quart. J. Roy. Meteor. Soc.*, **147**, 2691–2708, https://doi.org/10.1002/QJ.4045.

Metz, J. J., and D. R. Durran, 2023: Downslope windstorm forecasting: Easier with a critical level, but still challenging for high-resolution ensembles. *Wea. Forecasting*, **38**, 1375–1390, https://doi.org/10.1175/WAF-D-22-0135.1.

Molina, M. J., and Coauthors, 2023: A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artif. Intell. Earth Syst.*, **2**, https://doi.org/10.1175/aies-d-22-0086.1.

Muñoz-Esparza, D., R. D. Sharman, and W. Deierling, 2020: Aviation turbulence forecasting at upper levels with machine learning techniques based on regression trees. *J. Appl. Meteor. Climatol.*, **59**, 1883–1899, https://doi.org/10.1175/JAMC-D-20-0116.1.

Otero, F., and D. Araneo, 2021: Zonda wind classification using machine learning algorithms. *Int. J. Climatol.*, **41**, E342–E353, https://doi.org/10.1002/joc.6688.

Park, S. H., J. B. Klemp, and J. H. Kim, 2019: Hybrid mass coordinate in WRF-ARW and its impact on upper-level turbulence forecasting. *Mon. Wea. Rev.*, **147**, 971–985, https://doi.org/10.1175/MWR-D-18-0334.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830, https://dl.acm.org/doi/10.5555/1953048.2078195.

Potvin, C. K., M. L. Flora, P. S. Skinner, A. E. Reinhart, and B. C. Matilla, 2024: Using machine learning to predict convection-allowing ensemble forecast skill: Evaluation with the NSSL Warn-on-Forecast System. *Artif. Intell. Earth Syst.*, **3**, https://doi.org/10.1175/aies-d-23-0106.1.

Prein, A. F., and L. O. Mearns, 2021: U.S. extreme precipitation weather types increased in frequency during the 20th century. *J. Geophys. Res.:Atmos.*, **126**, e2020JD034 287, https://doi.org/10.1029/2020JD034287.

Prein, A. F., E. Towler, M. Ge, D. Llewellyn, S. Baker, S. Tighi, and L. Barrett, 2022: Subseasonal predictability of North American Monsoon precipitation. *Geophys. Res. Lett.*, **49**, e2021GL095 602, https://doi.org/10.1029/2021GL095602.

Reinecke, P. A., and D. R. Durran, 2009: Initial-condition sensitivities and the predictability of downslope winds. *J. Atmos. Sci.*, **66**, 3401–3418, https://doi.org/10.1175/2009JAS3023.1.

Samek, W., T. Wiegand, and K.-R. Müller, 2017: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*, https://doi.org/10.48550/arXiv.1708.08296.

Sandmæl, T. N., and Coauthors, 2023: The tornado probability algorithm: A probabilistic machine learning tornadic circulation detection algorithm. *Wea. Forecasting*, **38**, 445–466, https://doi.org/10.1175/WAF-D-22-0123.1.

Schumacher, R. S., A. J. Hill, M. Klein, J. A. Nelson, M. J. Erickson, S. M. Trojniak, and G. R. Herman, 2021: From random forests to flood forecasts: A research to operations success story. *Bull. Amer. Meteor. Soc.*, **102**, E1742–E1755, https://doi.org/10.1175/BAMS-D-20-0186.1.

Scikit-learn Developers, 2025: Scikit-learn user guide. Accessed 31 January 2025, https://scikit-learn.org/stable/user_guide.html.

Scorer, R. S., 1949: Theory of waves in the lee of mountains. *Quart. J. Roy. Meteor. Soc.*, **75**, 41–56, https://doi.org/10.1002/qj.49707532308.

Smith, R. B., 2019: 100 years of progress on mountain meteorology research. *Meteor. Monogr.*, **59**, 1–73, https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0022.1.

Taggart, R., N. Loveday, and D. Griffiths, 2022: A scoring framework for tiered warnings and multicategorical forecasts based on fixed risk measures. *Quart. J. Roy. Meteor. Soc.*, **148**, 1389–1406, https://doi.org/10.1002/qj.4266.

Trafalis, T. B., I. Adrianto, M. B. Richman, and S. Lakshmivarahan, 2014: Machine-learning classifiers for imbalanced tornado data. *Comput. Manage. Sci.*, **11**, 403–418, https://doi.org/10.1007/s10287-013-0174-6.

United States Geological Survey, 2021: United States Geological Survey 3D Elevation Program 1 arc-second digital elevation model. OpenTopography, accessed 25 February 2025, https://doi.org/10.5069/G98K778D.

Wilks, D. S., 2019: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, 4th ed., Elsevier, 369–483, https://doi.org/10.1016/B978-0-12-815823-4.00009-2.

Zoellick, C. L., 2025: CSU-WRF machine learning wind forecasts. Accessed 25 February 2025, https://schumacher.atmos.colostate.edu/zoellick/current_fcst.txt.

