DISSERTATION

INSIGHTS FROM MACHINE LEARNING-BASED FORECASTS OF CONVECTIVE HAZARDS AND ENVIRONMENTS

Submitted by

Alexandra Callahan Mazurek

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 S. Schumacher

Kristen L. Rasmussen
Susan C. van den Heever
Haonan Chen
Aaron J. Hill

ABSTRACT

INSIGHTS FROM MACHINE LEARNING-BASED FORECASTS OF CONVECTIVE HAZARDS AND ENVIRONMENTS

Severe convective thunderstorms and their associated hazards are costly, damaging, and difficult to predict. Machine learning (ML) techniques are rapidly being developed and deployed in an effort to predict severe thunderstorms more quickly and with greater accuracy than traditional methods. With these developments, there is a need to understand how ML-based weather prediction systems rely on atmospheric data and generate their forecasts. This work probes a number of ML-based convective thunderstorm-related forecasts over the contiguous United States to 1) understand how they make their predictions, 2) diagnose where their strengths and deficiencies may lie, and 3) explore how well their predictions resemble physical characteristics of the atmosphere. The insights gleaned from this research aim to support operational use of ML-based forecast guidance.

First, probabilistic ML-based forecasts of severe convective hazards (i.e., tornadoes, hail, and thunderstorm-driven winds) from the Colorado State University Machine Learning Probabilities (CSU-MLP) system are studied using an explainable machine learning technique known as Tree Interpreter (TI). TI provides context to the CSU-MLP forecasts by disaggregating its forecast probabilities into "contributions" by each of the environmental variables that are used to train the model. This technique allows one to see the extent to which each atmospheric "ingredient" contributes to the final predictions. Results of this work show that CSU-MLP uses environmental information to

make its predictions in ways that resemble the climatology and environments of severe storms, and the values of the contributions generally scale with values of the environmental inputs, effectively enhancing the interpretability of the ML system.

Second, CSU-MLP forecast performance is examined across different synoptic regimes in an effort to understand which types of environmental conditions tend to lead to skillful versus less-skillful forecast performance. Self organizing maps (SOMs), which are a type of ML, are employed to statistically diagnose regimes across two years of reanalysis data. The skill of day-2 CSU-MLP probabilistic tornado, wind, and hail forecasts are examined across the SOM-identified regimes. This work shows that SOMs are successful at identifying distinct atmospheric patterns using only surface-based convective available potential energy (SBCAPE) and vertical wind shear as inputs. At times, the best- and worst-performing CSU-MLP forecasts occur under highly similar atmospheric conditions, though the best-performing forecasts tend to be characterized by strong synoptic forcing and many storm reports.

Third, forecast output from three deep learning weather prediction (DLWP) models, GraphCast, Pangu-Weather, and FourCastNetv2, is studied to investigate how well they model severe storm environments and capture convection-related parameters. This work explores both native and derived fields from 22 months of daily forecasts from these three models, all of which were initialized with input conditions from the Global Forecasting System (GFS). The output is compared to ERA-5 reanalysis and GFS forecasts, both broadly and for specific convective events. Overarching results from this study show that the DLWP model forecasts tend to be characterized by less moisture and greater instability compared to ERA-5. For specific events, the DLWP

forecasts can reasonably capture convective environments at least a week in advance and are competitive with the GFS. However they tend to underforecast the vertical wind shear magnitude, and their limited vertical resolution can lead to overly smooth profiles that lack key details such as stable layers.

ACKNOWLEDGEMENTS

It is surreal to be writing these acknowledgments. I am unbelievably grateful to be surrounded by so many wonderful people in my life, each one of whom has played crucial roles in pulling me towards the finish line of my Ph. D. These past 5.5 years have been filled with joy, uncertainty, laughter, tears, and everything in-between, and I couldn't have made it through these ups and downs without these folks.

First and foremost, I'd like to thank my advisor, Dr. Russ Schumacher. I am so grateful that he was willing to take me on as a graduate student and for the many opportunities he's given me over the years. Russ has always believed me when I haven't believed in myself, and it is with his guidance and encouragement that I have made it here. I've learned so much from him over the past several years, and I'm very thankful for the role he has played in guiding this dissertation. Beyond research, Russ has always supported my career aspirations, and I admire his flexibility, humility, kindness, and dedication to his students. I have thoroughly enjoyed getting to learn from him during my graduate studies.

I'd also like to thank my doctoral committee: Dr. Kristen Rasmussen, Dr. Sue van den Heever, Dr. Haonan Chen, and Dr. Aaron Hill. I'm grateful for the support they've shown me throughout my graduate studies. This research is undoubtedly better because of the expertise that they've provided, and I'm thankful I've had the opportunity to learn from them. I especially thank Aaron for developing the CSU-MLP severe system; without his efforts, this work would not have been possible.

There are a number of other folks in the Department of Atmospheric Science that I'd like to thank. First, many thanks to the past and present members of the Schumacher Group with whom I've has the pleasure of getting to know: Stacey, Erik, Aaron, Faith, Jeremiah, Liu, Sam, Yi, Nathan, Eric, Jacob, JT, Casey, IT, Ayesha, Anastasia, Evan, Caleb, and Joseph. I look up to all of you, and I'm thankful for the conversations we've had together. I would also like to thank Dr. Imme Ebert-Uphoff for mentoring me throughout the third project of my dissertation. Our meetings have been extremely insightful, and I am excited to continue collaborating on this project after I defend. Additionally, I'd like to thank the Front Office Staff, especially Sarah, Nate, and Dinara, for their dedication to keeping our department running. I'd also like to thank the ATS professors that I've had the opportunity to learn from, as well as the students, postdocs, and research scientists that I've gotten to interact with. All of these people have made the department a warm and welcoming place that I've been able to grow in as a scientist.

There are a couple of professors beyond CSU that I'd like to thank as well. First, I send an immense thank you my undergraduate advisor at the University of Georgia, Dr. John Knox. Dr. Knox encouraged me to go to graduate school during my final year of undergrad, when I felt lost and confused about what was next. I am eternally grateful that he believed I was capable of pursing a higher degree, and I have no doubt that his recommendation helped me get to CSU. He has continued to be a mentor to me, both professionally and personally, and I am so thankful for all the wisdom he has shared with me. Second, I'd like to thank Dr. Jen Henderson at Texas Tech, who served on my Master's committee and has continued to mentor me since. She has been an enormous role model to me, and I have huge admiration for both her science as well as her kindness

and empathy. I am grateful for all that she has taught me about the social sciences (and academia in general). I feel so lucky to have her as a mentor.

Making the move to Colorado was very outside my comfort zone, but the friends I have made here have made it feel like home. I'd like to especially thank Chloe, Lilly, Bee, Charles, Marc, Mary Jo, Jill, and Grace for all the love and friendship they've shown me throughout my time here, especially throughout these last few months. I will cherish the laughs and adventures that I've shared with them forever (and hopefully there will be many more to come!) I would also like to thank Catharine, my "second sister", who I have been lucky to call a friend for over 20 years now. It's been a joy to grow up together!

I am also so thankful to my partner, Ryan, who has stayed by my side for many years. I am grateful that he followed me out to Colorado and for all the support he has shown me along the way. Ryan has taught me how to be curious, take risks, laugh more, and better balance work and life. He is smart and resilient, and I hope I can one day be half as good of a programmer as he is.

Lastly, I want to thank my family. It is no exaggeration to say that I would not have made it to this point without them, and I am eternally grateful to the love and support that they have shown me over my 28 years of life. My dad, Tom, inspires me by his innovation and creativity in everything he does. He has been unwavering in his support of my education, from waiting with me for the bus each morning before elementary school to helping me move 1500 miles away for graduate school. I'm glad life brought me to Colorado because our post-skiing pizza dinners in Steamboat will be among my all-time favorite memories. My mom, Joann, is my role model. She has a heart of gold and has a fierce passion for serving others. I'm inspired by her daily selflessness and I'm eternally grateful for everything she's done to take care of me over the years. I hope to be more like her. And

lastly I'd like to thank my sister, Cate, who embodies intelligence, strength, and thoughtfulness. She is my best friend and has helped me through numerous hardships over the past several years. I'm so thankful for the laughter and joy she brings me, even when we are hundreds of miles apart. And that we got to witness the cultural reset that was The Eras Tour together, of course.

The work presented in Chapters 2 and 3 is supported by the NOAA JTTI Grant NA20OAR4590350, and the work in Chapter 4 is supported by the Cooperative Institute for Research in the Atmosphere (CIRA). The model data used in Chapter 4 were generated and graciously provided by Dr. Jacob Radford and Dr. Imme Ebert-Uphoff, and the derived convective parameters analyzed in that study were computed by Dr. Ryan Lagerquist.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Severe thunderstorms

Convection in the atmosphere can fuel the development of thunderstorms, which can produce a multitude of hazards that pose danger to life and property, including tornadoes, damaging winds, hail, lightning, and flash flooding. Thunderstorm-driven hazards are not only dangerous and deadly, but they are also extremely costly. Through October 2024, there have been 24 billion-dollar disaster events this year in the United States alone, and 17 of those events were associated with severe weather (NOAA National Centers for Environmental Information, 2024). This statistic is not specific to 2024: for 20 of the previous 25 years (2000-2024), the majority of annual billion-dollar disasters have been attributed to severe storms (Fig. 1.1).

While all thunderstorm-driven hazards could undoubtedly be described as “severe" phenomena, “severe thunderstorms" in the United States are tied to a rather specific definition. The NOAA Storm Prediction Center (SPC) states that a thunderstorm is considered “severe" if it produces hail in excess of 2.54 cm or 1 inch diameter, severe wind stronger than 93 km h^{-1} or 50 knots, and/or a tornado of any intensity (NOAA Storm Prediction Center, 2023b). Such events are not uncommon in the United States, with thousands of reports of these hazards occurring each year (Fig. 1.2).

Predicting severe weather-driven hazards has been a long-standing forecast challenge. Historical approaches to forecasting severe weather have involved a combination of understanding local climatology, pattern recognition, and evaluating environmental fields or parameters–termed “ingredients-based forecasting" (e.g., Johns, 1984; Doswell, 1980; Johns and Doswell, 1992; Doswell et al., 1996; Brooks, 2007). These methods, which are still used in practice today, largely focus on

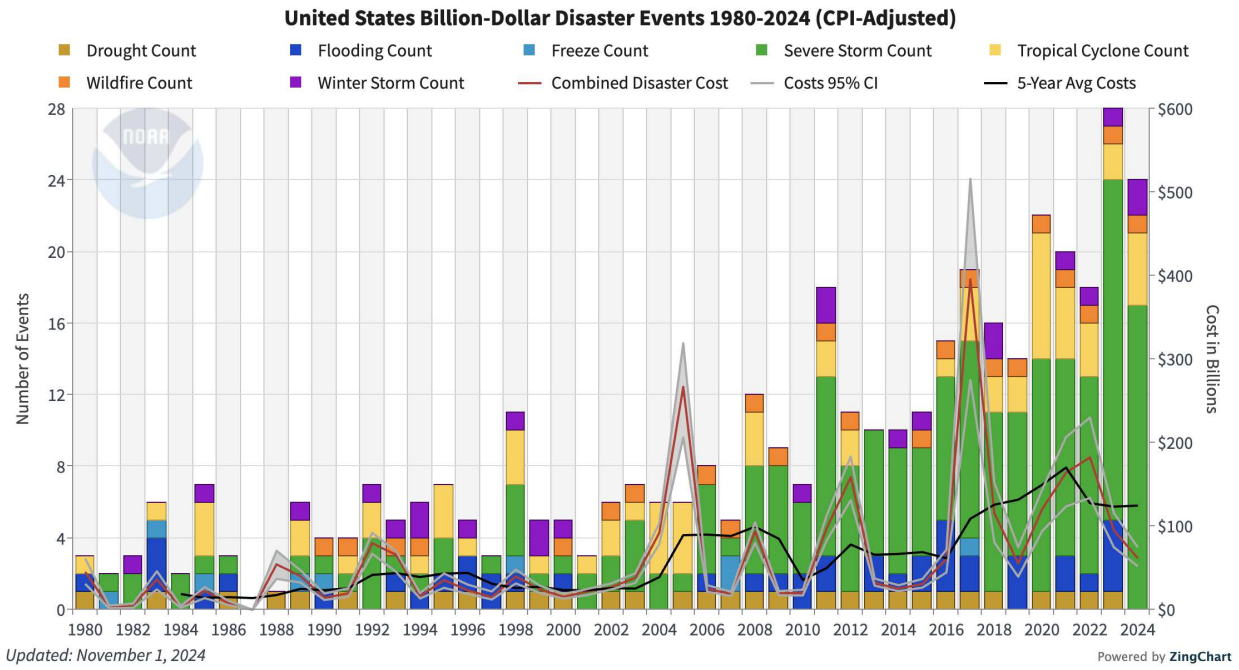

Figure 1.1: Consumer price index (CPI)-adjusted billion-dollar disasters from 1980 through 2024. Image from the NOAA National Centers for Environmental Information (NCEI) (NOAA National Centers for Environmental Information, 2024).

larger-scale environmental information (e.g., at the synoptic and mesoscale rather than the storm-scale).

Though severe thunderstorm hazards generally impact extremely local areas (which further complicates their predictability), the large scale environment can still provide rich data on where and when severe thunderstorms can occur. At its foundation, moisture, instability, and a lifting mechanism are crucial elements for deep moist convection to form (Doswell, 1987; Johns and Doswell, 1992; Doswell et al., 1996). For severe thunderstorms specifically, vertical wind shear has also been shown to be an important fourth “ingredient" (e.g., Brooks et al., 2003; Brooks, 2007). These four ingredients are typically described in terms of parameters, such as convective available potential energy (CAPE) for instability and precipitable water (PWAT) for moisture, as they provide measurements of the amount of the ingredient present (e.g., Doswell and Schultz,

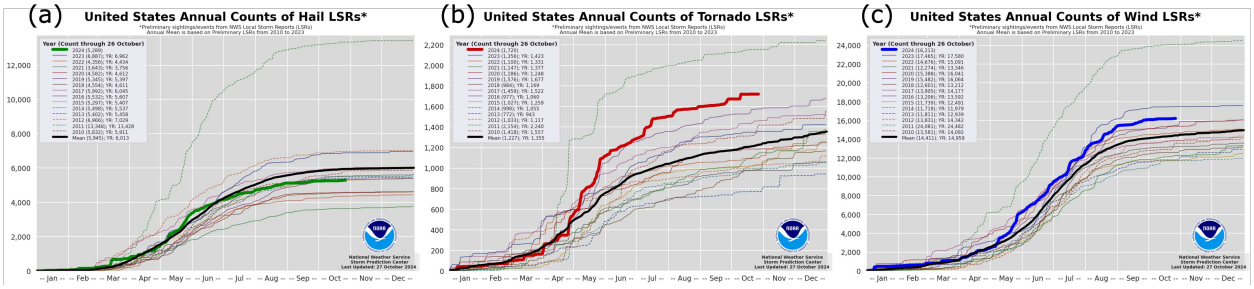

Figure 1.2: NOAA Storm Prediction Center annual cumulative counts of (a) severe hail (b) tornadoes, and (c) wind local storm reports (LSRs) since 2010. 2024 reports (green, red, and blue bolded lines, respectively) are accumulated through 26 October 2024 only. Images from the NOAA Storm Prediction Center (NOAA Storm Prediction Center, 2023a).

2006; Brooks, 2007) as well as synthesize meteorological data throughout the atmospheric column. Parameters are useful for climatological studies of severe storm environments (e.g., Brooks et al., 2003; Taszarek et al., 2020) as well as for model and observational analysis in operational forecasting.

1.2 Machine learning for severe weather forecasting

Since its inception, severe weather forecasting has strongly relied on observations and output from numerical weather prediction (NWP) systems. Raw output from NWP models can provide some insights about severe storm environments, but additional information is needed for more skillful predictions. Post-processed products can provide such guidance. These types of tools take output from NWP (and/or observations) and apply some kind of mathematical and/or statistical technique to the data to generate new output. The resulting guidance is often displayed in the form of parameters, probabilities, or proxies, which allow a forecaster to better calibrate where and when storms are imminent or occurring. Countless products have been developed in support of operational severe weather forecasting (e.g., Sobash et al., 2011; Schwartz et al., 2015; Gallo et al., 2016; Sobash et al., 2016; Smith et al., 2016; Gallo et al., 2018; Heinselman et al., 2024).

Machine learning (ML) has begun to be leveraged as an alternative post-processing technique (as well as in other ways) to aid in understanding and predicting severe weather (e.g., McGovern et al., 2023). For example, some work has utilized ML to generate probabilistic severe weather forecasts to inform operations (Burke et al., 2020; Hill et al., 2020; Loken et al., 2020; Flora et al., 2021; Loken et al., 2022; Clark and Loken, 2022; Hill et al., 2023). Others have used ML for classification tasks, including to to study properties of convective environments (e.g., Nowotarski and Jensen, 2013; Anderson-Frey et al., 2017; Nowotarski and Jones, 2018; Gensini et al., 2021; Hua and Anderson-Frey, 2022; Nixon et al., 2023) or to diagnose convective storm mode (e.g., Jergensen et al., 2019). ML has also been used in various capacities to emulate storms (e.g., Hilburn et al., 2021; Flora and Potvin, 2024; Pathak et al., 2024).

Many of these tools have shown promise for use in forecasting convective weather. However, there are many research opportunities that exist beyond the development and early testing phases of these systems. Some examples include utilizing explainability techniques to see how environmental information is used (e.g., McGovern et al., 2019), evaluating existing model performance for specific and/or complex weather system prediction tasks, and identifying potential biases in model training or prediction (e.g., McGovern et al., 2022a, 2024). In other words, with the plethora of ML weather prediction systems available presently, there is plenty of research to be conducted on these *existing* systems that does not require the training and testing of new systems (including for severe weather prediction applications specifically).

1.3 Dissertation outline and research objectives

Because of the substantial economic and and life-threatening impacts that thunderstorms can create, there is strong motivation to better understand and predict the environments that precede

thunderstorms. ML offers potential avenues to support these efforts. As such, this work broadly studies ML forecasts of convective environments and hazards. The overarching science objectives of this research can be summarized as follows:

1. to elucidate how environmental information is used to make ML-based probabilistic severe weather forecasts forecasts (Chapter 2)

2. to diagnose ML-based probabilistic forecast performance across various severe-weather-producing regimes (Chapter 3) and

3. to examine how derived convective parameters and environments in deep learning weather prediction (DLWP) models compare to reanalysis and operational forecasts (Chapter 4).

This dissertation proceeds as follows. Chapter 2 utilizes Tree Interpreter (TI), an explainable artificial intelligence (xAI) technique, to investigate two years of probabilistic random forest (RF)-based forecasts of severe weather hazards over the contiguous United States (CONUS). Specifically, TI is harnessed to disaggregate forecasts from the Colorado State University Machine Learning Probabilities (CSU-MLP) system into the meteorological features (i.e., inputs to the ML model) that were used to make forecasts. Separate examinations are conducted on CSU-MLP predictions of severe hail, tornadoes, and severe wind (as well as "any severe") and across lead times of two- to four-days. By using the TI technique, the ways in which environmental information is used by the RFs to make their predictions can be studied, providing important contextual information to their forecasts and supporting their use in operations. This work has been published in *Weather and Forecasting* [1] and is reproduced here without changes.

[1]Citation: Mazurek, A. C., Hill, A. J., Schumacher, R. S., and McDaniel, H. J. (2024). Can Ingredients-Based Forecasting be Learned? Disentangling a Random Forest's Severe Weather Predictions. *Wea. Forecasting*, **40 (2)**, 237–258, https://doi.org/10.1175/WAF-D-23-0193.1.

Chapter 3 also utilizes probabilistic hazard predictions from the CSU-MLP system, but attention is shifted from the ways in which they use environmental information to make predictions to the quality of their predictions across different environment types. In this work, a subset of the probabilistic severe weather forecasts studied in Chapter 2 (i.e., two-day forecasts only) are sorted by synoptic regimes that have been identified using self organizing maps (SOMs), and the forecast skill across each of these regimes is assessed. Using SOMs allows for regimes to be classified statistically/objectively (rather than manually/subjectively). The purpose of this section is to evaluate how CSU-MLP forecast performance may vary across different synoptic patterns, so that its output can be used more skillfully. This work is in preparation for publication.

In Chapter 4, forecasts from an entirely separate suite and variety of ML weather prediction models are studied. Specifically, output from three global NWP-emulating deep learning weather prediction (DLWP) systems (GraphCast, Pangu-Weather, and FourCastNetv2) from an archive developed by the Cooperative Institute for Research in the Atmosphere (CIRA) are studied in the context of severe weather. The archive includes raw output from these models as well as a number of derived parameters that have relevance to severe convective environments (such as instability, moisture, and shear parameters). Twenty-two months of forecasts and derived forecast parameters are examined over CONUS. This work builds on previous evaluations of these relatively new ML systems by assessing 1) how they capture convective environments relative to reanalysis and existing NWP forecasts and 2) if more complex environmental parameters (specifically those related to convection) can be extrapolated from their output. This work is in preparation for publication.

A summary of this dissertation, some avenues for future work, and final reflections are presented in Chapter 5.

Chapter 2

Can Ingredients-Based Forecasting be Learned? Disentangling a Random Forest's Severe Weather Predictions

2.1 Introduction

As machine learning (ML)-based guidance has become more commonplace in weather forecasting, it has become evident that there is a urgent need to curate resources to make these data-driven methods more understandable and trustworthy by forecasters and other end-users in the weather enterprise (Roebber and Smith, 2023). Educational tools such as course materials, review articles, tutorial-style papers, and programming modules that cover ML approaches to meteorology-specific problems have been one effort towards increasing literacy on the subject (e.g., Arcodia et al., 2022; Chase et al., 2022; McGovern et al., 2022b; Chase et al., 2023; ECMWF, 2023; McGovern et al., 2023; Molina et al., 2023; Flora et al., 2024). While these educational tools are foundational to building trust in ML methods, additional techniques are needed to build end-user confidence in specific ML-based forecasting guidance. In operational forecasting settings, emerging work has shown that having information about an individual ML model's internal architecture (e.g., features and training), development (e.g., developer affiliations), and performance can increase forecaster trust in ML products (Cains et al., 2024). Explainable artificial intelligence (xAI) techniques naturally support this transparency by explaining the internals of an ML prediction. xAI methods have already been applied across a number of ML tools that aim to forecast convective weather (e.g., Herman and Schumacher, 2018b; McGovern et al., 2019; Hill et al., 2020; Lagerquist et al., 2020; Hilburn et al., 2021; Mamalakis et al., 2022; Loken et al., 2022; Hill et al., 2023) and

has been helpful for research and development purposes, such as assessing scientific validity or optimizing compute time. Flora et al. (2024) acknowledges the increasing use of explainability techniques in atmospheric science-focused applications, and they provide a tutorial on several xAI methods using additional meteorological datasets.

Fewer efforts have been made to harness xAI specifically for assisting with operational interpretation and trust of ML-based forecast guidance, even though explainability has been shown to add value to forecasters in testbed settings (Clark et al., 2022). One potential reason for this exclusion may be due to the fact that many of the frequently-used xAI methods, such as permutation or impurity importances (e.g., Breiman, 2001; Lakshmanan et al., 2015; McGovern et al., 2019; Molnar, 2022; Flora et al., 2024) are classified as *global* explainability methods. Global explainability methods interrogate all of a trained ML model's predictions collectively by separating out its components to probe how the predictions are made (e.g., Molnar, 2022; Flora et al., 2024). When applied to ML models that are geared towards operational convective weather forecasting (e.g., Herman and Schumacher, 2018b; McGovern et al., 2019; Hill et al., 2020, 2023), global explainability methods can be useful for seeing if a model's predictions are overall consistent with what one might expect in the real world. Still, many of these global explainability methods measure feature *importance*. As a measure of global explainability, feature importances are useful for diagnosing performance or optimizing the purity of model's predictions (e.g., McGovern et al., 2019; Flora et al., 2024) by diagnosing how much each predictor influences the quality of the model's predictions. However, feature importances do not provide insights as to how features dictate *individual* predictions. This lack of detail, combined with the notion that forecasters are already tasked

with looking at too many products, offer additional explanations for why xAI has not been more readily introduced into operational forecasting alongside the new ML tools themselves.

One xAI method that has the potential to address some of the aforementioned roadblocks to introducing explainability in operations is tree interpreter (TI hereafter; Saabas, 2014; Loken et al., 2022; Flora et al., 2024). TI is a python package that allows for disaggregation of random forest (RF)-type (Breiman, 2001) ML predictions. There are a number of reasons that this particular method offers a promising avenue for harnessing xAI for operationally-focused purposes. For example, although TI can only be used for RF models specifically, there are already a number of promising RF-based products that have been developed for forecasting convective weather (e.g., Gagne et al., 2017; Herman and Schumacher, 2018a; Hill et al., 2020; Loken et al., 2020; Yao et al., 2020; Flora et al., 2021; Hill and Schumacher, 2021; Mecikalski et al., 2021; Schumacher et al., 2021; Clark and Loken, 2022; Loken et al., 2022; Hill et al., 2023; McGovern et al., 2023; Radford and Lackmann, 2023b), demonstrating that there are plenty of products that TI could theoretically be applied to, several of which some forecasters already have exposure to in testbeds (e.g., Clark et al., 2022, 2023; Trojniak and Correia, 2022) or in operations (e.g., Schumacher et al., 2021). In addition, TI presents itself as operationally useful because it is a *local feature attribution method*. In other words, TI can explain 1) how each feature contributes to an individual RF prediction (i.e., local explainability) and 2) how each predictor modulates the actual value of the prediction generated by the model (i.e., feature attribution) (Saabas, 2014; Loken et al., 2022; Flora et al., 2024).

Such attributes of TI could allow an end-user (i.e., the forecaster) to intuitively see how each meteorological predictor contributes to a single RF-based forecast, offering much needed context

to the RF's often opaque output. Further, TI's property of decomposing a prediction into meteorological parts presents itself as having similar properties to ingredients-based forecasting: a fundamental, decades-old methodology that is still widely used in operational meteorology. Forecasting with an ingredients-based approach involves an assessment of environmental conditions, often represented by parameters (e.g., convective available potential energy) or indices (e.g., lifted index), to build a forecast (e.g., Johns and Doswell, 1992; Doswell et al., 1996; Doswell and Schultz, 2006; Brooks, 2007). Ingredients-based forecasting was established (and continues to be frequently used) in the context of predicting thunderstorms and their associated hazards (e.g., Doswell, 1987; Johns and Doswell, 1992; McNulty, 1995; Doswell et al., 1996; Brooks, 2007), but it has also been applied to other forecast problems, such as predicting snowfall (Wetzel and Martin, 2001), illustrating that it is a transferable method. Forecasters have historically applied the ingredients-based approach to numerical weather prediction-based model output and observations in tandem with other forecasting methods to implicitly make probabilistic forecasts (such as the NOAA Storm Prediction Center (SPC) convective outlook; NOAA Storm Prediction Center, 2023b). However, this technique has yet to be formally applied to ML-based forecasts, largely because such products are typically not accompanied by interpretability or explainability information that yield insights as to why the model is generating a given prediction, which poses a challenge to real-time operations.

Little work has been done to formally apply TI to RF models that predict weather phenomena. However, one exception is work by Loken et al. (2022). They demonstrated that TI provided useful insights (both from a research and forecasting perspective) to their next-day convection allowing model (CAM)-based RF predictions for severe convective hazards. Their work establishes an

exciting precedent for using TI, and recent work on comparing xAI methods for weather-centric ML models has recommended that the package be explored further (Flora et al., 2022).

As such, we leverage TI here to investigate two years of RF-based probabilistic severe weather forecasts generated by the Colorado State University Machine Learning Probabilities (CSU-MLP) system (Hill et al., 2020, 2023): a tool that has been used by meteorologists at SPC and numerous NWS Weather Forecast Offices since at least early 2022 and became fully operational in Spring 2024. This work is distinct from Loken et al. (2022) in three key ways: model architecture, forecast lead time, and use of an operational (rather than experimental) forecast system. Specifically, *this work examines whether TI can provide insightful, ingredients-based insights to severe weather forecasts from CSU-MLP*, specifically those made between 36 and 72 hrs after initialization (outside the prediction window of most CAMs). The decision to apply TI to these particular forecast lead times is further motivated by 1) interest in augmenting severe weather forecasting information at lead times beyond the plethora of quality, high-resolution forecast data that exists within 36 h and 2) the choice of ML modeling system used in the study.

In this manuscript, the following questions are explored:

1. Which meteorological features tend to have the greatest impact on enhancing or constraining the CSU-MLP probabilities?

2. How do the TI-derived contributions by the most pertinent meteorological predictors vary in time, space, hazard type, and lead time?

3. Do the spatiotemporal patterns in the model probabilities and feature contributions tend to be consistent with current knowledge of severe storm environments and climatology?

Section 2 of this will provide an overview of the CSU-MLP architecture, TI algorithm, data selection and preprocessing, and aggregation techniques. Section 3 illustrates the results. Section 4 places the results in the context of severe storm climatology and environments, offers discussion on potential benefits to operational forecasting, and presents limitations of the study. A summary of this work, as well as conclusions and propositions for future research are covered in section 5.

2.2 Data and methods

2.2.1 Overview of the CSU-MLP system

Full details of the CSU-MLP system for severe weather prediction can be found in Hill et al. (2020) and Hill et al. (2023), but a summary of the model infrastructure is provided here (Fig. 2.1). CSU-MLP uses RF classification (Breiman, 2001) to generate daily 24-h probabilistic predictions of severe weather out to 8-day lead times, as well as individual severe hazards (i.e., tornadoes, severe convective wind greater than 93 km h^{-1} or 50 knots, and severe hail greater than 2.54 cm or 1 inch diameter) out to 3-day lead times. The finalized CSU-MLP products are intended to mimic the convective outlooks that are issued by the SPC (NOAA Storm Prediction Center, 2023b)[2].

Model architecture

The current system uses RFs trained on the ensemble median values of environmental predictor fields[3] (Table 2.1) from the 5-member Global Ensemble Forecast System v12 (GEFS) Reforecast Dataset (GEFS/R; Hamill et al., 2022; Zhou et al., 2022). The model training period is approximately 9 years long, incorporating daily GEFS/R initializations from 12 April 2003 through 11

[2]At the time of this manuscript submission, SPC convective outlooks for individual hazards are only issued for days 1 and 2, whereas CSU-MLP forecasts for individual hazards are issued out to day-3.

[3]Equations for derived variables can be found in the appendix of Hill et al. (2020).

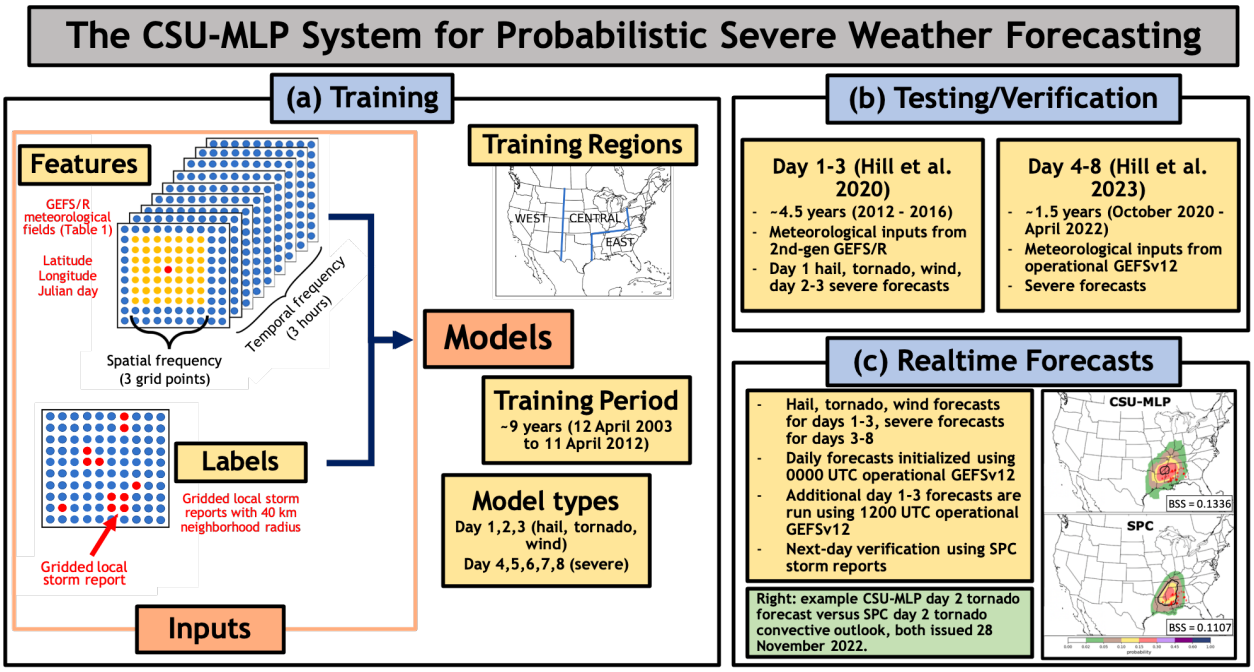

Figure 2.1: Overview of the CSU-MLP system training, historical testing and verification, and real-time forecasts. (a) summary of how the RFs are trained in the CSU-MLP system, including its feature and label assembly techniques, training regions, and training period. (b) summary of previous work that has been done to test and verify the probabilistic severe forecasts made by the CSU-MLP system. (c) overview of how the current real-time CSU-MLP forecasts are run. An example real-time CSU-MLP day-2 probabilistic tornado forecast is also shown, with the SPC day-2 tornado outlook valid for the same period below it. Both forecasts were issued 28 November 2022. Tornado reports are also overlaid as red triangles, and the black contours indicate regions of significant severe tornado probabilities.

April 2012 (Hill et al., 2020, 2023). The environmental predictors have a temporal resolution of 3 hours, so for a given 24-h forecast (valid 1200 UTC to 1200 UTC), 9 instantaneous environmental fields are used as model inputs (Fig. 2.1a). For example, a day-3 forecast would consider data from a 0000 UTC initialized GEFS/R run valid for forecast hours 60 through 84. Meanwhile, each forecast grid point (which has 0.5° grid spacing and matches the GEFS/R grid[4]) considers inputs from a horizontal radius of 3 grid points for each timestep (Fig. 2.1a). Thus, each CSU-MLP forecast grid point considers 6,615 dynamic predictors (i.e., 15 environmental variables * 49 spatial

[4]Some variables in the GEFS/R dataset that are used in training are on a 0.25° grid, so those variables are interpolated to the 0.5° grid to match the lower-resolution fields.

Table 2.1: Environmental predictors used in the CSU-MLP system for predicting severe weather as in Hill et al. (2020, 2023). Predictors with an asterisk are only used in the day 1 through day-3 models (i.e., the forecast lead times when tornadoes, wind, and hail are predicted separately). Predictors with a plus sign are derived variables that are not explicitly included in the GEFS output (see Hill et al. (2020) for equations used in these calculations).

Abbreviation	Description	Variable type
APCP	3-h accumulated precipitation	Thermodynamic
CAPE	Surface-based convective available potential energy	Thermodynamic
CIN	Surface-based convective inhibition	Thermodynamic
MSLP	Mean sea level pressure	Kinematic
PWAT	Total precipitable water	Thermodynamic
Q2M	2-m specific humidity	Thermodynamic
RH2M*+	2-m relative humidity	Thermodynamic
SHR500+	Bulk wind difference between 10 m and 500 hPa	Kinematic
SHR850+	Bulk wind difference between 10 m and 850 hPa	Kinematic
SRH*+	Storm relative helicity between the surface and 850 hPa	Kinematic
T2M	2-m air temperature	Thermodynamic
U10	Zonal component of 10-m wind	Kinematic
UV10+	10-m wind speed	Kinematic
V10	Meridional component of 10-m wind	Kinematic
ZLCL*+	Lifted condensation level height	Thermodynamic

points * 9 timestamps, plus three static predictors–latitude, longitude, and julian day) to make one 24-h prediction. Extensive feature sensitivity testing was done for the companion CSU-MLP excessive rainfall model (Herman and Schumacher, 2018a; Schumacher et al., 2021) and informally for the CSU-MLP severe model (Hill et al., 2020) to lead to the current choice in meteorological predictors, and thus those same inputs are used in this work.

Storm reports of tornadoes, severe convective wind, and severe hail from the NOAA Storm Data dataset (NOAA Storm Prediction Center, 2023a; NOAA National Centers for Environmental Information, 2023) are used as CSU-MLP training labels (i.e., historical observations; Fig. 2.1a). Reports are gridded using a 40-km neighborhood radius to the same 0.5° forecast grid that is used in the GEFS, meaning that each report will be mapped to at least one grid point (a 0.5° grid spacing

corresponds to a horizontal distance of approximately 55 km). Gridded reports are labeled separately as “severe" versus “significant severe" based on their intensity (per NOAA Storm Prediction Center, 2023b), allowing the RFs to predict 3 classes (with “no severe" being the third class). Only the “severe" predictions will be considered here.

To accommodate for spatiotemporal variability in severe weather climatology and computational limitations, separate RF models are trained over three CONUS regions: “west", “central", and “east" (Fig. 2.1a). Separate models are also trained for each of the eight forecast lead times, as well as for each of the individual severe hazards for the day 1-3 period.

The CSU-MLP system produces real-time forecasts by using environmental predictors from the *operational* GEFSv12 dataset (Zhou et al., 2022), since those fields are available nearly in real-time unlike those from GEFS/R (Fig. 2.1c). The 0000 UTC operational GEFS run is used to generate daily probabilistic predictions of severe weather out to 8-day lead times, and the 1200 UTC GEFS run is used to generate a second batch of forecasts for the day 1-3 individual hazards.

Objective performance

Previous work by Hill et al. (2020, 2023) has used multiple quantitative metrics to evaluate the skill of CSU-MLP severe forecasts at days 1-8 (Fig. 2.1b). Using Brier skill score BSS; (Brier, 1950), they show that at day-1 lead times, wind forecasts tend to have the best skill with respect to climatology, followed by the hail and tornado forecasts. In addition, while SPC outlooks were shown to be more skillful than CSU-MLP forecasts at day 1, CSU-MLP forecasts tended to outperform SPC convective outlooks at lead times between day-2 to at least day 5. However, these skill differences vary spatially, seasonally, and by hazard type.

2.2.2 Tree interpreter

To analyze the CSU-MLP forecasts, tree interpreter or TI (Saabas, 2014)–a python package that allows for disaggregation of RF predictions made in scikit-learn (Pedregosa et al., 2011)–is used. One of the most useful properties of TI is that it decomposes the probability made by a decision tree into a sum of the feature contributions and a bias term (Saabas, 2014; Loken et al., 2022). For a given decision tree in a RF, the final prediction $f(x)$ can be thought of as follows (using notation from Saabas (2014)):

$$f(x) = c_{full} + \sum_{k=1}^{K} contrib(x, k) \tag{2.1}$$

where c_{full} is the initial prediction value at the root of the decision tree, K is the number of features, and $contrib(x, k)$ is the contribution from the kth feature for a given xth feature vector (i.e., all the feature values that represent a given sample). To elaborate further, $contrib(x, k)$ for a given prediction $f(x)$ measures the sum of the changes in the training sample purity across the all nodes (excluding the root node c_{full}) where the feature k is used when a given testing sample (defined by a feature vector x) traverses the tree (see Saabas (2014) for an interactive example that illustrates this type of calculation). Thus, the contribution to the final prediction for a given feature k is dependent on the path taken along the decision tree, and this path is governed by the feature vector x. These contributions (or changes to the training sample purity) by a given feature can be positive or negative, meaning they can act to increase or decrease the value of the final prediction as the decision tree is traversed. Since a RF reflects the average of all predictions made by the decision trees in the forest (i.e., each $f(x)$), the collective prediction made by the RF, $F(x)$, can be written as such:

$$F(x) = \frac{1}{J}\sum_{j=1}^{J} f_j(x) \tag{2.2}$$

where J is the number of trees in the forest and $f_j(x)$ is the prediction made at the jth decision tree in the forest. Continuing the derivation from Saabas (2014), Eq. (2.1) and Eq. (2.2) can be combined:

$$F(x) = \frac{1}{J}\sum_{j=1}^{J} c_{j_{full}} + \sum_{k=1}^{K}(\frac{1}{J}\sum_{j=1}^{J} contrib_j(x,k)) \tag{2.3}$$

where $c_{j_{full}}$ is the initial prediction value at the root of a given decision tree j. Hence, TI demonstrates that the final prediction made by the RF can be viewed as a function of the average biases (i.e., the climatology of the training data in our application; Loken et al., 2022) and average feature contributions across all trees in the forest. CSU-MLP predictor architecture allows for feature contributions to be disaggregated in time, space, and by meteorological variable type at each gridpoint, yielding 6,615 contributions (i.e., the number of dynamic predictors).

It is worth mentioning that TI shares several similarities to a more well-known xAI method known as SHapley Additive Explanations (SHAP; Shapley, 1953; Molnar, 2022). Both methods are local explainability methods that describe the degree to which a final probability can be attributed to a given feature, but attributions are computed differently between the two techniques (Loken et al., 2022). SHAP calculates feature contributions by permuting the initial input data via subsetting the predictors, providing flexibility to explore scenarios with different predictor sets (Molnar, 2022), whereas TI requires that the input data remain fixed to calculate the feature attributions. Some tradeoffs between the TI and SHAP are transferability (TI can only be used with RFs, whereas SHAP can be used with many ML architectures), notoriety (TI is relatively new

and untested in research, while SHAP has been used across multiple disciplines for decades), and speed (SHAP values can be slower to compute than TI feature contributions) (Lundberg et al., 2020). Additionally, TI can be inconsistent with weighting the contributions, while SHAP is not. Contributions from feature spits at nodes near the decision tree leaves tend to be weighted more heavily than those closer to the roots of the trees in TI (Lundberg et al., 2020).

With these pros and cons considered, given that CSU-MLP uses RFs and has a large number of predictors, TI offers a performance advantage over SHAP for RFs, even when a tree-based SHAP approach (e.g., TreeExplainer) is used (Lundberg et al., 2020). Speed is particularly desirable in this study, given the authors' interest in investigating the potential utility of this approach in operations (where timely data is essential). Additional discussion on the similarities, differences, strengths, and weaknesses of TI and SHAP are provided by Loken et al. (2022) and Flora et al. (2024).

2.2.3 Forecast data selection and preprocessing

For this work, two years of daily day-2 through day-4 CSU-MLP probabilistic severe forecasts (initialized with 0000 UTC operational GEFS runs) between 1 January 2021 through 31 December 2022 are analyzed through the lens of TI. For day-2 and day-3, probabilistic forecasts for individual tornado, wind, and hail hazards are evaluated separately, and aggregate severe probabilities are assessed for the day-4 period. These particular model forecasts were chosen largely due to their objective (Hill et al., 2020, 2023) and subjective (Clark et al., 2021, 2022, 2023) performance at these lead times.