APPENDIX A: RF FEATURE IMPORTANCE MAPS

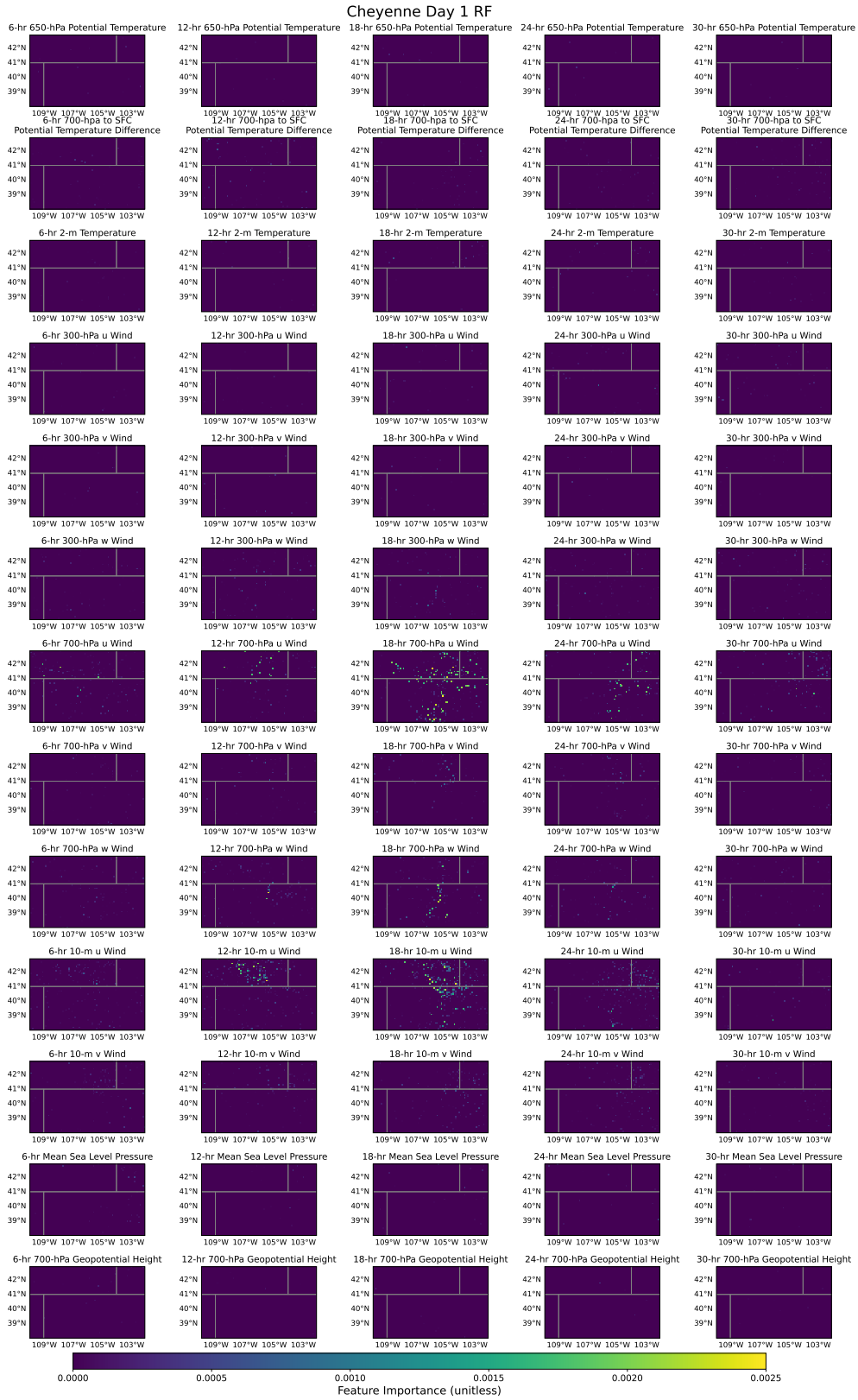

Figure A.1: Cheyenne Day 1 RF feature importances for each variable and forecast time comprising the input predictors.

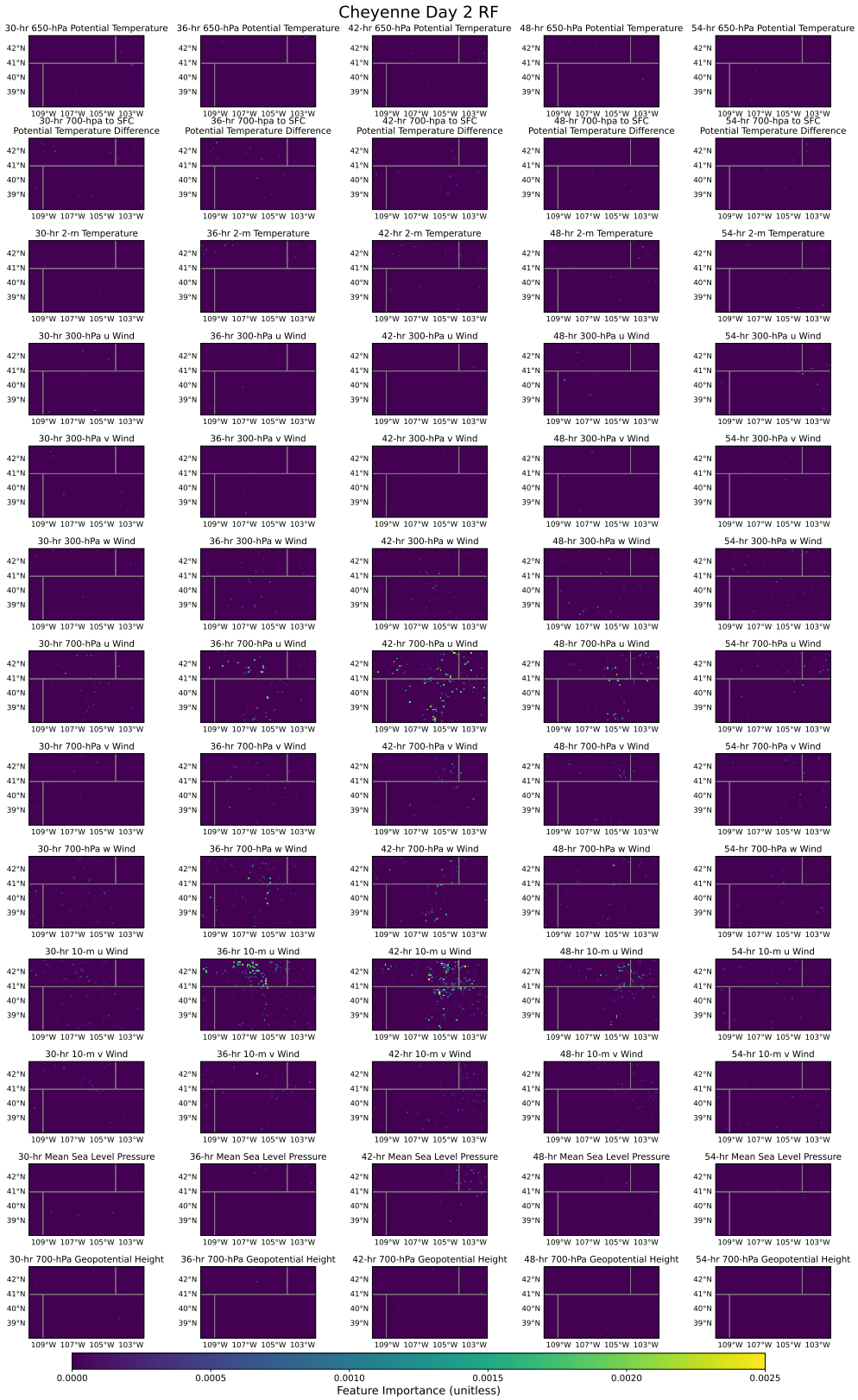

Figure A.2: Cheyenne Day 2 RF feature importances for each variable and forecast time comprising the input predictors.