TI feature contributions are calculated for each of the day-2 through 4 CSU-MLP forecasts in the 2-yr period. In the calculation, only the 6,615 dynamic (environmental) predictors are included:

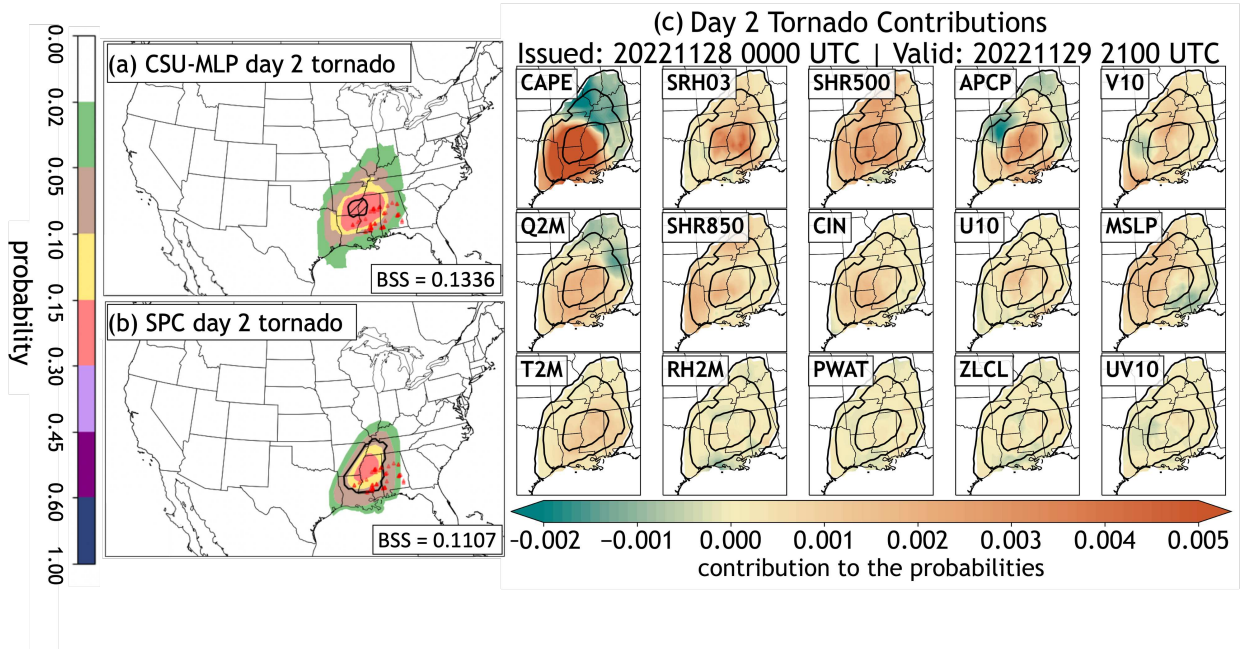

Figure 2.2: An example of the disaggregated feature contributions and masking procedure for a CSU-MLP day-2 tornado forecast. (a) the day-2 CSU-MLP tornado probabilities issued 28 November 2022 at 0000 UTC and valid for 29 November 1200 UTC to 30 November 1200 UTC. Red markers represent tornado reports. (b) the SPC day-2 tornado outlook issued 28 November 2022 at 1730 UTC. (c) feature contributions for each meteorological variable in the CSU-MLP day-2 tornado forecast shown in (a), masked according to the minimum discrete probability threshold (2%). Contributions in (c) are ordered from top left to bottom right by greatest absolute contribution value. Plots show feature contributions by variable for one timestamp (2100 UTC on 29 November 2022) within the 24-hour forecast period. Areas where the feature is contributing positively to the forecast probabilities are shown in orange shading, and areas where the feature is contributing negatively are shown in green shading. CSU-MLP forecast probabilities from (a) that are used to mask the contributions are overlaid in black using the same contour intervals. Feature abbreviations are defined in Table 2.1. This example forecast is discussed in detail in section 3d.

latitude, longitude, and Julian day are excluded. Feature importance values suggests that these three static predictors have negligible importance to the model (not shown), providing confidence in the decion to exclude them. Additionally, feature contributions occurring outside of the CONUS boundaries are excluded.

Masking procedures

Although the forecast probabilities exhibited by the system are continuous, minimum discrete thresholds are applied to the forecasts to emphasize contributions associated with probabilities that

are large enough to be operationally-relevant. These thresholds match the minimum thresholds used in SPC convective outlooks (NOAA Storm Prediction Center, 2023b) and vary by forecast lead time and hazard type (Table 2.2). Only contributions associated with daily probabilities above the minimum threshold are considered–see Fig. 2.2 for an illustration. This choice inherently focuses the results towards how the model uses environmental information to *generate* its probabilities (and uncovering which variables may enhance or constrain these above-threshold probabilities), rather than how the model uses these same inputs to ultimately *not* generate probabilities. In other words, this work emphasizes understanding how the CSU-MLP system makes its severe weather forecasts (not how it makes forecasts where probabilities are very low nor where it doesn't generate probabilities at all). Since the CSU-MLP system generates probabilities in excess of the minimum thresholds at a higher frequency than the SPC forecasters themselves (Hill et al., 2020, 2023), this decision still allows for a large sample size of forecasts to evaluate over the 2-yr analysis period.

Because of TI's additive properties, information about how each 3-h timestamp contributes to the 24-h CSU-MLP prediction window can be determined. To assess temporal variability in the feature contributions, TI is used to generate "3-h contributions" that are used to mask the contributions data on sub-daily timescales. To generate these 3-h contributions, the contributions are summed along the variable type and spatial neighborhood dimensions at each grid point for each forecast in the 2-yr period. This computation results in a single "3-h contribution" value at each grid point for each timestamp considered in a given 24-h forecast period. Similar to the previous probability-based masking procedure, minimum thresholds (Table 2.2) are applied to the 3-h contributions. These 3-h contribution thresholds are smaller than the daily probability thresholds,

Table 2.2: Minimum discrete thresholds set for the daily CSU-MLP forecast probabilities and 3-h contributions, which are used to mask the feature contributions on daily and sub-daily timescales respectively. 3-h contributions are defined as the sum of the contributions by all features for a given timestamp within the 24-h the forecast period; the sum of all 3-h contributions for a given CSU-MLP forecast is approximately equal to the daily probability for that particular forecast. Note that the daily probability thresholds match those used in the SPC convective outlooks (NOAA Storm Prediction Center, 2023b).

Forecast type	Minimum daily probability threshold	Minimum 3-h contributions threshold
Day-2 hail	5%	0.02
Day-2 tornado	2%	0.01
Day-2 wind	5%	0.02
Day-3 hail	5%	0.02
Day-3 tornado	2%	0.01
Day-3 wind	5%	0.02
Day-4 severe	15%	0.02

since the 3-h contribution values are fractions of the daily probabilities and are thus inherently smaller in value. Thresholds were chosen rather arbitrarily, largely because there are no official probability thresholds that are used operationally at SPC at sub-daily timescales. Results are sensitive to choosing a larger or smaller threshold (since the choice yields smaller and larger sample sizes, respectively), but the results remain qualitatively consistent across different thresholds.

The masking procedure using the 3-h contributions mimics the daily forecast masking procedure in Fig. 2.2, except that the 3-h contribution mask varies across each timestamp in each 24-h forecast period (rather than staying the same for the full forecast period). The purpose of applying the 3-h contribution mask to the contributions data in this way is to emphasize the contributions that are most important to the probabilities at various points in time: using the static daily probabilities mask would not allow for this temporal variability.

2.2.4 Using tree interpreter for global explainability

While TI is a local attribution method that provides insights on individual forecasts, it is not feasible to present results for two years of daily forecasts. Thus, in sections 3.1-3.2, the contributions are analyzed in aggregate over many forecasts, meaning that TI is scaled up to a global explainability method. This approach allows for a big-picture view of how the features in CSU-MLP generally act to augment or limit the probabilities. The contributions are aggregated to assess three aspects of them: 1) spatial distributions, 2) seasonality, and 3) diurnal patterns. Aggregation approaches are applied to each lead time and forecasted variable type separately.

To examine spatial distributions, the contributions are summed along the spatial neighborhood and forecast hour axes to yield a single daily contribution value for each grid point in the latitude/longitude space. Feature contributions are then summed across the two years of forecasts for each variable type, resulting in composite feature contribution maps for each CSU-MLP variable. These results are shown in section 3.1.

To assess the seasonal variability in the feature contributions (results shown in section 3.2), the contributions by each predictor to the daily forecast probabilities are computed for each forecast in the 2-yr analysis period using two aggregation approaches. The first approach analyzes feature contributions over time by their two respective variable type categories (i.e., thermodynamic versus kinematic; Table 2.1) over the 2-yr period. Here, the absolute values of the feature contributions are taken, then summed across the spatial neighborhood, forecast hour, latitude and longitude dimensions. This method provides a daily sum of the (absolute value of the) contributions to the forecast probabilities by the set of kinematic variables and by the set of thermodynamic variables. The relative contribution of each set variables to each day's forecast probabilities can subsequently

be computed. Note that while MSLP could be classified as both a kinematic and thermodynamic variable, it is categorized here as kinematic to be consistent with previous literature on the CSU-MLP system (Hill et al., 2020, 2023).

The second aggregation method used to assess seasonal variability in the contributions investigates the contributions by each variable individually (rather than categorically). For each variable, raw contribution values are summed along the forecast hour and spatial neighborhood axes, then a daily mean is computed along the latitude and longitude dimensions. The daily means are smoothed by taking a 30-day moving average for each variable.

Diurnal patterns in the feature contributions (section 3.2) are examined in two ways. First, frequency maps of the “3-h contributions" (described in the previous subsection) are plotted for each forecast hour to analyze which timestamps are contributing most to the probabilities and how that might vary spatially. Second, diurnal patterns in the contributions are analyzed across different variable types by summing over the spatial neighborhood axis, then masking the data at each forecast hour and forecast day according using the 3-h contributions threshold (see previous subsection; Table 2.2). Then, a mean contribution value across the latitude/longitude axes is computed for each forecast variable, hour, and day. This computation provides a distribution of daily mean contribution values for each variable and forecast hour, which are smoothed with gaussian kernel density estimators (KDE) that use the scott bandwidth estimator (Scott, 1979) in scikit-learn (Pedregosa et al., 2011).

2.2.5 Feature contributions versus GEFS inputs

In addition to assessing the feature contributions on their own, it is also useful to compare the values of the feature contributions to the values of the GEFS inputs that they are associated with.

This analysis can illustrate whether the relationships between the contributions and GEFS values are intuitive (e.g., does a large CAPE contribution correspond to a large value of CAPE in GEFS?).

Comparisons are made between the TI contributions and GEFS inputs on a grid point-by-grid point basis for a select number of environmental fields in section 3.3. Given the large number of grid points and for clarity of presentation, the datasets are reduced further to only include forecasts valid in the spring (March-May 2021 and 2022) and at a single GEFS forecast hour (forecast hour 45, which corresponds to 2100 UTC in the day-2 period). In an effort to fully capture the relationship between the contributions and raw NWP data at these dates/times, the probability-based mask is not applied here. In other words, contributions data that are associated with forecast probabilities below the minimum thresholds *are* included in this analysis.

2.3 Results

2.3.1 Spatial patterns in feature contributions

When the feature contributions are summed across each variable for the two years of forecasts, similar results were found for the day-2 and 3 forecasts. Therefore, spatial patterns in the day-2 and day-4 forecasts are emphasized, and a few select variables are focused on (Fig. 2.3). Because aggregated contribution values are dependent on the magnitude and frequency of the probabilities[5], interpretation of the raw contribution values is slightly complex. Thus, interpretation of Fig. 2.3 is kept more qualitative rather than quantitative.

Accumulated precipitation (Fig. 2.3a-d) and surface-based CAPE (Fig. 2.3e-h) are generally positive contributors to the CSU-MLP probabilities for day-2 through 4 forecasts for all hazard

[5]More feature contribution values will be accumulated in locations where more CSU-MLP probabilities have occurred, and the values of the most pertinent contributions to the model forecasts will inherently be higher in locations where probabilities were higher.

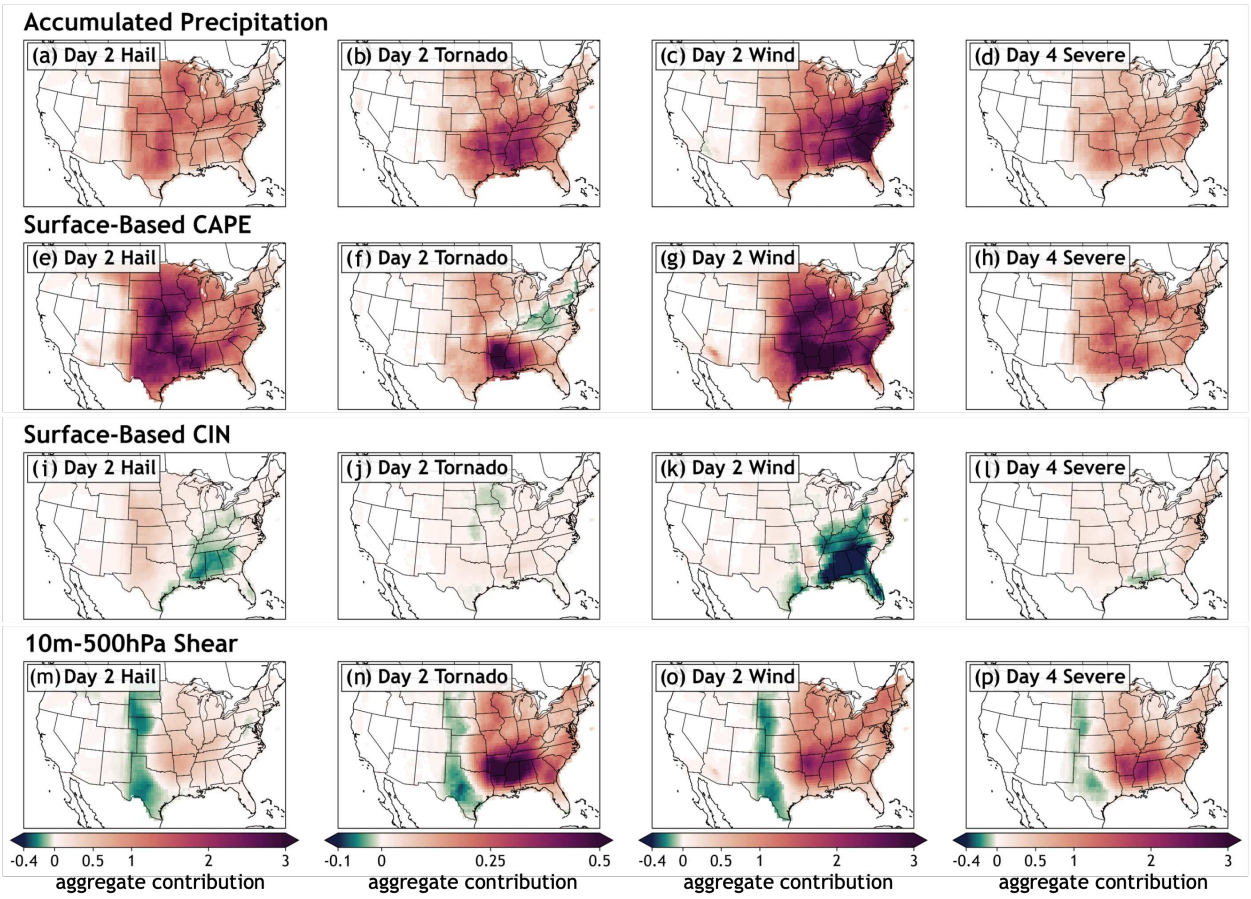

Figure 2.3: Aggregated feature contributions to the CSU-MLP forecast probabilities by (a)-(d) accumulated precipitation, (e)-(h) surface-based CAPE, (i)-(l) surface-based CIN, and (m)-(p) 10m-500hPa shear for two years of forecasts spanning January 2021 to December 2022 period. From left to right column, the CSU-MLP forecasts that these contributions are derived from are day-2 hail, day-2 tornado, day-2 wind, and day-4 severe. Note that scaling on the figures for the day-2 tornado forecasts is different than for the other figures.

types. That is, accumulated precipitation and CAPE tend to enhance (rather than constrain) the model predictions in most locations. Previous work has shown surface-based CAPE to be one of the most important (if not *the* most important) predictors in CSU-MLP forecasts across most forecasts and lead times according to Gini impurity-calculated feature importances (Hill et al., 2020, 2023), though accumulated precipitation has shown much less importance according to this

metric. This result is interesting considering that features usually tend to rank similarly across both of these explainability methods (Flora et al., 2022).

Spatial patterns in the aggregated CIN (Fig. 2.3i-l) and aggregated 10m-500 hPa shear (Fig. 2.3m-p) fields are different than those for CAPE and accumulated precipitation. In general, CIN strongly constrains the forecast probabilities in parts of the Southeast and Ohio Valley, particularly in the day-2 hail (Fig. 2.3l) and wind (Fig. 2.3k) forecasts. It is unclear why this variable may be so important to the forecast probabilities in this area and not other regions. Aggregate contributions by 10m-500hPa shear show the prevalence of negative aggregated contributions across the Great Plains in both the day-2 and day-4 forecasts (Fig. 2.3m-p). A possible explanation for this pattern could be related to the terrain in these regions: elevation increases in the High Plains with proximity to the Rockies, and elevation is also greater along the Appalachian Mountains. Increasing elevation decreases the depth of the shear layer between 10m and 500 hPa, and this reduced depth could influence how the GEFS-derived shear calculations are used in the random forests. Additional model training and testing would be needed to confirm this hypothesis. Outside of the Great Plains, 10m-500hPa shear largely augments the forecast probabilities, particularly over the Deep South.

Overall, APCP, CAPE, and SHR500 have the largest and most extensive positive contributions to the forecast probabilities across all or most lead times and hazards. However, there are a few additional variables that are not shown here that have a role in influencing probabilities across certain regions that are worth noting. For the day-2 and 3 hail forecasts, SHR850 has a profound role in enhancing probabilities over the Great Plains (a region where SHR500 largely constrains those forecast probabilities; Fig. 2.3m). MSLP (not shown) tends to positively enhance hail probabilities

over this same region–usually to a lesser extent than SHR850 but to a greater extent than SHR500. LCL enhances probabilities in most locations, though its positive impact is most noticeable in the day-2 and 3 wind and tornado forecasts across the Appalachians and parts of the Deep South (not shown). Meanwhile, PWAT tends to limit hail probabilities over the Midwest and eastern Texas for day-2 hail forecasts, while 2-m specific humidity positively impacts tornado probabilities across parts of the Southeast (though to a lesser extent than other environmental fields).

Spatial patterns in the feature contributions illustrate that there are a few key variables in CSU-MLP that tend to have the most dominant role in increasing forecast probabilities. Interestingly, these variables that are most crucial to the CSU-MLP forecasts represent ingredients that are fundamental for severe convection: CAPE (instability), APCP (moisture and lift), and SHR850/500 (shear) (e.g., Doswell et al., 1996; Brooks and Craven, 2002; Brooks et al., 2003). Without these key variables in place, it seems that the RFs have learned that severe convection is very unlikely (if not impossible), which is also true in the real world.

2.3.2 Temporal patterns in feature contributions

Here, patterns in the feature contributions are investigated seasonally and diurnally. For brevity, the day-2 CSU-MLP forecasts are discussed in greatest depth.

Seasonality

Fig. 2.4 illustrates the relative degree to which the thermodynamic versus kinematic variables influence the forecast probabilities. It is evident that there is noticeable day to day variability, but taking the 30-day running means of the relative contributions from category reveals longer-term patterns. This smoothing illustrates that the thermodynamic features typically have a larger role in

influencing the daily forecast probabilities compared to the kinematic features, and this observation is consistent across all three day-2 forecasted hazard types (Fig. 2.4), and these results are consistent in the day-3 forecasts (not shown). The dominance of the thermodynamic variables over the kinematic variables is most pronounced in the warm season–especially mid- to late summer into early fall– for all three hazards.

In the cool season, there are some subtle differences among the relative contributions by the thermodynamic and kinematic features for each hazard type. For example, in the day-2 hail forecasts, the relative contributions by each feature category are much closer to being equal compared to the warm season, though the thermodynamic variables still mostly outweigh the kinematic variables (Fig. 2.4a). Similar patterns are found in the day-2 tornado (Fig. 2.4b) and day-2 wind (Fig. 2.4c) forecasts, though the 30-day moving averages of the relative contributions by each variable type are closer to being evenly weighted in the cool season for these forecasts compared to the hail forecasts. Thus, while thermodynamic variables tend to have a larger impact on the model's probabilities relative to the kinematic variables, the extent to which this is true varies by hazard type and season.

When the day-2 CSU-MLP probabilities are disaggregated by individual features (see methods for details), several notable patterns emerge. For example, accumulated precipitation (APCP) is a meaningful contributor to the probabilities year-round across all three hazards. Surface-based CAPE also has an important role in increasing CSU-MLP probabilities: its contributions are most apparent in the warm season hail probabilities (Fig. 2.5a), followed closely by wind (Fig. 2.5c), then tornado probabilities to a lesser extent (Fig. 2.5b). LCL height augments the tornado probabilities to a greater extent than the other hazards, especially in the late spring to summer (Fig. 2.5b).

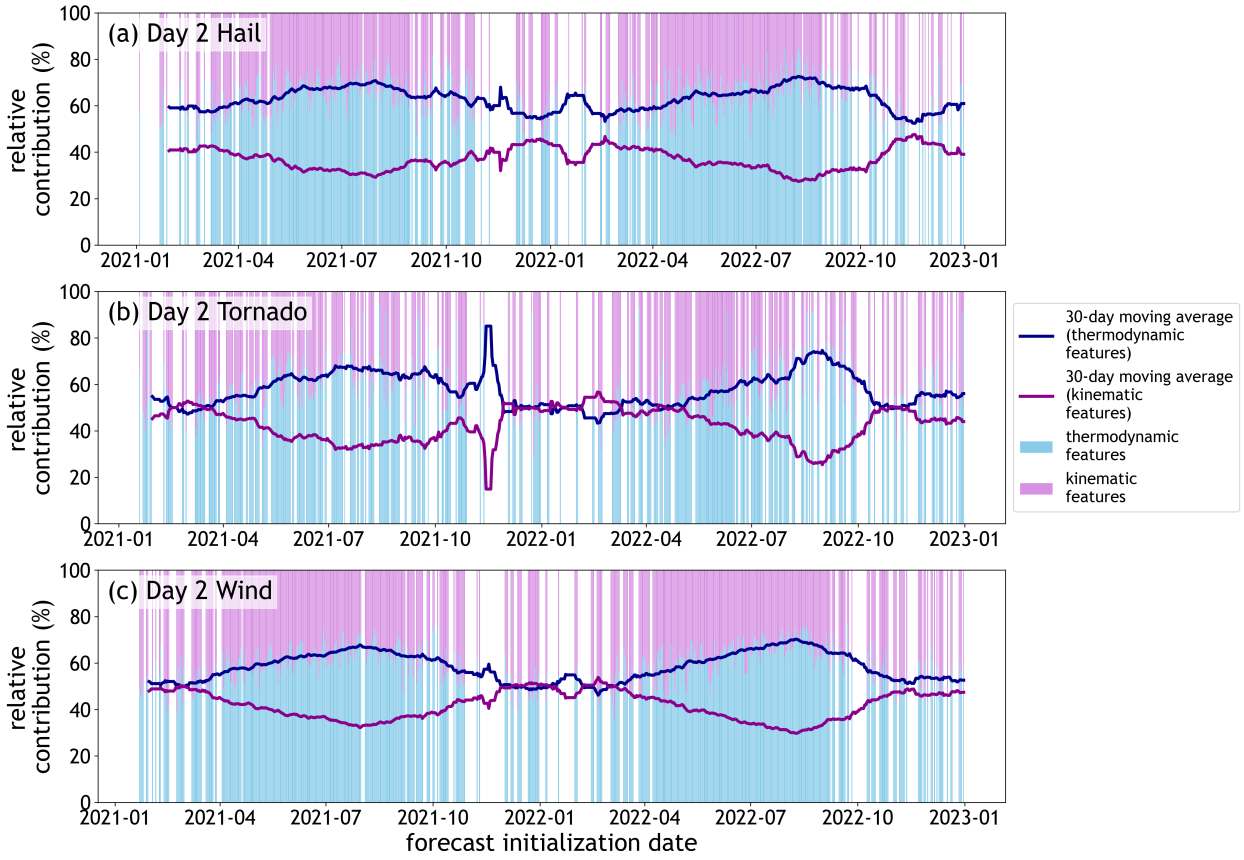

Figure 2.4: Relative contributions to the forecast probabilities by the daily-summed absolute value of the feature contributions by thermodynamic features (blue) and kinematic features (purple) for the daily day-2 (a) hail, (b) tornado and (c) wind CSU-MLP forecast probabilities over the January 2021 to December 2022 period. Blue and purple bars indicate the daily relative contributions by the thermodynamic and kinematic features, respectively, to the daily forecast probabilities (see text for more details on the aggregation approach). The 30-day rolling means of these relative contributions by thermodynamic and kinematic features are shown by the blue and purple lines respectively.

This feature is consistent with real-world knowledge of tornadic environments, as LCL height has been shown to be a key discriminator between environments that do or do not produce tornadoes (e.g., Brooks and Craven, 2002; Brooks et al., 2003; Thompson et al., 2012), whereas it is not quite as informative for other severe hazards such as hail (Brooks and Craven, 2002; Thompson et al., 2003; Nixon et al., 2023).

The vertical wind shear variables (SHR850 and SHR500) have the largest role in dictating the seasonal variability among the kinematic features, though the magnitude of their relative contributions vary by forecasted hazard type. For example, in the day-2 tornado and wind forecasts, SHR500 has a prominent role in influencing the forecasts during the cool season (i.e., December through April; Fig. 2.5b,c), whereas this feature has less influence on the hail probabilities. On the other hand, SHR850, has an important role in the hail forecast probabilities (Fig. 2.5a). SHR850 is also a notable contributor to the day-2 wind probabilities, though to a lesser extent than SHR500. Both shear variables instrumental in causing the kinematic variables to have greater relative importance in the cool season wind probabilities (Fig. 2.4c).

Physically, the differences in the importance of SHR500 and SHR850 to the model forecasts for each hazard are intriguing. It is well-known that shear in the lowest levels of the atmosphere (below 1 km) tends to be crucial for tornado formation, whereas deep-layer shear and shear at higher levels (above 1 km) tends to have greater importance for hail formation (Nixon and Allen, 2022). Yet surprisingly, the feature contributions here seem to suggest that the shallower shear variable (SHR850) is used by the model to delineate hail probabilities more than SHR500 is, whereas the opposite is true for the tornado probabilities. Without further analysis, one can only speculate on why CSU-MLP uses the shear information in these ways. Possible explanations range from flaws in the model training to potential gaps in our understanding of how shear correlates with severe weather environments. It is also possible that not holding the shear depth constant (which will inherently affect the shear values and therefore the values of the feature contributions) introduces a confounding factor in the dataset that the model is not aware of. Regardless, this example illustrates

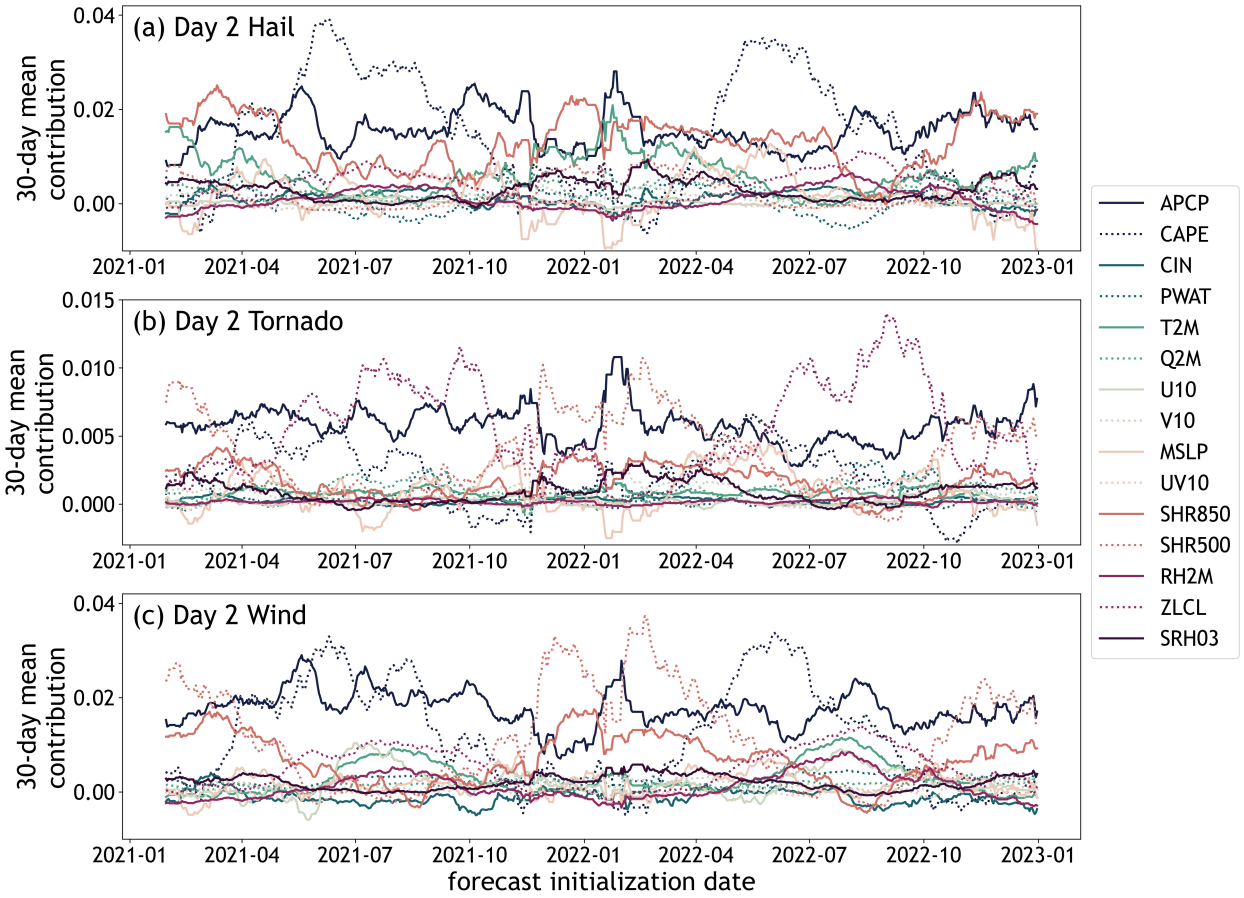

Figure 2.5: 30-day rolling means of the mean daily contributions of each feature to the day-2 CSU-MLP (a) hail, (b), tornado, and (c) wind probabilities over the January 2021 to December 2022 period. See text for full details on the aggregation approach. Feature abbreviations are defined in Table 2.1.

that while TI can provide powerful insights into how a model makes a prediction, there are limits to how much information it can provide.

There are several predictors that contribute very little to the forecast probabilities. For all three hazards, these features include Q2M, UV10, U10, V10, and CIN (Fig. 2.5). This result suggests that these variables are generally unimportant to increasing the severe hazard predictions made by the CSU-MLP system, though they may still hold some relevance in individual forecasts (e.g., Fig. 2.2) and limiting the probabilities (i.e., negative contributions).

Diurnal patterns

Frequency plots of severe reports (aggregated at 3-h intervals) (Fig. 2.6) are used to contextualize frequency plots of the 3-h contributions for day-2 and day-4 forecasts (Figs. 2.7; 2.8). While the sub-daily day-3 contributions will not be discussed explicitly, spatial patterns in those forecasts strongly resembled those in the day-2 forecasts.

In general, the day-2 sub-daily contributions occur most frequently during local mid afternoon to late evening (i.e., 2100 UTC to 0300 UTC; Fig. 2.7d-l) and relatively infrequently in the early morning (not shown), though there is variability amongst the three hazard types. In the day-2 hail forecasts, 3-h contributions gradually ramp up in frequency throughout the daytime, and peak in occurrence during the late evening to overnight hours over the Great Plains (Fig. 2.7j,m). Meanwhile, across the lee of the Appalachians, the 3-h contributions peak in frequency during the afternoon to early evening (Fig. 2.7a,d) as well as overnight (Fig. 2.7m).

When compared to storm report patterns over the 2-yr study period, hail observations in the lee of the Appalachians peak during a similar time frame to the peak in the contributions, with most reports occurring between 1800 to 0000 UTC in this area (Fig. 2.6d,g). This timing is consistent with diurnal patterns in longer term hail observations for this particular region (Wendt and Jirak 2021). Meanwhile, spatial patterns in the 3-h contributions over the central to southern Great Plains align reasonably well with locations of reports, though the hail report frequency peaks earlier than the peak frequency in 3-h contributions (i.e., 2100 to 0300 UTC in the reports (Fig. 2.6g,j) versus 0300 to 0600 UTC in the sub-daily contributions). Diurnal patterns in the two years of reports are consistent with longer term hail records, which show a peak in reports around 0000 UTC (Allen and Tippett, 2015; Wendt and Jirak, 2021), suggesting that the CSU-MLP system may

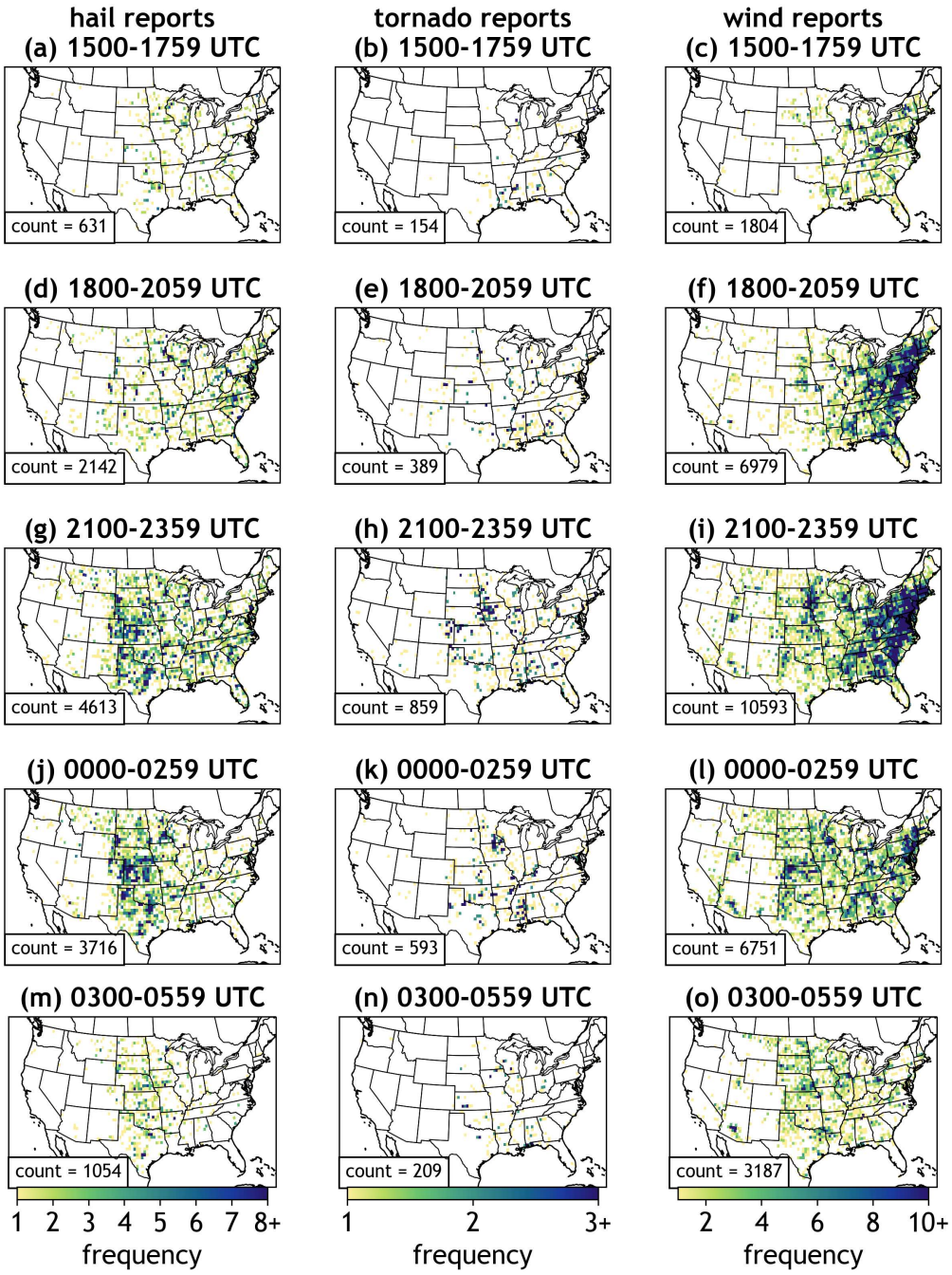

Figure 2.6: Gridded reports of hail (left column), tornadoes (center column), and wind (right column) from January 2021 through December 2022. Reports are compiled over select 3-h increments: (a)-(c) 1500-1759 UTC, (d)-(f) 1800-2059 UTC, (g)-(i) 2100-2359 UTC, (j)-(l) 0000-0259 UTC, and (m)-(o) 0300-0559 UTC to illustrate diurnal variability. Each report is gridded to the nearest 0.5° grid point in the GEFS dataset (Hamill et al., 2022; Zhou et al., 2022). Storm reports are sourced from the NOAA Storm Data dataset (NOAA Storm Prediction Center, 2023a; NOAA National Centers for Environmental Information, 2023).

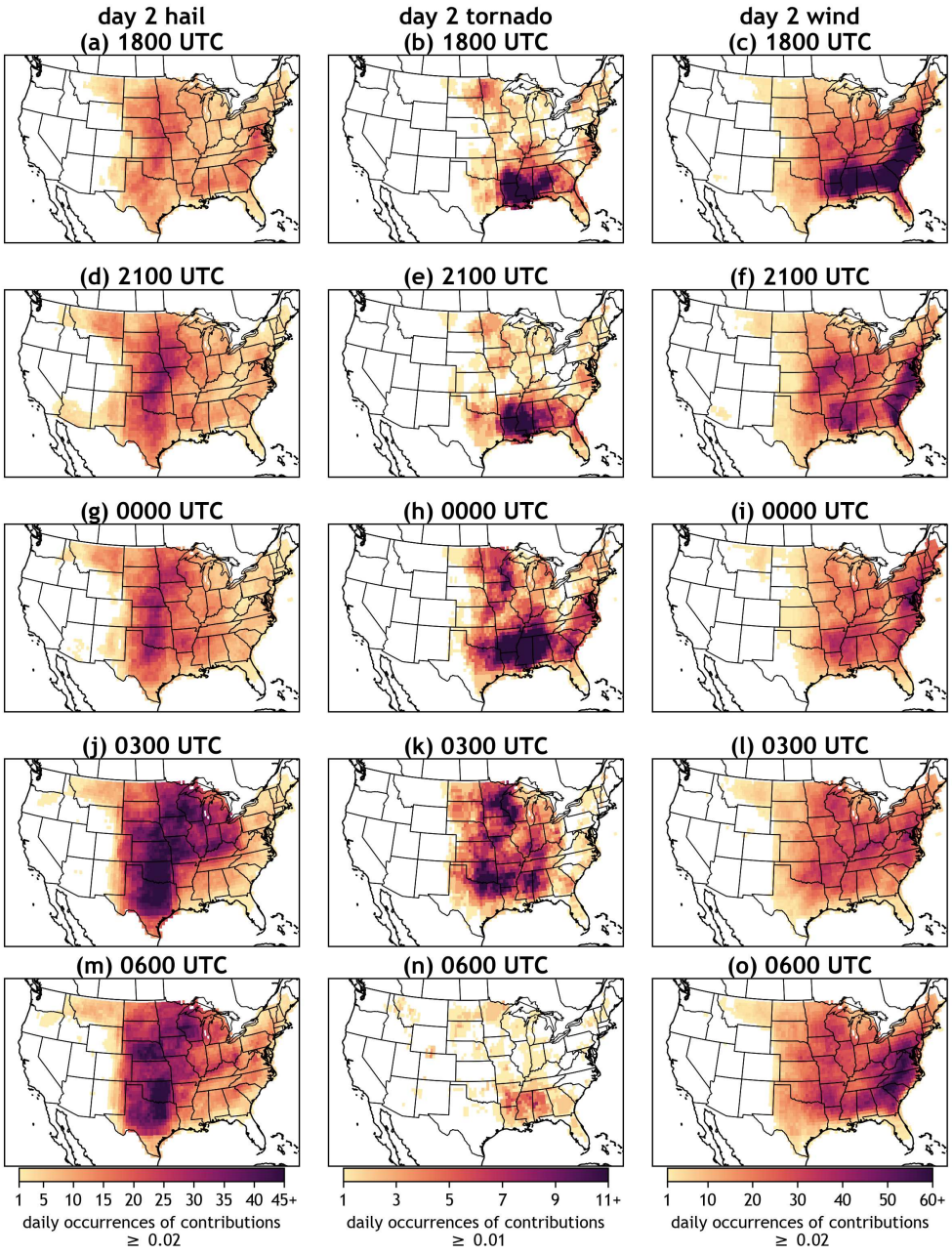

Figure 2.7: Heatmaps showing the number of days in the January 2021 to December 2022 period when the 3-h contributions for the day-2 CSU-MLP forecasts exceed 0.02 in hail or wind forecasts (left and right columns, respectively) or 0.01 in tornado forecasts (center column). Heatmaps are shown for the daily frequencies of the 3-h contributions exceeding these thresholds at forecast periods ending at (a)-(c) 1800 UTC, (d)-(f) 2100 UTC, (g)-(i) 0000 UTC, (j)-(l) 0300 UTC, and (m)-(o) 0600 UTC within the day-2 period. These timestamps correspond to forecast hours 42, 45, 48, 51, and 54 respectively. Forecast times before 1800 UTC and after 0600 UTC are not shown to conserve space and because the 3-h contributions at these timestamps are small and infrequent compared to the rest of the forecast period.

have a systematic tendency to rely on the GEFS information at timestamps that extend later into the overnight hours (after the climatological maximum in hail reports has occurred) to generate its hail forecast probabilities.