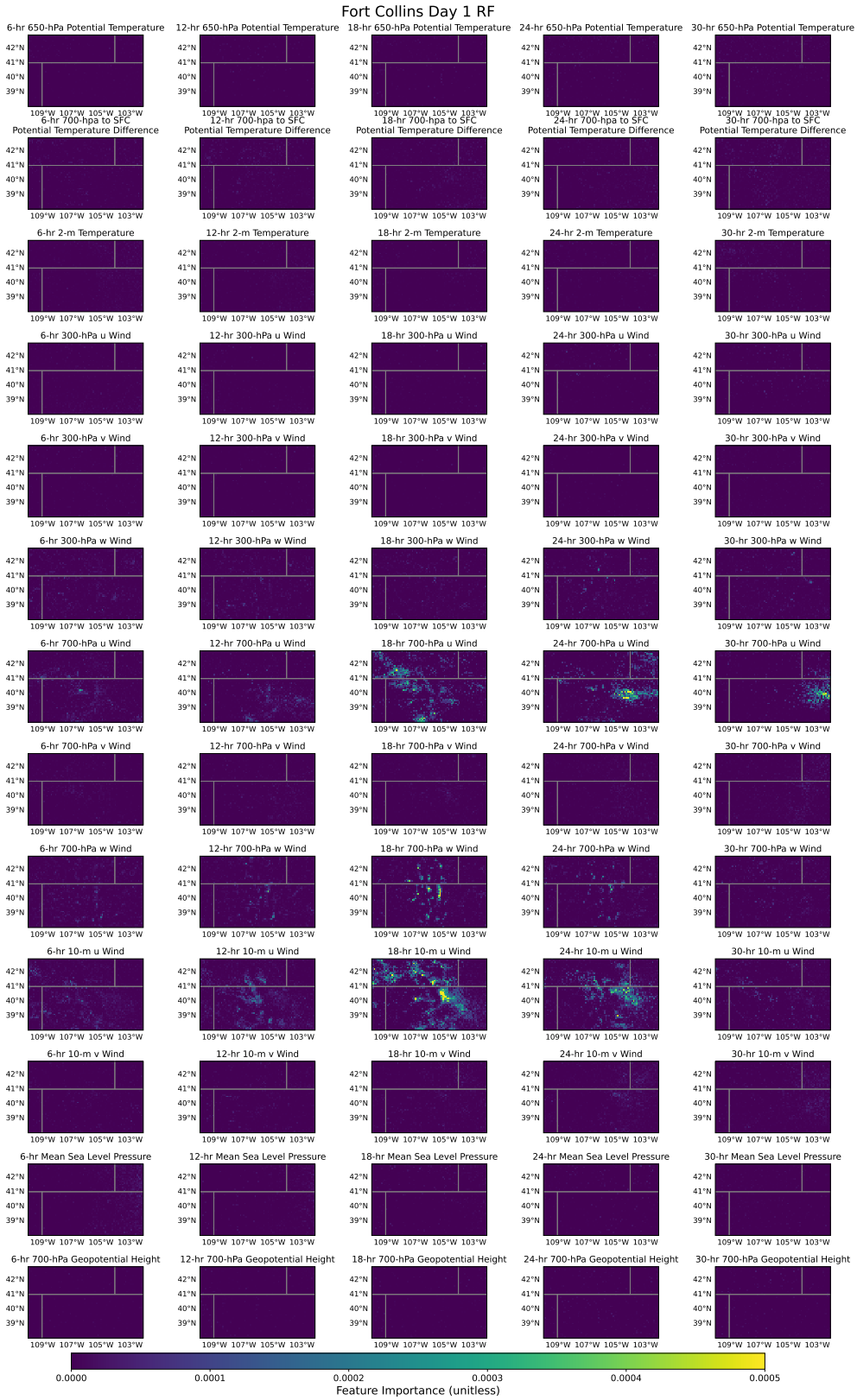

Figure A.3: Fort Collins Day 1 RF feature importances for each variable and forecast time comprising the input predictors.

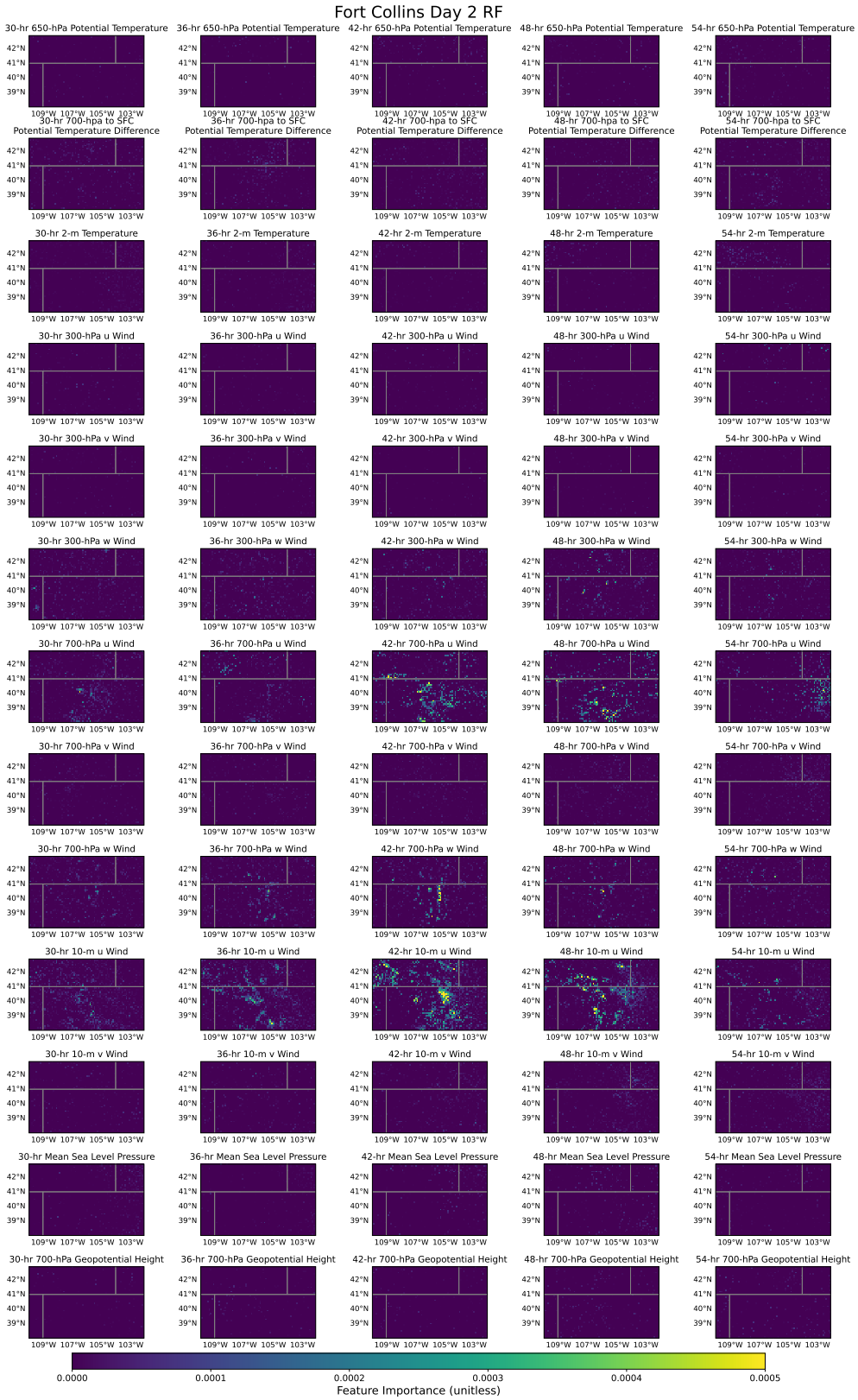

Figure A.4: Fort Collins Day 2 RF feature importances for each variable and forecast time comprising the input predictors.

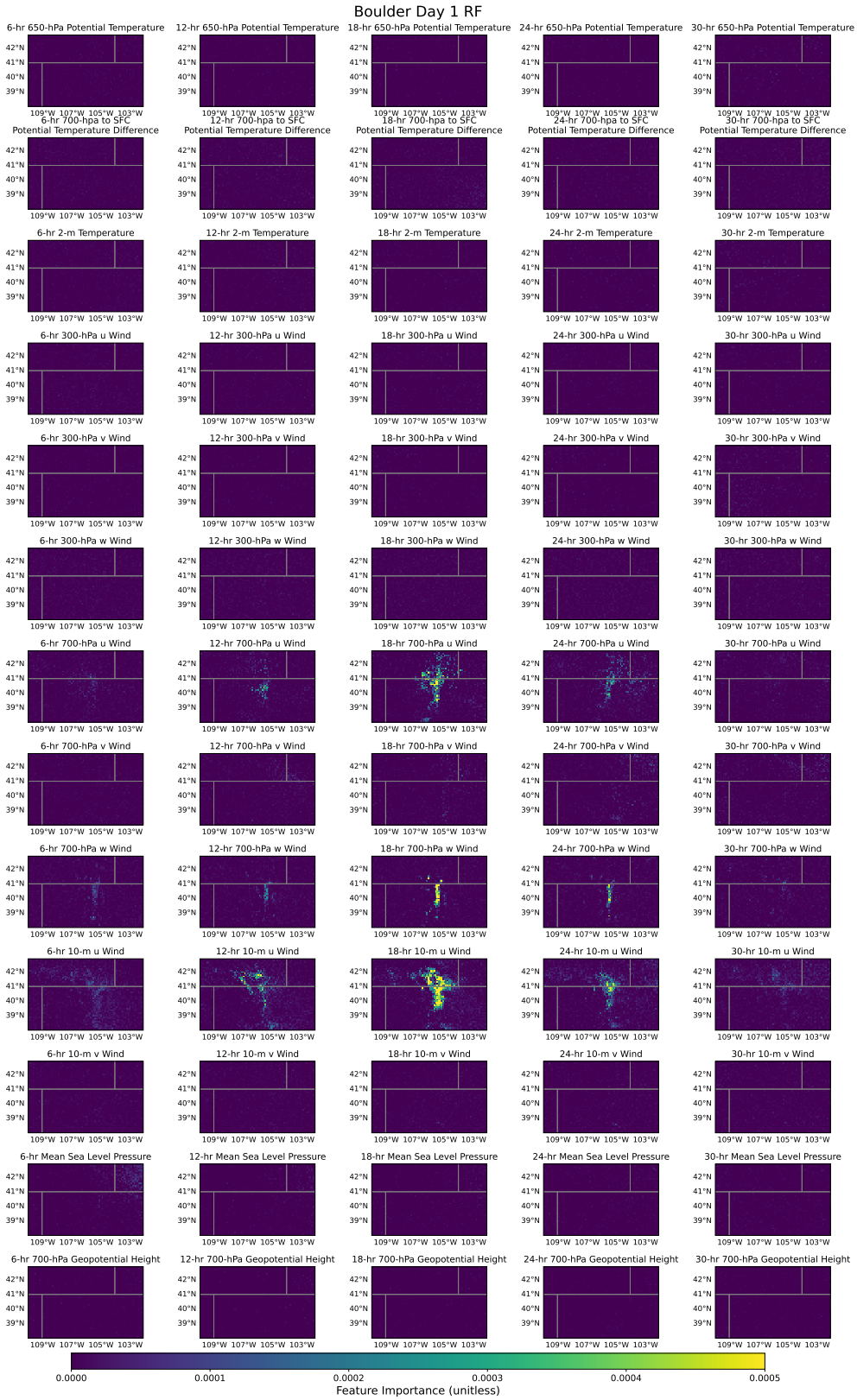

Figure A.5: Boulder Day 1 RF feature importances for each variable and forecast time comprising the input predictors.