The day-2 sub-daily tornado contributions occur over parts of the Deep South (i.e., Arkansas, Louisiana, Mississippi, and Alabama) relatively often over the 2-yr period for all forecast hours (Fig. 2.7b,e,h,k,n), including in the morning forecast hours (not shown). Other locations see sub-daily tornado contributions on a similarly frequent basis as the Southeast, such as the Midwest, parts of the Great Plains and the Mid-Atlantic (Fig. 2.7h,k), though contributions over those regions peak in frequency over a more narrow time frame within the diurnal cycle (i.e., evening to early overnight). The peak in tornado reports over the 2-yr period occurs in the late afternoon to late evening hours when examining CONUS as a whole (i.e., 2100 to 0300 UTC; Fig. 2.6h,k), though they are not rare later or earlier in the day, particularly across the Southeast and parts of the Midwest (e.g., Fig. 2.6e,n). It appears that CSU-MLP does utilize information from the GEFS (albeit infrequently) over those locations during these times of day, and there are essentially never sub-daily tornado contributions outside of this region during the late night or morning hours. These characteristics are consistent with previous work that has examined regional variability in diurnal patterns of tornado reports. Tornado activity in the Southeast tends to be distributed throughout the day, whereas it peaks in the late afternoon to evening for most other regions (e.g., Krocak and Brooks, 2018), due to regional variability in tornadically favorable environments during the day versus night (e.g., Kis and Straka, 2010; Sherburn et al., 2016). Thus, despite having no information on diurnal timing on either the storm reports or the GEFS meteorological fields that are used to train the RFs, it appears that the CSU-MLP system is capable of learning the times of day

when environments tend to be most favorable for tornadic activity (including how that might vary spatially).

The 3-h day-2 wind contributions occur most frequently in the early afternoon across much of the Southeast and Mid-Atlantic compared to other times of day (Fig. 2.7c), though it is not uncommon for GEFS information to influence forecast probabilities in those areas during other times of day such as the evening (Fig. 2.7f,i), overnight (e.g., Fig. 2.7o) or late morning (not shown). The high frequency of reports across the Mid-Atlantic coincide well with the frequent 3-h wind contributions in the afternoon through late evening hours (Fig. 2.6c,f,i,l), though the increased frequency in sub-daily contributions after 0600 UTC (Fig. 2.7o) is not reflected in the wind reports (Fig. 2.6o). Further, the large number of reports across New England do not seem well-represented by the frequency the sub-daily contributions in this region. This discrepancy could be attributed to choosing to plot the number of reports at each grid point versus the number of days with severe wind reports (i.e., it is not unusual to get several hundred wind reports along the East Coast in a single 24-h period). The Midwest and Ohio Valley lack an obvious diurnal pattern in the frequency of the sub-daily contributions, despite relatively frequent wind probabilities those areas (not shown). However, wind gust reports increase in frequency slightly later in the day across these locations compared to the East Coast (Fig. 2.6i,l,o).

The spatial patterns in the day-4 sub-daily contributions (Fig. 2.8) combine some features of the sub-daily contributions for day-2 individual hazard forecasts, but there are also some differences. For example, the Mid-Atlantic displays frequent 3-h contributions in the late morning to early afternoon compared to other locations (Fig. 2.8b,c) much like the patterns shown in the day-2 sub-daily contributions to the wind probabilities (e.g., Fig. 2.7c). Parts of the Great Plains also have

frequent 3-h contributions at day 4 in the late evening and overnight hours (Fig. 2.8f,g), similar to the day-2 sub-daily hail contributions (Fig. 2.7j,m). A notable difference between the day-2 and day-4 sub-daily contributions is that the Midwest frequently sees 3-h contributions at 2100 UTC in the day-4 forecasts (Fig. 2.8d), whereas the 3-h contributions for this area were more common in the late evening to early overnight hours in the day-2 individual forecasts (e.g., Fig. 2.7h,j,k,m). One hypothesized explanation for these differences is related to the GEFS ability to predict the timing of severe storms across increasing lead times. That is, as lead time decreases, the GEFS may be able to better diagnose the timing of when severe hazards are most likely to occur, whereas it might struggle to capture favorable environments outside the diurnal heating maximum at longer lead times.

One perplexing feature in the day-4 sub-daily contributions that is also present in the day-2 and 3 forecasts (not shown) is that there are some areas with frequent contributions at the first forecast hour (1200 UTC; Fig. 2.8a) that are not similarly highlighted during the last forecast hour (1200 UTC; Fig. 2.8i). Possible explanations for these discrepancies could be 1) the model has learned in training that the pre-storm environment is more important than the post-storm environment, 2) there is smaller ensemble spread in the GEFS at the start versus end of the period, which varies the predictor and contribution values at those times, and/or 3) it is an unknown statistical artifact of the TI or RF algorithms.

To further diagnose diurnal variability in the feature contributions, a couple key environmental predictors (surface-based CAPE and SHR500) can be analyzed temporally (see methods and Fig. 2.9 caption for more details). For simplicity, this analysis is restricted to only day-2 forecasts (though there are again similarities in the results from the day-3 probabilities). Distributions of

3-h mean contributions by CAPE show a strong diurnal pattern across predictions for all three hazards (Fig. 2.9). The greatest mean 3-h CAPE contributions tend to occur in the local afternoon to evening hours (i.e., 1800 UTC to 2100 UTC), which is consistent with when atmospheric instability tends to be at its greatest. From overnight to mid-morning, (e.g., 0600 UTC to 1200 UTC), the mean CAPE contributions are almost always small. Surface-based instability is typically eroded significantly (if not altogether) overnight due to stability induced by nocturnal cooling in the boundary layer, so it seems that contributions by SBCAPE tend to be larger during times of day that SBCAPE values are typically larger. While all three hazards show diurnal variability in 3-h mean contributions by CAPE, smaller contribution values (near zero) in the afternoon to evening hours are not as uncommon for the tornado probabilities as they are for the wind and hail probabilities (c.f., Fig. 2.9b and 2.9a,c). This pattern suggests that CAPE may have a greater role in enhancing the hail and wind probabilities than it does for the tornado probabilities. Tornado events are not uncommon in low CAPE environments particularly at night and as long as substantial shear is present (e.g., Guyer and Dean, 2010; Sherburn et al., 2016; Nixon and Allen, 2022), so it is notable that the contributions reflect this physical attribute of tornadic environments.

Distributions of 3-h mean contributions by SHR500 show different patterns across the day-2 forecasts (Fig. 2.10) compared to the CAPE contributions. Overall, the mean values of the contributions by SHR500 are much lower than the 3-h distributions of CAPE in the day-2 hail forecasts (Fig. 2.10a). This difference is not too surprising, as SHR500 was not shown to be a particularly important predictor for the hail probabilities, particularly compared to CAPE (e.g., Figs. 2.3m; 2.5a). For the day-2 tornado probabilities, however, 3-h mean contributions by SHR500 are similar in magnitude to the 3-h mean CAPE contributions (Fig. 2.10b), though SHR500 contributions

peak at a slightly later time in the diurnal cycle (0000 UTC to 0300 UTC). The difference in the diurnal peak between the contributions by these two predictors is also apparent in the day-2 wind forecasts, where the mean 3-h contributions by SHR500 tend to be greatest between approximately 0000 UTC to 0300 UTC (Fig. 2.10c), but the mean 3-h contributions by CAPE tend to occur between 1800 UTC to 2100 UTC (Fig. 2.9c). While the largest mean contributions by SHR500 occur most often in the evening for all three hazards, there is a less clear diurnal cycle than there is in the CAPE contributions, as large SHR500 values are also somewhat common at other times of day. This finding is not too surprising from a physical perspective, as large-scale kinematic variables tend to have less variability between day and night compared to thermodynamic variables (with one exception being the nocturnal low-level jet).

2.3.3 Comparing contributions to GEFS fields

There are strong relationships between the raw GEFS data and the contribution values, though the nature of these relationships vary with variable and hazard being forecasted (Fig. 2.11). To start, correlation between the GEFS PWAT values and the TI contributions by PWAT shows an interesting pattern across each forecasted hazard type. For the day-2 hail forecasts, PWAT contributions seem have the capacity to be greatest for GEFS values around 30 mm (Fig. 2.11a). Above this threshold, the contributions have the tendency to become negative. The tornado contributions show a slightly similar pattern, though there is much less variability because the contribution values for tornado probabilities are much smaller (Fig. 2.11b). The wind contributions show a roughly opposite pattern from the hail contributions in the PWAT data (Fig. 2.11c), where the values of the contributions seem to minimize with a GEFS input of around 25 mm, then steadily become more positive thereafter. Perhaps these differences could be related to characteristics of wind-producing

versus hail-producing storms (e.g., large PWAT can limit updraft speeds due to hydrometeor loading, reducing hail growth potential). CAPE contributions and GEFS CAPE values show a more straightforward relationship, where negative contribution values are found in low-CAPE regimes, particularly for the hail and wind forecasts, then generally become positive above roughly 1500 J kg^{-1} (Fig. 2.11d,f). This pattern is less linear in the tornado contributions (Fig. 2.11e), which could be a function of there being frequent tornado probabilities in the Southeast (not shown), where high shear-low CAPE tornadic environments tend to be common (e.g., Sherburn and Parker, 2014). For CIN, the TI contributions appear to be negative or very near zero for the majority of grid points for day-2 forecasts of all three hazards, though the greatest positive contributions occur with low GEFS values of CIN (Fig. 2.11g-i). This pattern makes sense physically, as severe hazard-producing storms are often able to overcome small amounts of convective inhibition (e.g., "loaded gun" soundings) through lifting or surface heating, while large amounts tend to fully suppress severe convection. Lastly, contributions by SHR850, similar to those by CAPE, show a generally positive correlation with GEFS shear values (Fig. 2.11j-l), suggesting that larger values that are input to the system tend to lead to enhanced probabilities. Again, this pattern has consistencies with observed storm environments, where higher shear tends to favor more organized convection, which has greater tendency to produce severe hail, wind, and tornadoes.

While this analysis is limited to only a subset of the forecasts examined in this paper, it illustrates that there are profound ties between the GEFS data and the TI contributions. More importantly, the relationships between the two values largely follow properties of real-world expectations of severe environments (e.g., larger CAPE and shear signifies greater potential for severe weather). It is also worth taking note of the concentration of grid points with TI contributions that

are near or very near zero with highly variable corresponding GEFS values (Fig. 2.11). A hypothesis for this pattern is that the CSU-MLP system has learned about dependencies of the various environmental inputs on one another. For example, if the system is given 4000 J kg^{-1} of CAPE, but there is no accumulated precipitation and/or shear, perhaps it has learned that instability alone is not enough to support severe convection.

Knowledge of the ways in which ML systems such as CSU-MLP make their predictions and how it relates to the model data that it relies on offers valuable information to a forecaster. Specifically, knowing how CSU-MLP uses GEFS inputs can help a forecaster deduce whether the model output should be trusted for a given forecast: if a forecaster suspects a bias in a particular GEFS field, and they know how much that field tends to influence CSU-MLP probabilities, they can decide whether the ML output can be trusted.

2.3.4 Feature contributions for an example forecast

In addition to analyzing the contributions in aggregate, it is important to consider the information that TI can provide for a single CSU-MLP forecast, as fine details in the contributions over small spatial and temporal scales become illuminated. The example forecast and contributions shown in Fig. 2.2 are discussed here.

Overall, the location and magnitude of the of the SPC forecast and the CSU-MLP probabilities are similar for this case (Fig. 2.2a,b), with both products highlighting northern Louisiana, southeast Arkansas, and Mississippi for the greatest risk for tornadoes. The CSU-MLP probabilities are more spatially extensive compared to the SPC forecast, as the probabilities for the former extend farther north and west compared to the latter. While the greatest density of tornado reports ultimately occurred southeast of where the highest tornado probabilities were in both forecasts, both

the human and machine-learned forecasts still encompassed nearly all reports within the forecast period, and their predictions were skillful relative to climatology based on their positive BSS (Fig. 2.2a,b).

TI can dissect the CSU-MLP probabilities into contributions by each feature to contextualize how environmental information from the GEFS informed the prediction. As an illustration, Fig. 2.2c shows the feature contributions from each of the 15 environmental predictors for one forecast timestamp (2100 UTC) in the 1200 UTC to 1200 UTC day-2 forecast window. In this example, CAPE, 0-3km SRH, 10m-500 hPa shear, accumulated precipitation have the greatest overall influence on the forecast probabilities. Meanwhile, other predictors (such as precipitable water and LCL height) influence the probabilities very little according to their contributions (at least at this particular timestamp). Still, there is spatial variability in the sign of the contributions of many features, demonstrating that features can augment and offset each other to influence the final probabilities.For instance, there is a sharp gradient in the CAPE contributions to the forecast probabilities, where the variable is positively contributing to the forecast probabilities over east Texas, southern Arkansas, and much of Louisiana and Mississippi, while it is limiting the probabilities over areas northeast of this region.

2.4 Discussion

2.4.1 Resemblances to severe hazard climatology and environments

The results of the aggregate feature contribution analysis show that the CSU-MLP predictions rely most strongly on just a handful of predictors in the model to make its forecasts: CAPE, SHR500, SHR850, LCL height, and APCP (e.g., Fig. 2.5). Collectively, these five variables can

serve as indirect or direct measures of instability, lift, moisture, and shear, which are all fundamental large-scale ingredients required for severe convection (e.g., Doswell et al., 1996; Brooks and Craven, 2002; Brooks et al., 2003). In other words, the RFs generally get most information from just a handful of predictors, and the variables that it relies on most heavily to make these predictions are crucial for severe convection. While some predictors may influence the probabilities very little and add “redundant" information to the algorithm in a broad sense, they may provide key information in individual forecasts. Further, redundancy in the predictors does *not* imply that these environmental fields are physically unimportant for predicting severe weather–just that they are not really used by this particular algorithm.

Analyzing the feature contributions in aggregate also showed that there is some variability in how salient each of the key predictors are to the CSU-MLP probabilities in terms of hazard type, time of year, and time of day. For instance, thermodynamic variables tended to be more important to the RFs in the warm season compared to the cool season when kinematic variables tended in influence the probabilities more (Fig. 2.4). This seasonal variability is consistent with previous studies that have illustrated that dynamically-dominant regimes (i.e., “high shear, low CAPE") more frequently characterize cool season events compared to the warm season (Sherburn and Parker, 2014; Sherburn et al., 2016). Another example shown in the results is that the RFs depend more heavily on LCL for generating its tornado probabilities compared to other hazards (Fig. 2.5), which is in agreement with previous work that has shown it to be crucial for diagnosing tornado environments but is perhaps not as useful for other hazard types (e.g., Brooks and Craven, 2002; Brooks et al., 2003; Thompson et al., 2012; Nixon et al., 2023). In terms of diurnal variability, the aggregate feature contributions showed that thermodynamic features (such as surface-based

CAPE; Fig. 2.3) tended to contribute more to probabilities during the day compared to the night, whereas kinematic variables (such as SHR500; Fig. 2.3) did not have such a prominent diurnal cycle. These differences in the diurnal pattern amongst kinematic versus thermodynamic variables in the feature contributions is similar to patterns found in the feature *importances* associated with the CSU-MLP system (Hill et al., 2023), illustrating that this finding exists across multiple explainability metrics. This result is also consistent with what one might expect physically, where surface-based instability shuts down at night with boundary layer cooling, large-scale forcing remains relatively unchanged.

This sub-daily (3-h) disaggregation TI revealed one of the more remarkable results of this study. In Fig. 2.7, it was shown that the contributions have a discernible diurnal cycle, with larger contributions occurring with greater frequency during the late afternoon through evening and fewer sub-daily contributions occurring in the late night through mid-morning. This variability was found to some extent across all lead times and hazard types in ways that are mostly consistent with the diurnal patterns in the reports during the study period (Fig. 2.6). This result is notable because there is no information that denotes time of day in the CSU-MLP system's training: that is, the RFs are not "aware" of the timestamps in the GEFS fields nor in the storm reports. Yet, because the system has learned relationships between atmospheric conditions and observed severe weather, it is able to recognize that conditions at certain times of day are generally more favorable for severe weather hazards than others.

Patterns in the 3-h contributions were also found to vary spatially in ways that are consistent with previous literature on severe storms climatology. For example, sub-daily tornado contributions in parts of the Deep South tended to exist during the overnight and morning hours, whereas

other locations (such as the Plains) did not see the 3-h contributions at these times of day. These differences in likelihood for daytime versus nocturnal tornadoes between various locations have been documented in the literature both in terms of reports (e.g., Krocak and Brooks, 2018) as well as environments (e.g., Kis and Straka, 2010). Additionally, while 3-h hail contributions tended to maximize later in the night than the diurnal maximum in hail reports, it is possible that report biases could be contributing to this discrepancy. Radar-derived hail estimates suggest that the actual diurnal maximum in reports may be slightly later in some locations due to undereporting biases after dark (Wendt and Jirak, 2021), and so it is possible that the system has learned the larger-scale ingredients that tend to favor after-dusk hail-producing storms, despite the fact that there tend to be fewer reports during this time of night.

Relatedly, because CSU-MLP is trained on storm report observations, there may be limits on how well the system is able to capture the “truth" in its probabilities because of the underlying issues that exist within the storm events database. Some of these issues (particularly with hail and wind reports) can be tied to population biases, spatial variability in observing platform siting, and varying land surface conditions just to name a few (e.g., Doswell et al., 2005; Smith et al., 2013; Allen and Tippett, 2015; Edwards et al., 2018; Bunkers et al., 2020; Wendt and Jirak, 2021). Patterns in the CSU-MLP probabilities reflect some of these reporting characteristics, such as the very high frequency of wind probabilities in the Mid-Atlantic (where dense population and vegetation favor frequent *estimated* gusts) compared to the Great Plains (which is where *measured* severe wind most often occur; e.g., Doswell et al., 2005; Smith et al., 2013; Edwards et al., 2018; Bunkers et al., 2020). While these issues are well-known in the meteorology community, progress towards managing these characteristics of estimated severe wind gusts is being made (Tirone et al.,

2024). It is important to be aware of the observation characteristics when interpreting ML-based weather predictions like CSU-MLP, as issues in the model input cascade into issues in the model output.

2.4.2 Utility in operational forecasting

Understanding the TI feature contributions in aggregate can have useful applications to operational forecasting. This work has demonstrated that the CSU-MLP probabilities are made using learned statistical relationships amongst the meteorological predictors that are consistent with characteristics of real-world environments. This evidence can offer trust and confidence that the ML forecasts are generated in ways that follow scientifically meaningful patterns rather than relationships that are inconsistent with physical processes. Additionally, this work suggests that the CSU-MLP predictions are largely reliant on just a handful of predictors. Awareness of these characteristic variables could help a forecaster parse down the number of key environmental fields they might want to consider when examining the CSU-MLP output alongside the raw GEFS fields. More broadly, these results show that there is a vast amount of information for severe weather forecasting that can be extracted from the large-scale environment alone (and without CAM data), and ML is capable of skillfully learning these patterns and their nuances.

Applying TI to individual forecast probabilities offer several benefits to the forecaster. First, in a real-time forecasting setting, disaggregated contributions allow a forecaster to “interrogate" the model’s prediction to gain insights on otherwise vague model probabilities. The forecaster can use the contributions to see how the probabilities are constrained and enhanced by different features across various locations and times, allowing them to see which variables are most relevant to the prediction. This approach presents itself similarly to a familiar technique to forecasters,

ingredients-based forecasting (e.g., Johns and Doswell, 1992; Doswell et al., 1996; Brooks, 2007), where each environmental variable (or ingredient) known to be important for a particular weather phenomenon are taken into consideration to make a forecast. The forecaster can use contributions to dissect a machine learning forecast in similar ways that they analyze an NWP forecast. This information may be particularly advantageous at longer lead times, when a forecaster has less information available. Second, contributions can be used by forecasters as a confidence-booster or checkpoint on their forecast intuition. For instance, if the contributions are agreeable with other data, they can provide additional confidence in the forecast. If the contributions disagree with a forecaster's thinking, they could offer motivation for looking more closely at certain variables in the NWP models. Third, using the contributions over time–especially in retrospective event analyses–could also help a forecaster identify biases in the model probabilities and allow them to use it more effectively. Lastly, forecasters can use TI contributions to assess whether a ML model is making a prediction in ways that are consistent with physical intuition, which could enhance their trust in the output.

2.5 Summary and conclusions

The work presented in this paper leverages tree interpreter (TI), a local explainable artificial intelligence (xAI) method, to investigate how severe weather forecasts are made at lead times of 2 to 4 days. TI is used to disaggregate daily probabilistic forecasts for severe weather hazards (hail, tornadoes, and convective wind) generated by the Colorado State University Machine Learning Probabilities (CSU-MLP) system: a random forest (RF)-based model that has undergone thorough evaluation in testbeds and is now a fully operational forecasting product. The daily probabilistic forecasts are disaggregated in time, space, and by variable type into "feature contributions".

When the TI-generated feature contributions are analyzed in aggregate across two years of forecasts, the results show that the RFs are able to learn that certain meteorological fields from the Global Ensemble Forecast System Reforecast Dataset (GEFS/R) have a crucial role in influencing severe weather probabilities generated by the model. These fields (such as CAPE, wind shear, and LCL height) are also known via observations to be key in generating severe weather, so the model is able to recognize patterns that are consistent with physical expectations. Further, patterns in the feature contributions over this 2-yr period vary significantly in time and space in ways that resemble severe storm climatology and environments. These results provide confidence that the CSU-MLP system is learning scientifically-meaningful relationships, despite the absence of high resolution data and convection-permitting information in the model.