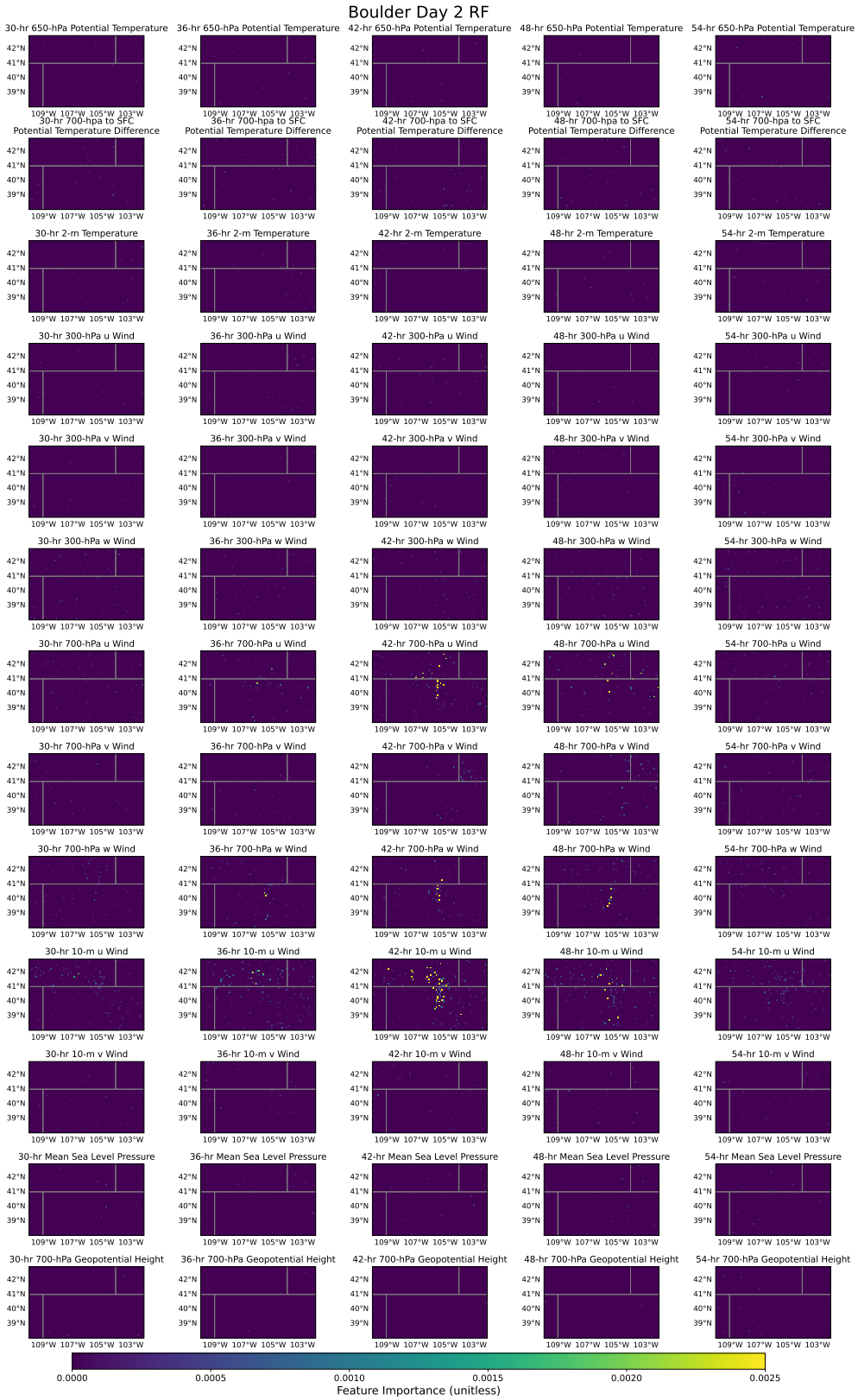

Figure A.6: Boulder Day 2 RF feature importances for each variable and forecast time comprising the input predictors. Note the change in scale from the Boulder Day 1 RF plots (Figure A.5).

APPENDIX B: DRAGMM DAY 2 FEATURE AND FULL COMPOSITES

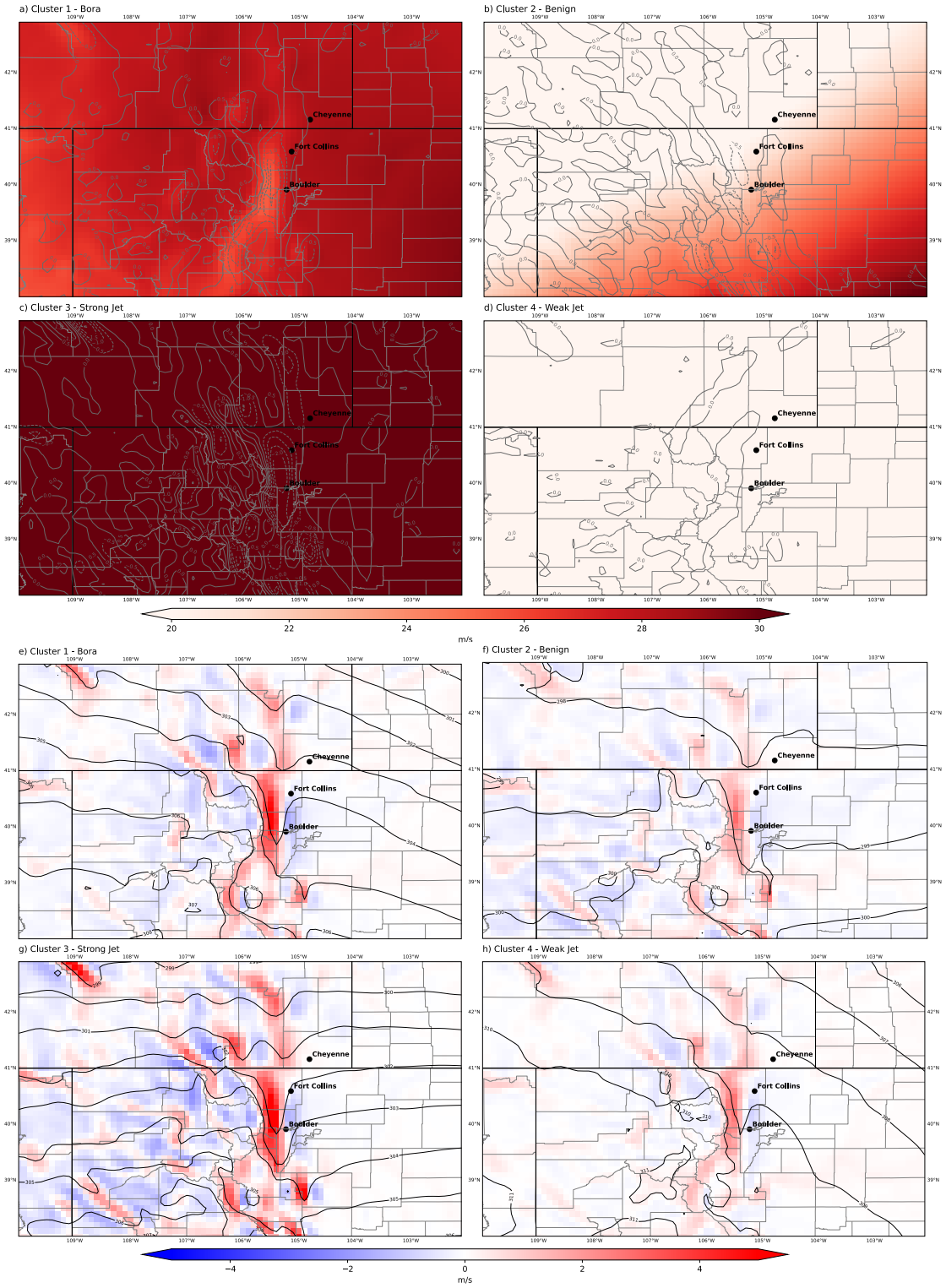

Figure B.1: Day 2 feature composites for each cluster of (a-d) 300-hPa horizontal wind magnitude (m s^{-1}, red shading) and vertical motion (m s^{-1}, grey contours) and (e-h) 700-hPa geopotential height (dam, black contours) and vertical motion (m s^{-1}, red and blue shading) over the input feature domain. Cheyenne, Fort Collins, and Boulder are labeled for geographical reference in this figure and subsequent feature composite figures.

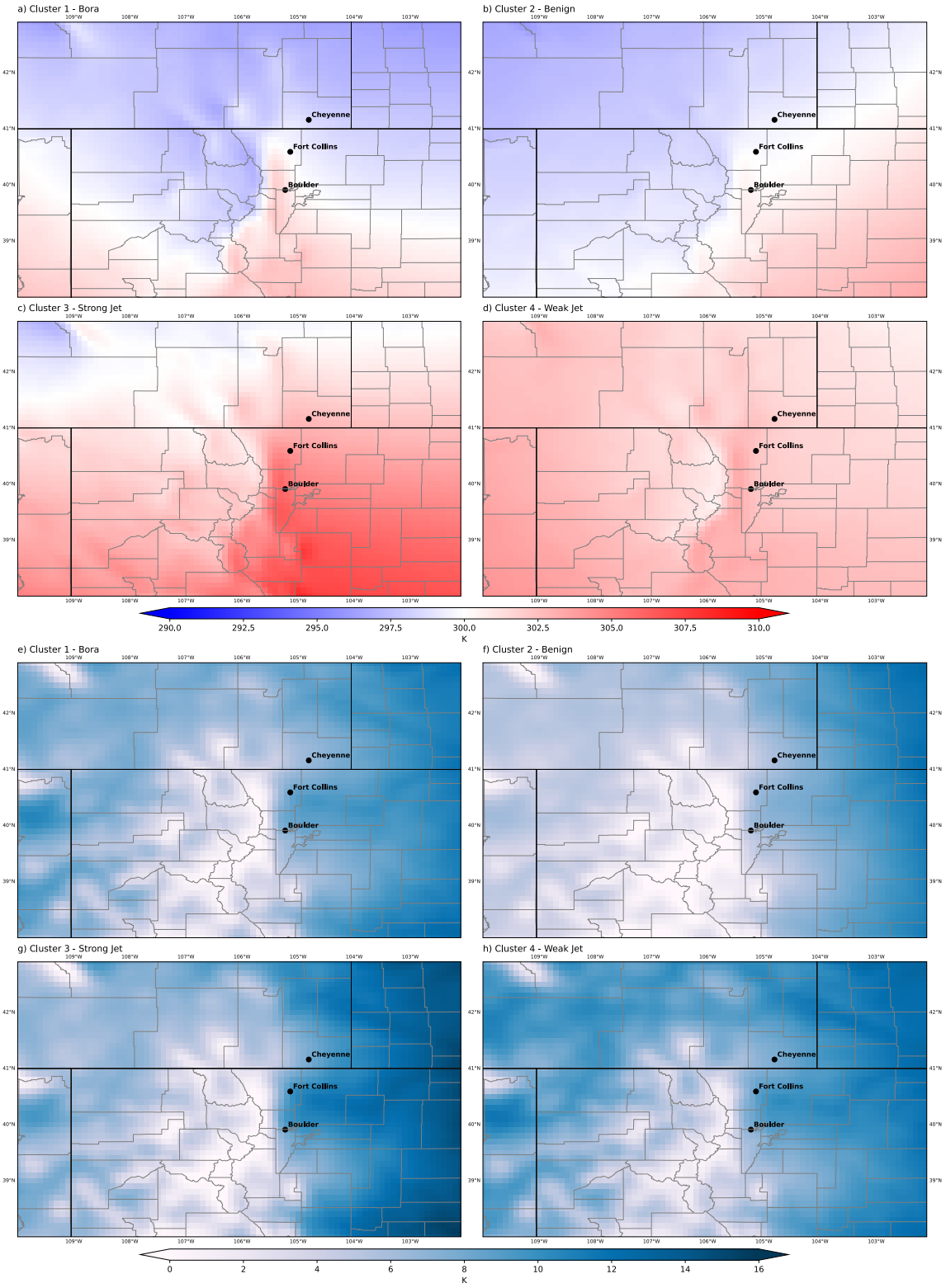

Figure B.2: Day 2 feature composites for each cluster of (a-d) 650-hPa potential temperature (K, red and blue shading) and (e-h) the potential temperature difference between 700-hPa and the surface (K, turquoise shading). Negative values in the 700-hPa to surface potential temperature difference are masked as zero as much of these negative values result from the model terrain intersecting the 700-hPa level.