For individual daily forecasts, disaggregating the CSU-MLP probabilities by feature can illustrate how each meteorological input variable is acting to increase or constrain the forecast probabilities in time and space throughout the prediction period. Given these properties, as well as the fact that the feature contributions generated by TI sum to the daily forecast probabilities, this approach can be thought of as an ingredients-based forecasting approach (Brooks, 2007), but it is applied to the output of a machine learning model, rather than a numerical weather prediction model.

From an operational forecasting perspective, this work demonstrates that TI can supply insightful information on machine learning guidance across a variety of applications. For individual forecasts, TI can provide an ingredients-based perspective for the probabilistic guidance from CSU-MLP. At extended lead times (e.g., for forecast periods beyond the current availability of typical convective-allowing models) when there are less data available to consider, TI could provide

valuable information to a forecaster considering the ML-based output in real-time. This information could also be helpful in post-event analysis, allowing a forecaster to look back on the ingredients that most influenced the ML guidance for a given forecast period, and those fields can be compared to what was seen in observations. Finally, demonstrating how the CSU-MLP forecasts are generally made by analyzing the TI feature contributions in aggregate could help build a forecaster's trust in the model as a whole, particularly considering that the model's predictions tend to be consistent with physical patterns and relationships. Given the familiarity and use of the CSU-MLP guidance across much of the weather forecasting community, a deeper understanding of the probabilities put forth by the model is likely to be of interest to those who use it.

Future work will investigate the connections between attributes of the TI feature contributions (e.g., magnitude and sign) and actual values of the meteorological parameters that are used in the GEFS/R. If the feature contributions are correlated with the values of the meteorological predictors in ways that make physical sense (e.g., high positive CAPE contributions being correlated with high CAPE values), it would build confidence that CSU-MLP is learning meteorologically-relevant patterns and further motivate its use as an ingredients-based forecasting tool that could be applied to other ML guidance. Potential connections between TI feature contributions and model skill could also be explored. Lastly, near-term research aims to diagnose regimes amongst the CSU-MLP forecasts that are tied to successful and unsuccessful performance. That work, in addition to guidance from the TI contributions, would provide an additional layer of transparency to the CSU-MLP probabilities, allowing a forecaster to better understand how heavily the model output should be taken into account in their day-to-day forecasting practices.

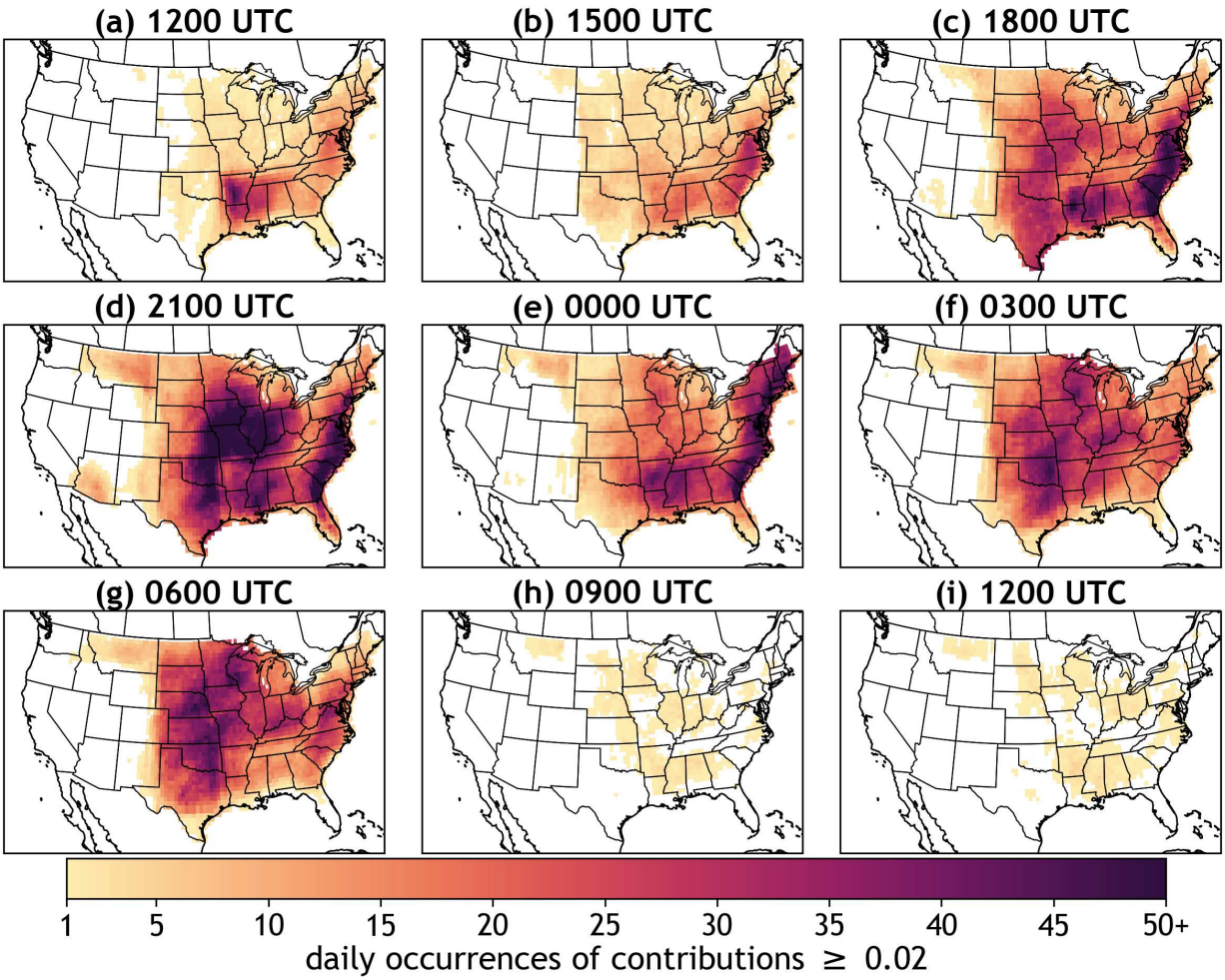

Figure 2.8: As in Fig. 2.7, but for the 3-h contributions exceeding 0.02 in the day-4 CSU-MLP severe forecasts. Heatmaps are shown for the daily frequencies of the 3-h contributions exceeding 0.02 at forecast periods ending at forecast hour (a) 84 or 1200 UTC, (b) 87 or 1500 UTC, (c) 90 or 1800 UTC, (d) 93 or 2100 UTC, (e) 96 or 0000 UTC, (f) 99 or 0300 UTC, (g) 102 or 0600 UTC, (h) 105 or 0900 UTC and (i) 108 or 1200 UTC within the day-4 period.

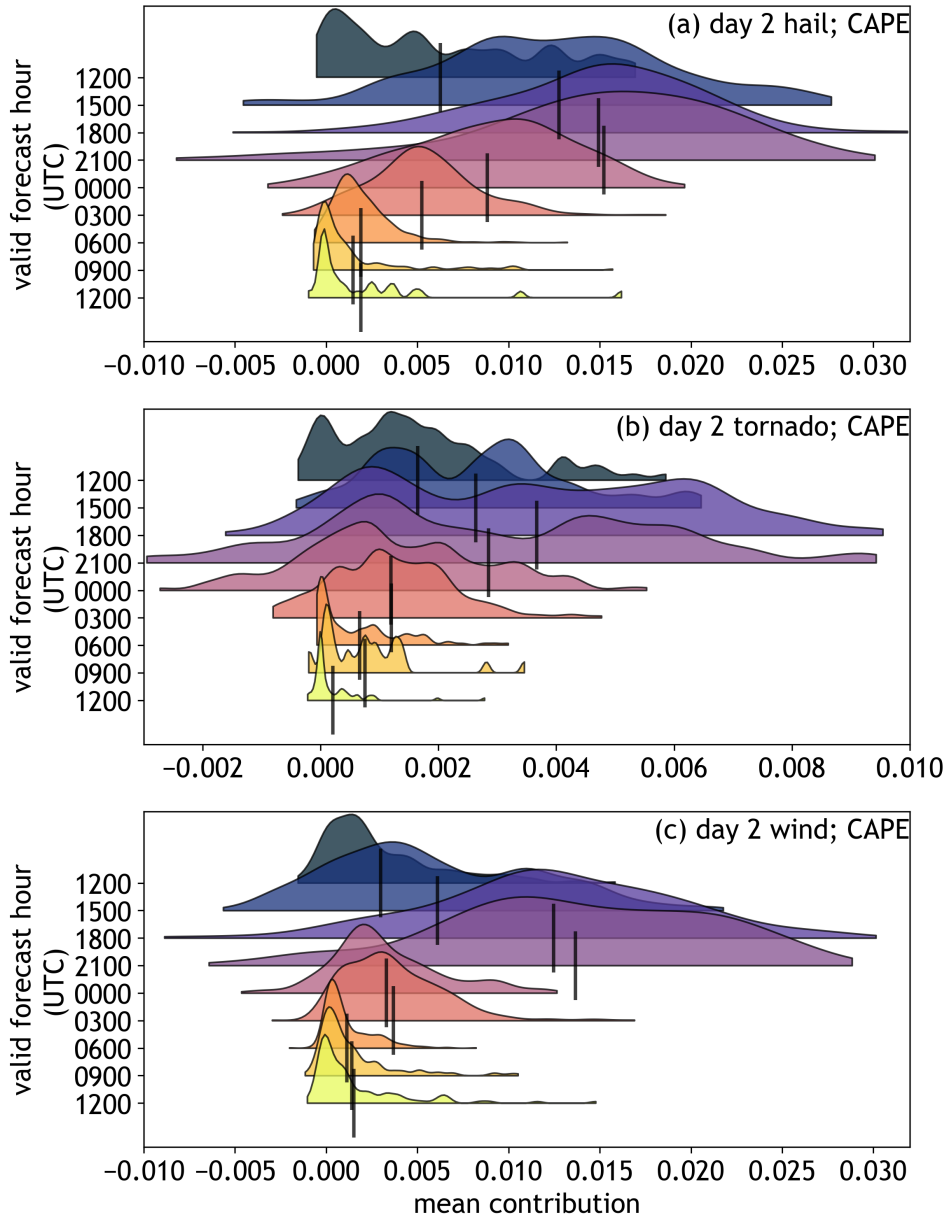

Figure 2.9: Gaussian kernel density estimator (KDE) plots illustrating the smoothed distributions of the sub-daily mean 3-h CAPE contributions to the CSU-MLP day-2 (a) hail, (b) tornado, and (c) wind forecast probabilities over the January 2021 to December 2022 period. Forecast timestamps and KDE plots are listed in chronological order from the top to the bottom of the plots. That is, the KDE plot on the upper side of the figure represents the distribution of the daily mean CAPE contributions at the earliest valid timestamp (1200 UTC, dark gray) in the day-2 period, followed by the distribution at the next valid timestep (1500 UTC, dark blue), and so on. Vertical black bars along each KDE plot mark the mean of the distribution of mean CAPE contributions at each 3-h forecast period. Note the differences in scaling along the x-axis.

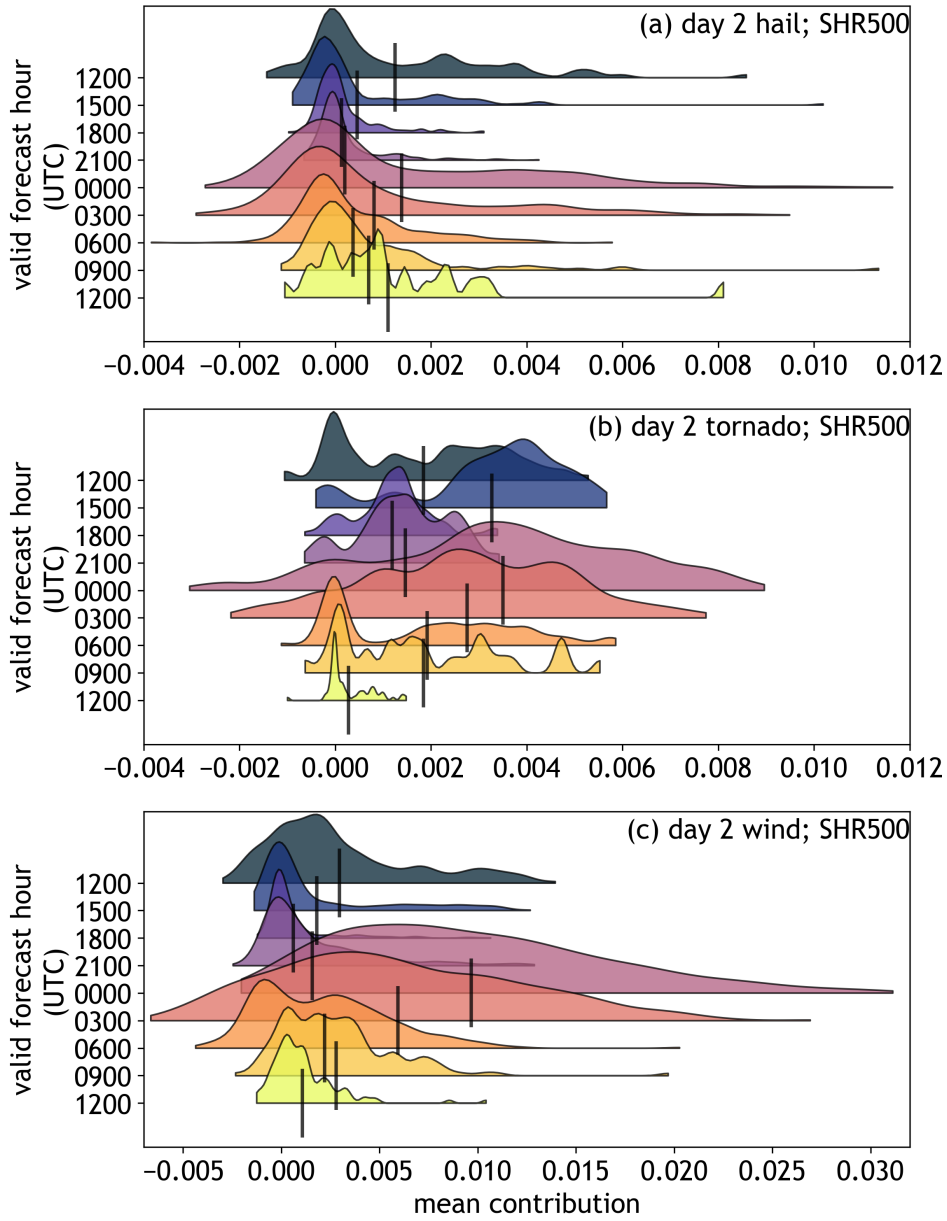

Figure 2.10: As in Fig. 2.9, but for 10m to 500 hPa bulk wind shear (i.e., SHR500).

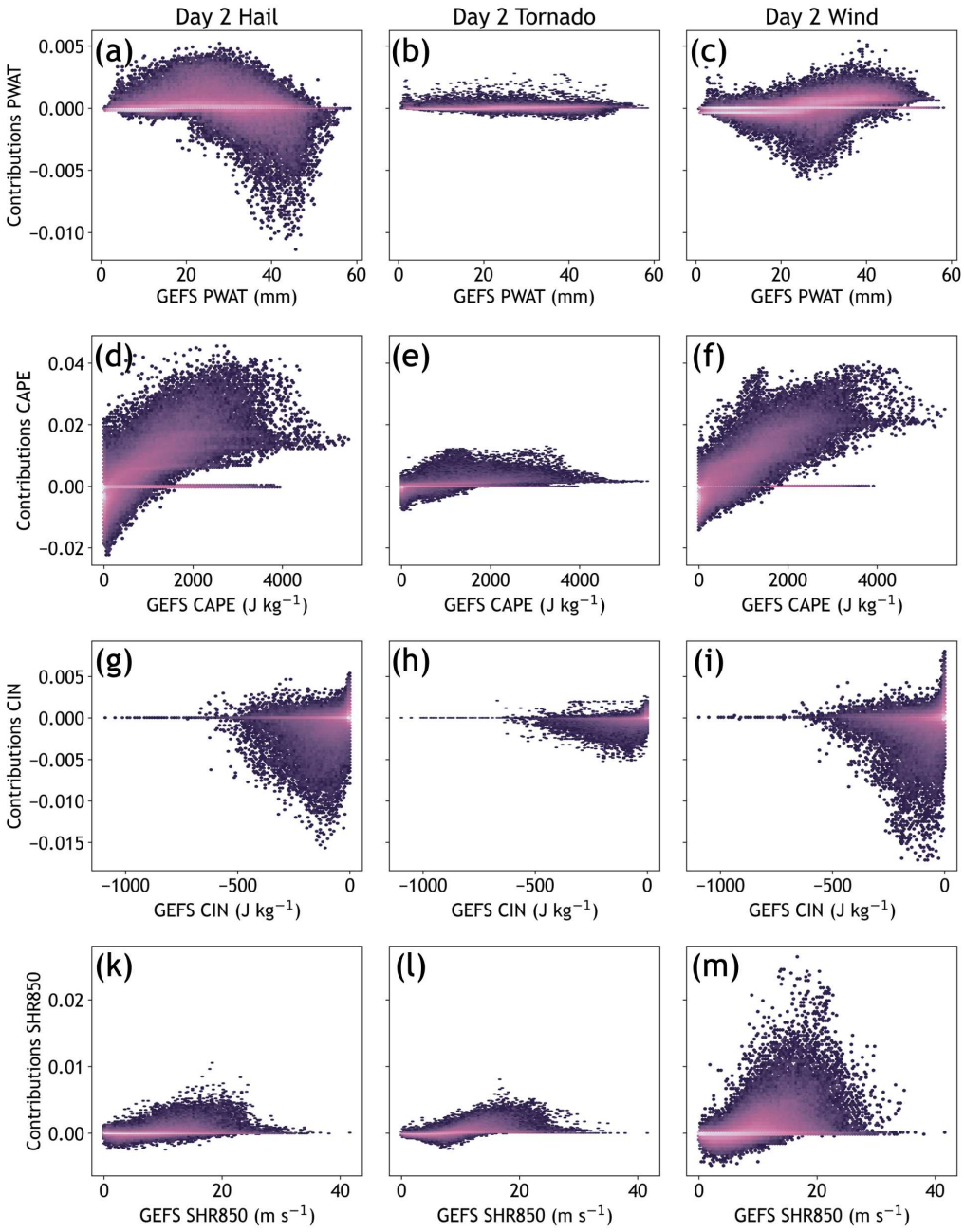

Figure 2.11: Hexbin plots comparing operational GEFS forecasted values (x-axes) versus TI contributions (y-axes) of (a)-(c) PWAT, (d)-(f) CAPE, (g)-(i) CIN, and (j)-(l) SHR850 on a grid point-by-grid point basis. Specifically, day-2 GEFS forecasts valid at hour 45 (i.e., 2100 UTC in the day-2 period) are compared to the CSU-MLP contribution values valid for 2100 UTC for the day-2 hail (left column), day-2 tornado (center column), and day-2 wind (right column) forecasts. Data are only shown for forecasts valid for 1 March to 31 May 2021 and 2022 . The number of values in each bin is on a log scale, with the lighter colors representing a larger number of points falling within that bin.

Chapter 3

Investigating Skill of Probabilistic Severe Weather Forecasts Across Self Organizing Map-Diagnosed Regimes

3.1 Introduction

Despite improvements to modeling systems, communications, and forecasting practices, predicting severe weather remains a complex forecast challenge. Among the plethora of upgraded modeling systems and guidance products that have contributed to improved forecasting capabilities, machine learning (ML) has become an increasingly relied-upon tool for forecasting convective-related hazards (McGovern et al., 2023). At convective time scales, accurate and timely forecast information is of utmost importance, and ML has shown promise in accomplishing both of these needs. Unlike the requirements of numerical weather prediction (NWP)-based models, ML approaches typically rely on data (either from past forecasts, reanalysis, and/or observations). By side-stepping the need to solve hundreds of complex prognostic equations to produce a forecast, ML-based weather prediction systems also typically also offer speed advantages over traditional methods. Recent work suggests that it may even be possible to effectively model storms at convection-allowing model scales with these tools (e.g., Flora and Potvin, 2024; Pathak et al., 2024).

While these numerical weather prediction (NWP)-emulating systems, particularly at fine resolutions, are relatively new to the community, other ML-based forecasting products that focus on convective weather prediction have been around for longer. Among the tools that have been developed are probabilistic ML-based forecasts. These types of systems typically rely on two key

datasets: 1) atmospheric variables (such as from NWP models, observations, and/or reanalysis) and 2) observations of some convective weather phenomenon. Such ML models are trained to learn statistical relationships between the “predictors"–i.e., the input atmospheric data (such as humidity or temperature)–and the “labels" (i.e., the severe weather observations). Once satisfactorily trained on historical data, the ML model can then be used to make future severe weather predictions using new atmospheric data. Probabilistic ML forecasts have been developed for all kinds of convective weather prediction tasks, including tornadoes, hail, strong winds, and heavy rainfall (e.g., Lagerquist et al., 2017; Herman and Schumacher, 2018a; Loken et al., 2019; Hill et al., 2020; Hill and Schumacher, 2021; Flora et al., 2021; Schumacher et al., 2021; Clark and Loken, 2022; Loken et al., 2022; Hill et al., 2023).