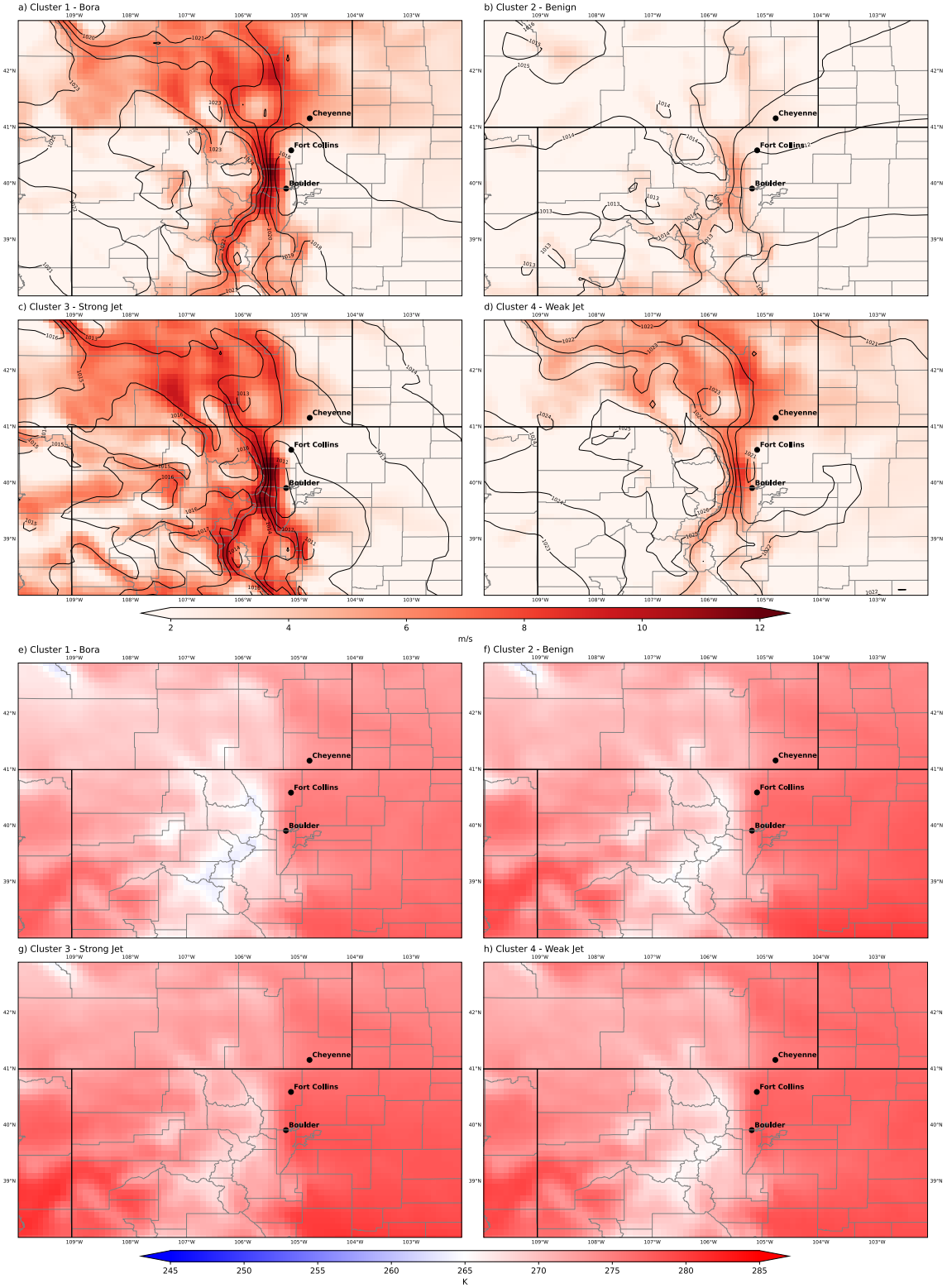

Figure B.3: Day 2 feature composites for each cluster of (a-d) MSLP (hPa, black contours) and 10-m wind magnitude (m s^{-1}, red shading) and (e-h) 2-m temperature (K, red and blue shading).

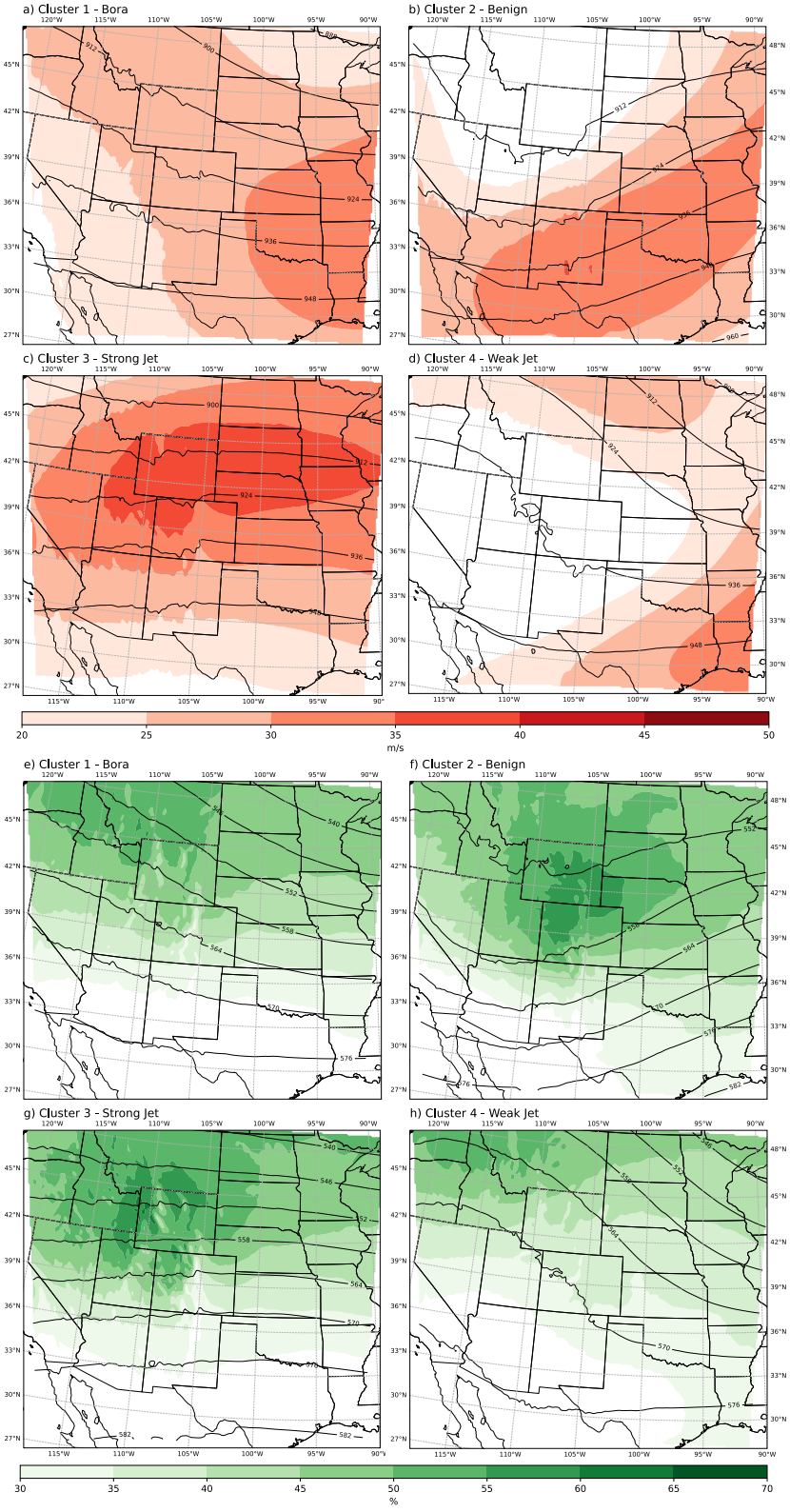

Figure B.4: Day 2 composites of (a-d) 300-hPa geopotential height (dam, black contours) and wind speeds (m s^{-1}, red shading) and (e-h) 500-hPa geopotential height (dam, black contours) and relative humidity (%, green shading) for each cluster across the CSU-WRF domain.

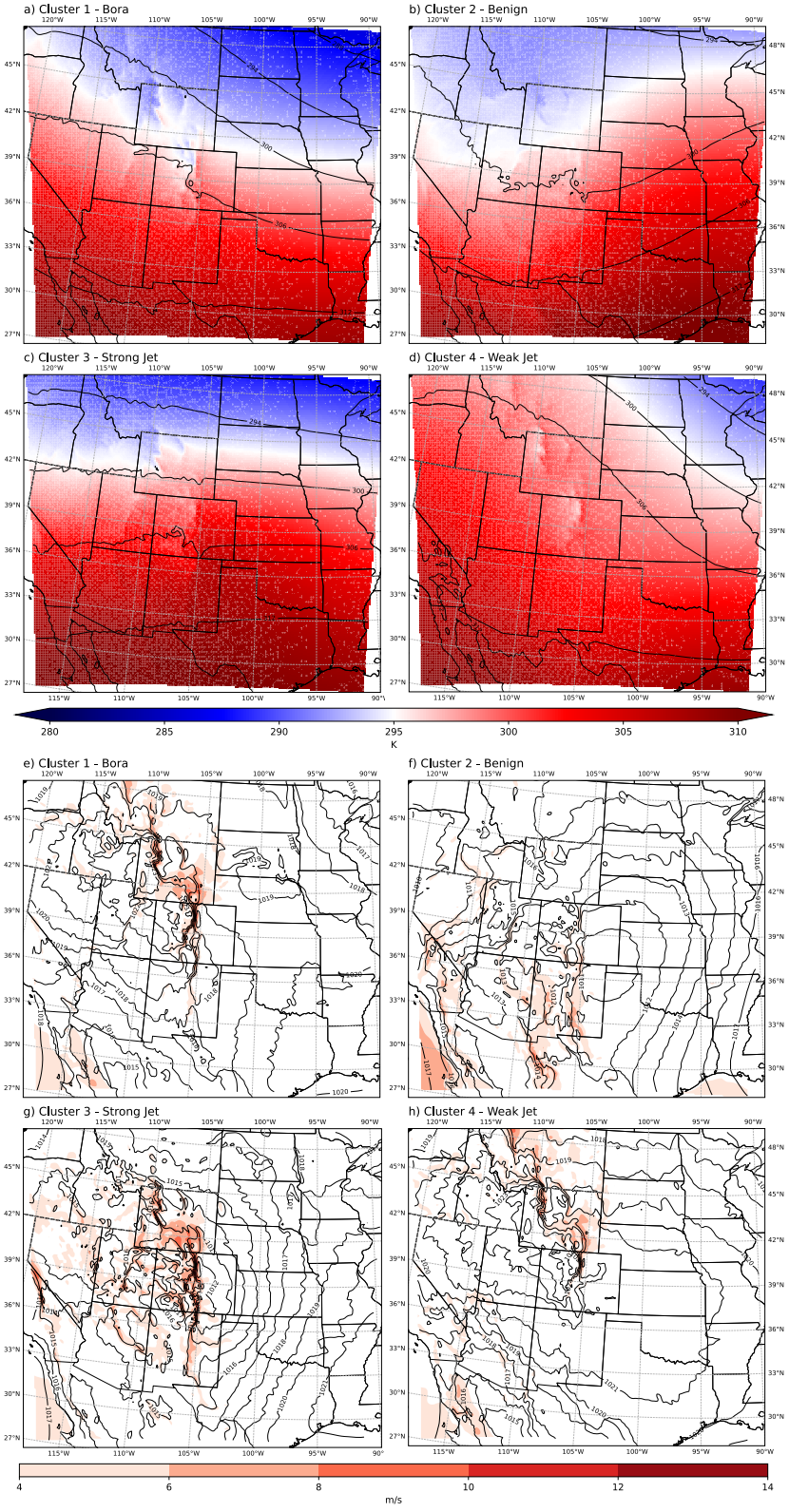

Figure B.5: Day 2 composites as in Figure B.4 but for (a-d) 700-hPa geopotential height (dam, black contours) and potential temperature (K, blue and red shading) and (e-h) MSLP (hPa, black contours) and 10-m wind speed (m s^{-1}, red shading).