ML is undoubtedly changing the landscape of forecast products that are available for operational weather prediction. However, another (perhaps quieter) way that it has been used in the context of convective weather is to better understand their environments. Self-organizing maps, or SOMs (Kohonen, 1982), are one ML method that has been used for these purposes. SOMs are a type of artificial neural network that employ competitive learning, which is an unsupervised ML technique. SOMs are powerful tools for conducting dimensionality reduction and clustering on datasets that are described by many parameters. Thus, they can be useful for identifying and visualizing patterns in complex multivariate datasets.

SOMs have been used for at least two decades to describe and categorize atmospheric patterns (Hewitson and Crane, 2002), offering an alternative to manual classification that had been done for many decades prior. For convective storms environments specifically, SOMs have been used in a number of studies. For example, Nowotarski and Jensen (2013) and later Nowotarski and Jones

(2018) used SOMs to classify proximity soundings in an effort to distinguish between tornadic and non-tornadic supercell environments. Anderson-Frey et al. (2017) evaluated relationships between variability in environments and attributes of associated tornado warnings and reports using SOMs. Hua and Anderson-Frey (2022) also examined characteristics of tornadic environments with SOMs, though their work focused on describing diurnal variability in such environments across different locations and seasons. Meanwhile, Nixon et al. (2023) leveraged SOMs to classify different types of hail environments. SOMs have also been used to describe variability in environments associated with specific convective storm types, including right-moving supercells (Warren et al., 2021) and mesoscale convective systems (Song et al., 2019).

Probabilistic ML-based severe weather forecasts and SOMs both have relevance to the research conducted here: in this work, output from a ML-based probabilistic forecasting model, the Colorado State University Machine Learning Probabilities (CSU-MLP) system, is studied with the help of SOMs. The CSU-MLP system was originally developed as a tool for aiding in predictions of excessive rainfall (Herman and Schumacher, 2018a; Schumacher et al., 2021), eventually becoming an operational product used at the NOAA Weather Prediction Center. Soon after, the model architecture was leveraged to predict severe weather events–tornadoes, severe wind, and hail– in support of needs at the NOAA Storm Prediction Center or SPC (Hill et al., 2020, 2023). These probabilistic severe weather forecasts, which will be the focus of this work, are now also operational at NOAA. Additional details on this model can be found in the methods section.

The CSU-MLP severe probabilities have been shown to be skillful (Hill et al., 2020, 2023), though additional work is needed to support forecaster confidence and trust in its predictions.

Recent work by Mazurek et al. (2025) (Chapter 2 herein) contributed to these efforts by showing that the CSU-MLP system uses simulated environmental information to make its forecasts in ways that agree with the physical and climatological characteristics of severe storm environments, demonstrating that the predictions align with conceptual understandings of the atmosphere. *This work furthers the goal of optimizing CSU-MLP's use in operations by using SOMs to investigate the skill of their forecasts across different environments.* It builds on previous work by Escobedo (2022) (Escobedo and Schumacher, 2024, submitted to Wea. Forecasting), which examined performance of the CSU-MLP excessive rainfall forecasts across different regimes, though this study employs SOMs (rather than hand-labeling) for regime identification and examines severe weather environments (rather than excessive rainfall environments). SOMs have been shown to be a useful tool for studying relationships between forecast skill and associated environments in physics-based models (e.g., Kolczynski and Hacker, 2014; Radford and Lackmann, 2023a), and it is hypothesized that they could be useful for studying ML-based forecast skill as well.

The overarching objective of this work is *to establish an understanding of which type of synoptic patterns tend to be associated with good versus poor-performing CSU-MLP forecasts.* SOMs are used here as a regime diagnostic tool to describe the various ambient synoptic and mesoscale characteristics that were in place during two years of day-2 CSU-MLP probabilistic severe weather forecasts. Relationships between the different environmental conditions that are present in each regime are compared to the skillfulness of the model's tornado, hail, and wind predictions. The broader purpose of this work is to identify atmospheric patterns that correlate to strong and weak predictability in the CSU-MLP system, which will allow a forecaster to use it more effectively as guidance.

This chapter proceeds as follows. Section 2 provides background on the CSU-MLP system, SOM development, and forecast evaluation techniques. Section 3 highlights the results, where characteristics of the SOM-identified regimes and forecasts that fall within each of them are provided. Forecast skill of the CSU-MLP probabilities are also discussed across the various regimes. In section 4, the findings in section 3 are put into context to describe overall strong and weak points of CSU-MLP predictability. Future work and a summary of the project are provided in section 5. Additional figures for this work can be found in Appendix A at the end of this dissertation.

3.2 Methods

3.2.1 CSU-MLP system overview

The CSU-MLP system is a random forest (RF)-based (Breiman, 2001) machine learning model that produces probabilistic forecasts of tornadoes (of any magnitude), severe hail (in excess of 2.54 cm or 1 inch diameter) and severe wind (stronger than 93 km h^{-1} or 50 knots) at one-through eight-day lead times. In essence, these probabilistic forecast products are designed to resemble the SPC's convective outlooks (i.e., predicting the probability of tornadoes, severe wind, and/or severe hail within 40 km of a location) and provide forecasters with an objective "first-guess" at where severe hazards may occur. The products have undergone several years of evaluation in NOAA's Hazardous Weather Testbed (e.g., Clark et al., 2021, 2023) as well as internally at the SPC (personal comm. with Israel Jirak) and are now operational at NOAA. Throughout this evaluation period, forecasts have been used in operations by SPC and local Weather Forecast Offices (since at least early 2022; personal comm.). The CSU-MLP probabilistic forecasts became operational in spring 2024.

The system is trained on nine years (spanning April 2003 to April 2012) of environmental fields from the Global Ensemble Forecast System v12 Reforecast (GEFS/R) dataset (Hamill et al., 2022)

and gridded observed hail, wind, and tornado reports from NOAA's Storm Data (NOAA Storm Prediction Center, 2023a; NOAA National Centers for Environmental Information, 2023). These datasets are used as the predictors and training labels, respectively, to train separate RF models at each of the eight lead times. At day 1 through 3 lead times, three separate models are trained to predict separate tornado, hail, and wind probabilities (Hill et al., 2020)[6], whereas a single model is trained to predict "aggregate severe" forecast probabilities at days 3 though 8 (Hill et al., 2023). The day 1-3 models share the same environmental predictors with each other, while the day 3-8 models are trained on a slightly smaller subset of these predictors (see Table 1 in Mazurek et al. (2025)). Predictors were informed via formal sensitivity tests conducted in its predecessor model for excessive rainfall prediction (Herman and Schumacher, 2018a; Schumacher et al., 2021) as well as informally with early iterations of the severe weather version of the system (Hill et al., 2020). The day 3-8 models use a smaller number of predictors than the day 1-3 models as testing showed that the additional predictors did not enhance model skill (personal comm. with Aaron Hill).

Real-time forecasts are run twice-daily[7] at 0000 UTC and 1200 UTC using forecasted environmental fields (the same variables as those used in training) from the operational Global Ensemble Forecast System v12 (Zhou et al., 2022). These probabilistic forecast products are designed to resemble the SPC's convective outlooks (i.e., they predict the daily probability of tornadoes, severe wind, and/or severe hail within 40 km of a location) and provide forecasters with an objective "first-guess" look at where severe hazards may occur. Additional details of the CSU-MLP architecture (including data, preprocessing, and predictor assembly techniques) can be referenced in

[6]Forecasts of individual hazards for days 2 and 3 have been added since the publication of Hill et al. (2020).

[7]The day-1 forecast is only run at 0000 UTC.

Hill et al. (2020) (see (Mazurek et al., 2025) for more recent updates to the day 1-3 forecasts) and (Hill et al., 2023).

Two years (January 2021 to December 2022) of day-2 CSU-MLP probabilistic severe weather forecasts of tornadoes, severe hail, and severe wind–encompassing lead times of 36 to 60 hours from the 0000 UTC GEFS initialization–are the focus of this work. This lead time was selected for a few reasons. First, the day-2 forecast period encompasses lead times that near the limits of present operational convective allowing model guidance, where model skill tends to degrade, promoting the utility of the CSU-MLP system at these lead times in particular. Additionally, at shorter lead times (i.e., day 1), there already exists a crowded field of skillful forecast guidance that can offer more detailed information on where severe hazards are likely than the CSU-MLP output can. At longer lead times, CSU-MLP forecast skill degrades, and forecast probabilities that exceed the minimum SPC-defined threshold (15% for days 4-8) become fewer in number, reducing the sample size of forecasts viable for analysis. Thus, day-2 forecasts are selected in an effort to probe a “sweet spot" in the model’s predictions. Plus, it has been shown that the trained models for days 2 and 3 tend to use environmental information in similar ways (Mazurek et al., 2025), which suggests that the findings surrounding the day-2 forecasts in this study would potentially be applicable to day 1 and 3 forecasts that utilize the same model architecture.

3.2.2 Self organizing map (SOM) development

This work uses the SOMPY library (Moosavi et al., 2014) to train the SOMs in Python[8]. Vesanto et al. (1999) and Nowotarski and Jensen (2013) provide thorough overviews of how SOM algorithms work, but a brief overview is provided here, along with specifications on this work’s

[8]SOMPY models the algorithm of the long-running SOM toolbox in Matlab (Vesanto et al., 1999).

specific training approach (using details from the SOMPY documentation and source code). SOM training generally begins with defining a hexagonal or rectangular lattice (that is typically two-dimensional) with some number of user-specified nodes or neurons and their orientation. Each node is initialized with a weight vector with dimensions that match the input data. Various initialization methods exist, but this work uses random initialization, where the initial weights represent random values that fall within confines of the input dataset. Then, an input data vector is passed through the SOM and compared to the current weight vectors at each node via some distance calculation (Euclidean distance is most common and used here). The “winning node" or Best-Matching Unit (BMU) is the node that is the shortest Euclidean distance from the input vector. The SOM learns by shifting the weight vector at each node towards the input vector values. Because SOMs use a competitive learning approach, the weight vector at the winning node is nudged the most, and weights at neighboring nodes are nudged only slightly. The degree of nudging outside the winning node is determined by a neighborhood function (a Gaussian is used here). A batch training approach is used in this work, so all input vectors are seen by the SOM at the same time, and the weights are updated together (rather than over each input, as is done in sequential training). The distance-weighted nudging is repeated many times in two training phases (rough and fine-tuning). At the conclusion of training, the SOM sorts all of the input vectors into their appropriate final-weighted BMUs.

SOMs are used as a regime classification technique in this work. Using SOMs for this purpose allows for regimes to be sorted more quickly and statistically (which are two advantages over subjective manual classification). Two SOMs are trained, and their configuration is kept rather

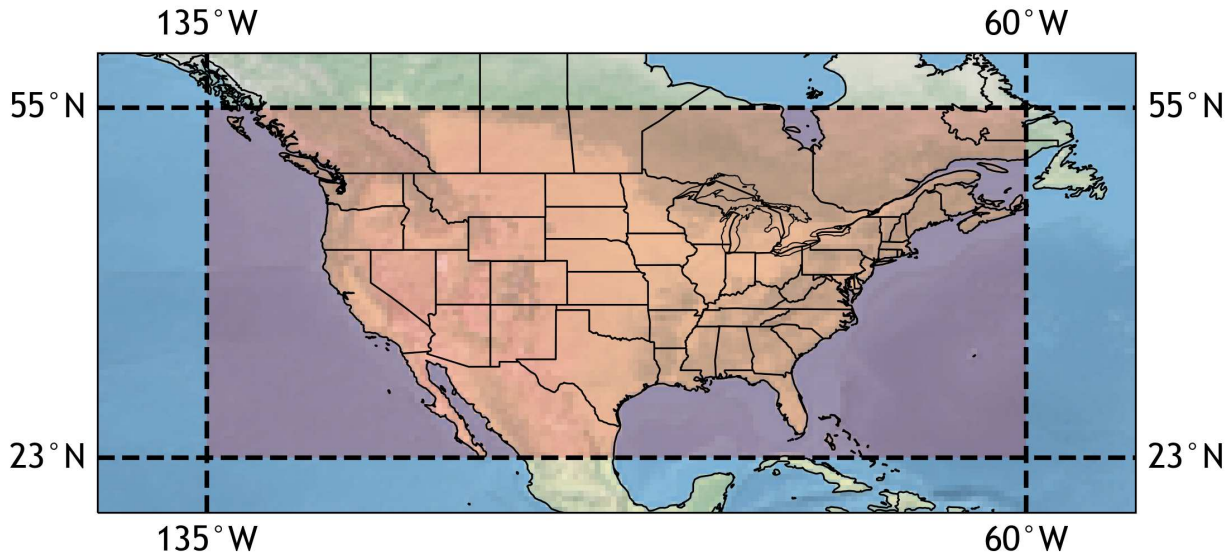

Figure 3.1: The SOM training area.

simple. SOMs with few variables have shown to be successful in previous work (e.g., Anderson-Frey et al., 2017). Training two SOMs allows for the analysis and conclusions to be interpreted somewhat more broadly by illustrating how sensitive the results might be to the SOM configuration (i.e., it would be difficult to make major conclusions about the relationship between environmental characteristics and forecast skill if the regimes were vastly different between the two SOMs).

Both SOMs utilize fields from ERA-5 reanalysis (Hersbach et al., 2020): one is trained on a combination of surface-based convective available potential energy (SBCAPE) and 10m to 850 hPa shear (SHR850), and the other is trained on SBCAPE and 10m to 500 hPa shear (SHR500). These SOMs will be referred to as SOM0 and SOM1 respectively. SBCAPE and shear fields are selected for SOM training because they are well-known to be associated with severe convection when they are co-located with each other (e.g., Brooks et al., 2003; Taszarek et al., 2020). Training on SOMs with a large number of variables (e.g., adding precipitable water, mean sea level pressure, and others) seemed to lead to less-clearly defined regimes during informal testing, thus these two key

variables are relied on here. Because CSU-MLP forecasts only exist over the contiguous United States (CONUS), and this work focuses on synoptic characteristics of the environment, the SOM training region is restricted to areas encompassing CONUS and slightly beyond (Fig. 3.1). One ERA-5 timestamp, valid at 2100 UTC, is considered for each valid date in the two-year analysis period. The 2100 UTC timestamp is used in an effort to best align the reanalysis data with the timing of when the diurnal maximum in severe storms tends to occur over CONUS (e.g., Krocak and Brooks, 2020).

There is some preprocessing done to the SBCAPE and shear data before SOM training. First, because shear is not a native variable in the ERA-5 dataset, bulk wind shear difference is computed over the 10m-850hPa and 10m-500 hPa layers. Because these parameters can be noisy, a small amount of Gaussian smoothing is applied to smooth the fields spatially[9]. Then, daily standardized anomalies are computed for the two years of data in the analysis period. Anomalies are computed relative to the 30-yr (1991-2020) ERA-5 reanalysis climatology of these variables. The fields used in the daily 30-yr mean and standard deviation calculations have the same spatial smoothing as the fields in the two-year analysis period. Additionally, some temporal smoothing is also applied to the 30-yr climatological fields (via taking a 15-day moving average). Anomalies are computed in this way to remove seasonality from the environmental fields. Finally, because SOMs require each of their training cases to be one-dimensional data vectors, each of the daily standardized anomalies of SBCAPE and shear (which are two-dimensional arrays that are a function of latitude and longitude) are reshaped into one-dimensional arrays.

[9]The smooth gaussian function from MetPy (May et al., 2022) is used, and the degree of smoothing used here is subjectively set to 2, which adjusts the standard deviation of the Gaussian to determine the degree of smoothing (2 is very little smoothing). Additional details can be found in the function's documentation.

A summary of the parameters used to train the SOMs is provided in Table 3.1. A number of sensitivity tests were conducted to tune the parameters, such as training length and training radius, in an effort to maximize the SOM skill. Two metrics were used to test SOM skill in this tuning phase. The first is topographic error, which is a measure of how well the SOM retains the original data structure by computing the proportion of inputs (data vectors) for which the best node and second best node are not adjacent to each other on the SOM (Kiviluoto, 1996; Moosavi et al., 2014). Quantization error captures how much (on average) the data vectors differ from their winning nodes on the SOM. It is optimal to keep both of these scores as small as possible. Some of these sensitivity tests can be found in Appendix A. These tests illustrate that changes to the random seed, training radii, and training length (beyond a certain number of epochs) generally did not yield substantial differences in quantization and topographic errors, suggesting that the SOM regime diagnostics are not highly sensitive to these parameters. Still, this testing did help with adjusting some of the parameters and with building confidence in the SOM approach. For instance, sensitivity tests showed that a smaller fine-tune training radius generally favored more skillful results, though the effects of the rough training radius on skill were less clear[10].

The trained SOMs use a rectangular 3x3 lattice structure. This training architecture classifies the ERA-5 reanalysis data into 9 nodes or regimes. As shown by Anderson-Frey et al. (2017), there are advantages and drawbacks to increasing or decreasing the number of nodes: more nodes can elucidate additional details on regime characteristics but could ultimately results in nodes with redundant patterns, whereas fewer nodes reduce the risk of redundancy but may smooth the regimes

[10]In the literature, it is generally recommended that SOMs use a wide radius during rough training and a smaller radius during fine-tune training (e.g., Kohonen, 1990; Vesanto et al., 1999). Because our data are standardized and our SOM is relatively small (Kohonen, 2013), these numbers are kept relatively small.

Table 3.1: Parameters used to train the SOMs.

parameter	value
number of nodes	3x3 (9 total)
lattice structure	rectangular
neighborhood	Gaussian
rough training length	50
fine-tune training length	10
rough training radius	3
fine-tune training radius	1

too much. In this study, 9 nodes subjectively seemed to strike a reasonable balance across these considerations.

3.2.3 Assessing regime characteristics and forecast skill

While all two years of ERA-5 reanalysis data are included for the training of the SOMs and the construction of the regimes, “null" cases are removed after the SOM regimes are identified. Null cases are taken to be forecast days when the maximum CSU-MLP system probabilities are less than 15% for hail *and* wind *and* less than 5% for tornadoes, which is the “slight" risk probability threshold for severe storms as defined by the SPC. The daily maximum CSU-MLP probabilities for *all three* hazards must be less than this threshold in order for the case to be removed. Removing null cases allows for the analysis to be restricted to cases when the CSU-MLP system predicted that severe weather was likely (rather than when the model probabilities suggested that it was unlikely).

This approach resulted in a sample size of 381 cases over the two-year period. Table 3.2 illustrates the number of cases that were sorted into each node before and after null cases were removed. There is substantial variability in the fraction of cases that are removed from each node anywhere in CONUS, though the majority of nodes retain a reasonable number of cases for compositing. It

Table 3.2: Summary of the number of total cases and non-null cases in each node, as well as the percentage of cases that were retained after removing null cases.

node number	case count (SOM0)	non-null case count (SOM0)	percent of cases retained (SOM0)	case count (SOM1)	non-null case count (SOM1)	percent of cases retained (SOM1)
node0	107	83	77.6%	115	70	60.9%
node1	57	33	57.9%	59	30	50.8%
node2	161	61	37.9%	151	41	27.2%
node3	44	34	77.3%	46	38	82.6%
node4	21	11	52.4%	20	13	65.0%
node5	54	16	29.6%	81	29	35.8%
node6	92	76	82.6%	100	82	82.0%
node7	67	30	44.8%	42	30	71.4%
node8	124	37	29.8%	113	48	42.5%
total	727	381	52.4%	727	381	52.4%

is worth noting that this methodology inherently focuses the analysis on hits and false alarms of the CSU-MLP system rather than its misses and correct negatives.

While only SBCAPE and shear parameters are used to train the SOMs, once the cases are sorted among the nodes, additional environmental fields can be examined by constructing composites. Compositing offers a way to analyze environmental characteristics of the non-null cases that fall into each node. It is worth emphasizing that the composites are composed using the ERA-5 reanalysis from the cases themselves (as opposed to any raw SOM node output, which do not represent physical characteristics of the environment). Composite fields are shown as mean standardized anomalies, which are computed relative to the 30-year climatology (following the same procedure as the calculations for the SOM input fields). Regime characteristics are discussed qualitatively.

The day-2 CSU-MLP forecasts for tornadoes, severe wind, and severe hail are sorted into their appropriate SOM regime (i.e., each forecast is matched with the ERA-5 case that falls in that valid

forecast period). Spatiotemporal patterns of the forecast probabilities are analyzed across each node. Results are also framed in the context of the ERA-5 composites across the different regimes.

Brier skill score, or BSS, is used as the main verification method to evaluate the skill of the CSU-MLP forecasts across various regimes. BSS is computed as follows:

$$BSS = 1 - \frac{BS_{fcst}}{BS_{ref}} \tag{3.1}$$

where BS_{fcst} is the Brier score (Brier, 1950) of the CSU-MLP forecast, and BS_{ref} is the Brier score of the reference climatology. The reference climatology is a smoothed daily 30-year climatology (1990-2019) of severe weather reports from the SPC that has been used in past CSU-MLP evaluations (see Hill et al. (2023) for details). BS describes how accurate a probabilistic forecast is at predicting a binary event (typically using mean squared error), and the BSS describes how good a probabilistic forecast is relative to climatology. A positive score would indicate that the probabilistic forecast is more skillful than climatology, whereas a negative score implies the opposite.

BSS is computed for CSU-MLP forecasts as well as SPC forecasts, so that the ML forecasts can be compared both to climatology as well as operational human-generated guidance. Day-2 SPC forecasts are sorted into nodes in the same manner as the day-2 CSU-MLP forecasts. All SPC day-2 forecasts that match the dates of the non-null day-2 CSU-MLP forecasts are retained (regardless of whether they would be considered null cases or not). In order to create a fairer comparison between the CSU-MLP forecasts (which have continuous probabilities) and the SPC probabilities (which have discrete probabilities), probabilities from both products are discretized to the midpoint of the SPC probability bins (see Hill et al. (2023) for more information).

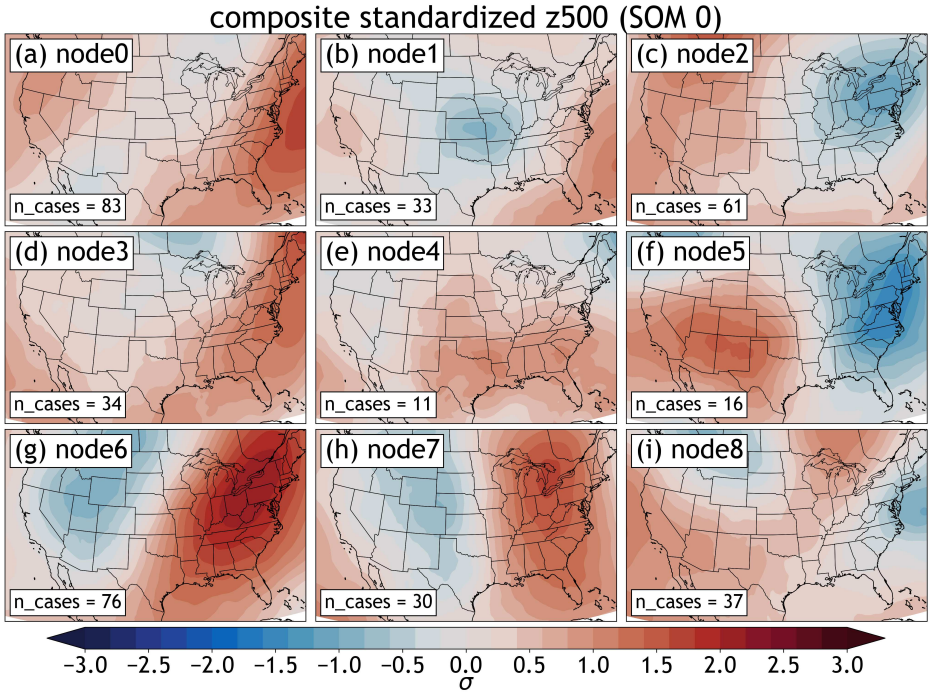

Figure 3.2: Mean standardized daily 500 hPa height anomalies (see methods for details), sorted by each node in SOM0. Node numbers and the number of non-null forecast cases in each node are annotated on each panel.

To analyze variability in skill across cases in each of the nodes, daily BSS is used to discriminate between “best" and “worst" forecasts. Best forecasts are considered to be forecasts with a daily BSS in the top 25% (75th percentile or greater) of all 381 forecasts in the dataset, whereas worst forecasts are characterized by a daily BSS in the bottom 25%. Best and worst forecasts are compared across the nodes.

3.3 Results

3.3.1 SOM-identified regime characteristics

Examining the overall synoptic pattern using 500 hPa composite ERA-5 reanalysis standardized geopotential height anomalies (z500; Figs. 3.2; 3.3) shows that there are distinct regimes across the nodes identified by both SOM0 and SOM1. In SOM0 (Fig. 3.2), two of the nodes

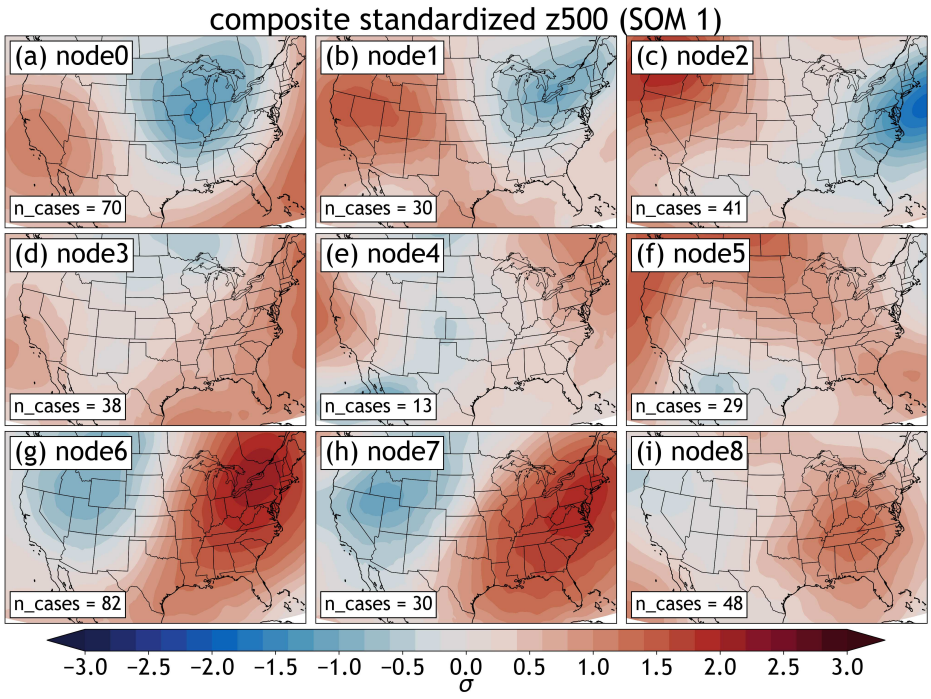

Figure 3.3: As in Fig. 3.2, but for the SOM1 node configuration.

(nodes 2 and 5) show anomalous mid-level ridging over the western CONUS coupled with anomalous mid-level troughing over the eastern CONUS. In SOM1, nodes 0, 1, and 2 match this pattern most closely (Fig. 3.3a,b,c). Meanwhile several of the nodes in both SOMs show a nearly opposite pattern, with an anomalous ridge in the eastern CONUS and an anomalous trough in the western CONUS (i.e., nodes 6 and 7 in SOM0 and SOM1). Node 3 in the SOMs resemble each other, with above-average heights over most of CONUS and subtle troughing over Canada. Node 1 in SOM0 (Fig. 3.2b) and node 4 in SOM1 (Fig. 3.3e) are perhaps less similar to each other compared to other nodes, but they share some resemblance to each other with respect to below-average heights in the center of CONUS with slightly above-average heights in other locations. The node similarities suggest that both SOMs pinpoint characteristically similar regimes.

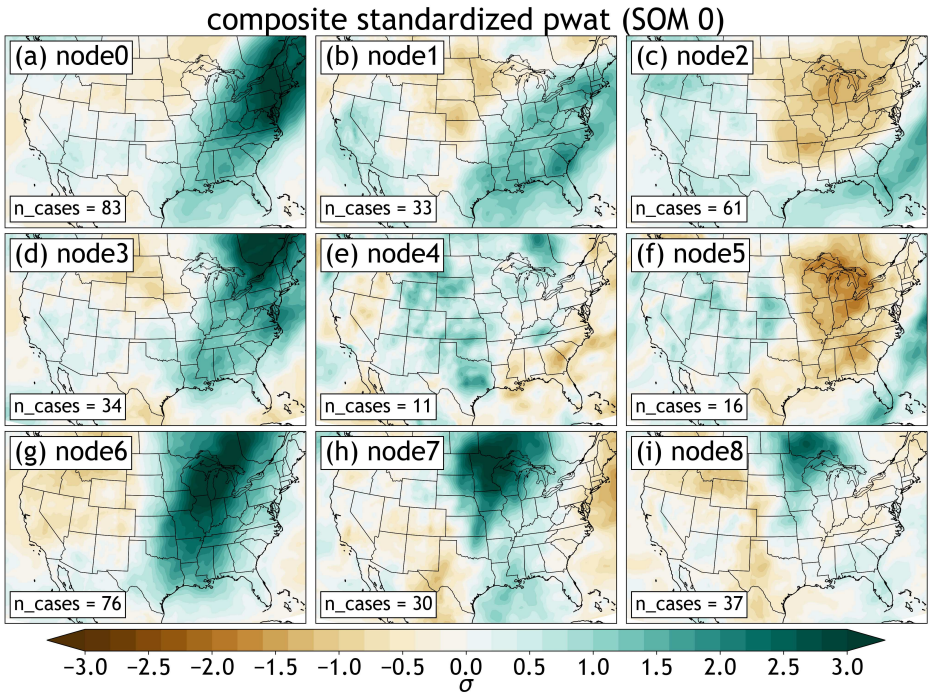

Figure 3.4: As in Fig. 3.2, but for precipitable water (PWAT).

However, despite the number of similarities in the 500 hPa height anomalies composited across the nodes in each SOM, there are some differences in the synoptic regimes between the two SOMs. For example, node 8 in SOM0 (Fig. 3.2i), which shows below-average heights over the Mid-Atlantic and the northern Intermountain West, does not share obvious similarities to any of the nodes in SOM1. Similarly, the anomalously low heights over the Pacific Northwest coupled with above-average heights in the eastern third of CONUS in node 8 of SOM1 (Fig. 3.3i) seems to be a distinct regime from the nodes in both SOMs. Node 0 in SOM0 (Fig. 3.2a) and node 5 in SOM1 (Fig. 3.3f) both suggest subtle troughing over the far southwest CONUS and northern Mexico but are otherwise dissimilar from the other nodes.

Additional details can be gleaned about the regimes in SOM0 and SOM1 by examining other composite variables, such as precipitable water anomalies (PWAT; Figs. 3.4; 3.5). For example,

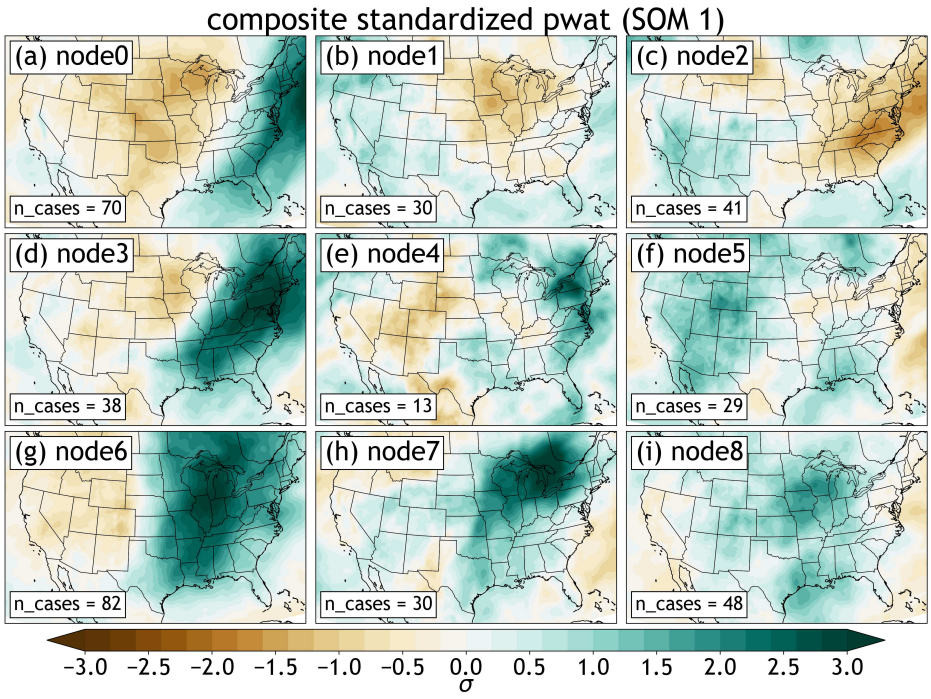

Figure 3.5: As in Fig. 3.3, but for precipitable water (PWAT).

the 500 hPa height anomaly composites show nearly-identical patterns in nodes 6 and 7 in SOM1 (Fig. 3.3g,h), yet the precipitable water composites for these nodes appear different (Fig. 3.5g,h). Node 6 shows anomalously moist conditions over the eastern half of CONUS with anomalously dry conditions over the western half. Yet, node 7 focuses the largest PWAT anomalies over the Great Lakes region, with a mixture of slightly-above and slightly-below average conditions elsewhere. Some nodes with similar height anomaly patterns do, however, share similar moisture anomalies. For example, nodes 2 and 5 in SOM0, which are characterized by a western CONUS ridge/eastern CONUS trough pattern in the 500 hPa height anomalies, show similar composite PWAT anomalies (Fig. 3.4c,f), with below-normal moisture accompanying the lower heights and above-normal moisture accompanying the higher heights. Additionally, some of the nodes that are characterized

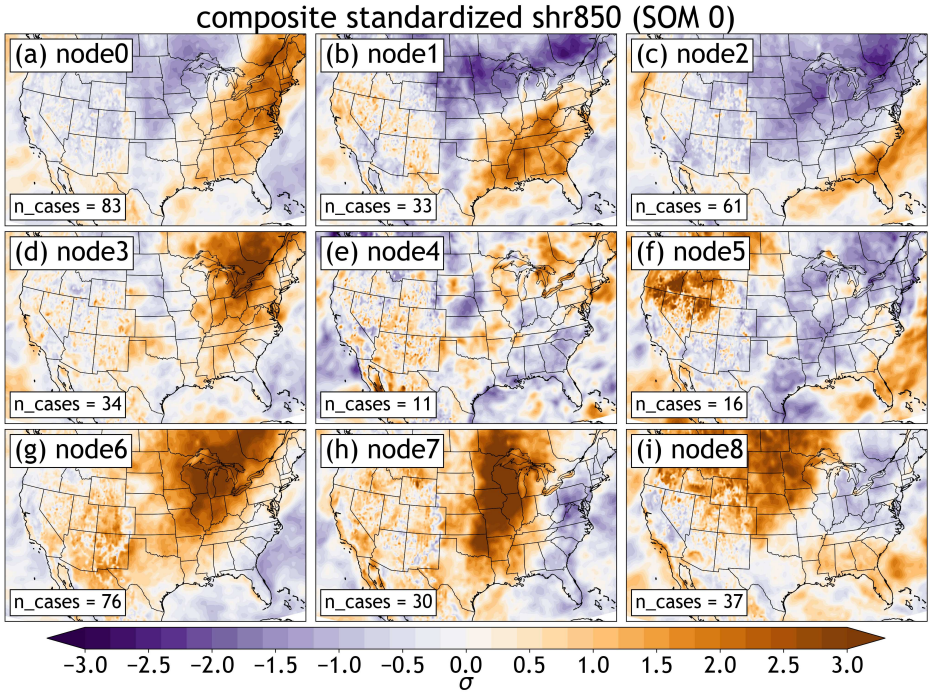

Figure 3.6: As in Fig. 3.2, but for 10-m to 850 hPa vertical wind shear.

by similar 500 hPa height anomaly patterns across both SOM0 and SOM1 (e.g., Figs. 3.2c; 3.3b) also have similar composite PWAT anomalies (Figs. 3.4c; 3.5b).

Examining vertical wind shear anomalies, both in the 10-m to 850 hPa layer (SHR850; Fig. 3.6) and in the 10-m to 500 hPa layer (SHR500; Fig. 3.7) in SOM0 provides additional insights about the regimes and may also elucidate which shear parameter is a better regime diagnostic. Equivalent plots for SOM1 can be found in Appendix A (Figs. A.16; A.15). One notable regime that stands out among the SOM0 nodes is node 6, which is characterized by anomalously high low-level shear values over the Midwest (Fig. 3.6g) along with anomalously high mid-level shear farther upstream (Fig. 3.7g), near where the 500 hPa height anomaly gradient exists (Fig. 3.2g). Node 5 shows anomalously large SHR850 and SHR500 values over the Northwest CONUS (Figs. 3.6f; 3.7f) atop an anomalous ridge over the Southwest (Fig. 3.2f). Notably, SHR500 composite anomalies

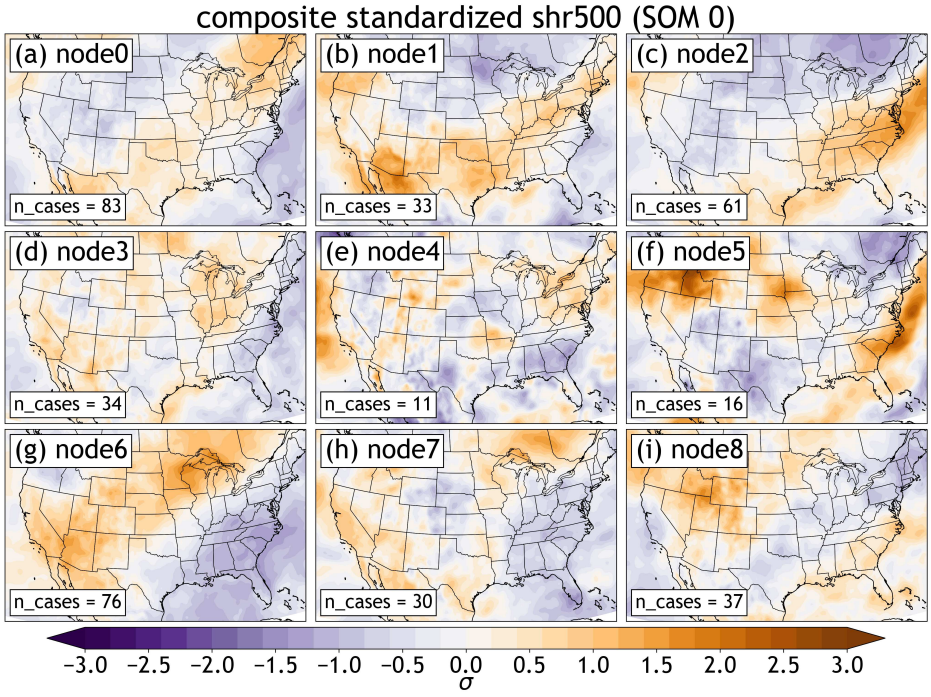

Figure 3.7: As in Fig. 3.2, but for 10-m to 500 hPa vertical wind shear.

are much smaller in magnitude overall compared to SHR850 composite anomalies. This finding is also consistent in SOM1 (Figs. A.16; A.15), even though SOM1 uses SHR500 (not SHR850) in its training. The stronger SHR850 anomalies among the different nodes suggests that it is a more useful parameter for distinguishing between different synoptic environments than SHR500, which could be attributed to differences in the widths of their distributions (i.e., the SHR850 distribution could be wider than SHR500, providing opportunity for more extreme anomalies).

Another way to make sense of the SOM nodes is to study which seasons they are associated with. In Fig. 3.8, most nodes are featured across most of the seasons, though there are some notable features. For example, DJF is dominated by node 6 in both SOM0 and SOM1, with at least a quarter of wintertime cases being classified under that regime. Node 6 also describes a large proportion of fall (SON) events. Recall node 6 is very similar in SOM0 (Fig. 3.2g) and SOM1 (Fig. 3.3g),

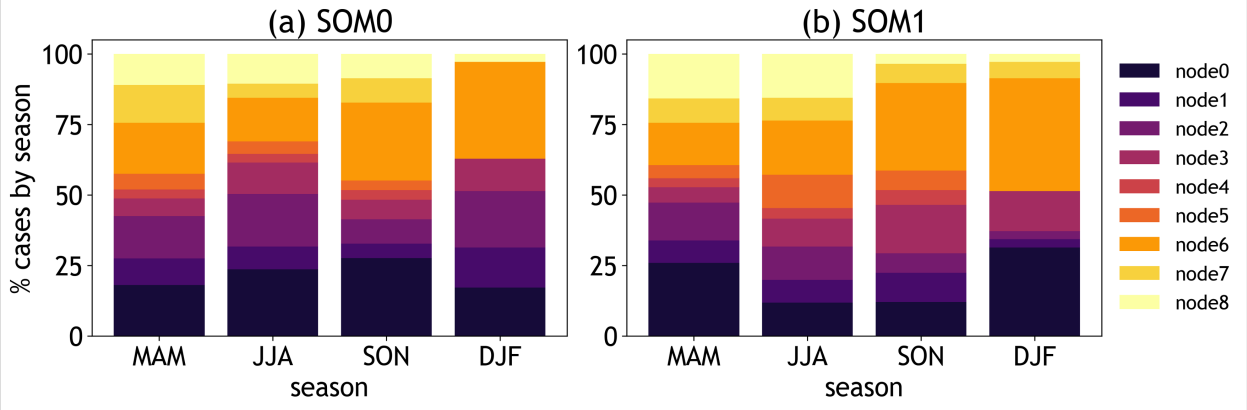

Figure 3.8: Regime composition of each season for (a) SOM0 and (b) SOM1.

featuring prominent regions of below-normal and above-normal anomalies in the 500 hPa heights. Node0 is also characterized by a longwave trough pattern over CONUS (specifically the Midwest), and cases in that node also tend to comprise a large percentage of wintertime severe weather events (Fig. 3.8). It is well-known that cool season severe weather events tend to rely on strong synoptic forcing (e.g., Sherburn and Parker, 2014; Sherburn et al., 2016), so these observations agree with physical reasoning. However, not all regimes fit this rationale. For example, node 7 in SOM1 exhibits a 500 hPa height anomaly pattern similar to node 6, yet node 7 cases account for a similar fraction of severe weather events across all seasons (Fig. 3.8b). That is, despite sharing synoptic characteristics with node 6, node 7 SOM1 cases do not seem as strongly tied to the fall and winter season as node 6 cases do (though it is worth mentioning that there are nearly three times as many forecasts in node 6 than node 7). Node 7 in SOM0 shares similar z500 composite characteristic to node 7 in SOM1 and despite them having the same number of cases, SOM0 node 7 cases account for a larger fraction of MAM events than the SOM1 node 7 does.

The spring, summer, and fall months feature a more diverse set of regimes in both SOMs compared to winter months. In both SOMs, MAM, JJA, and SON all feature cases from all nine

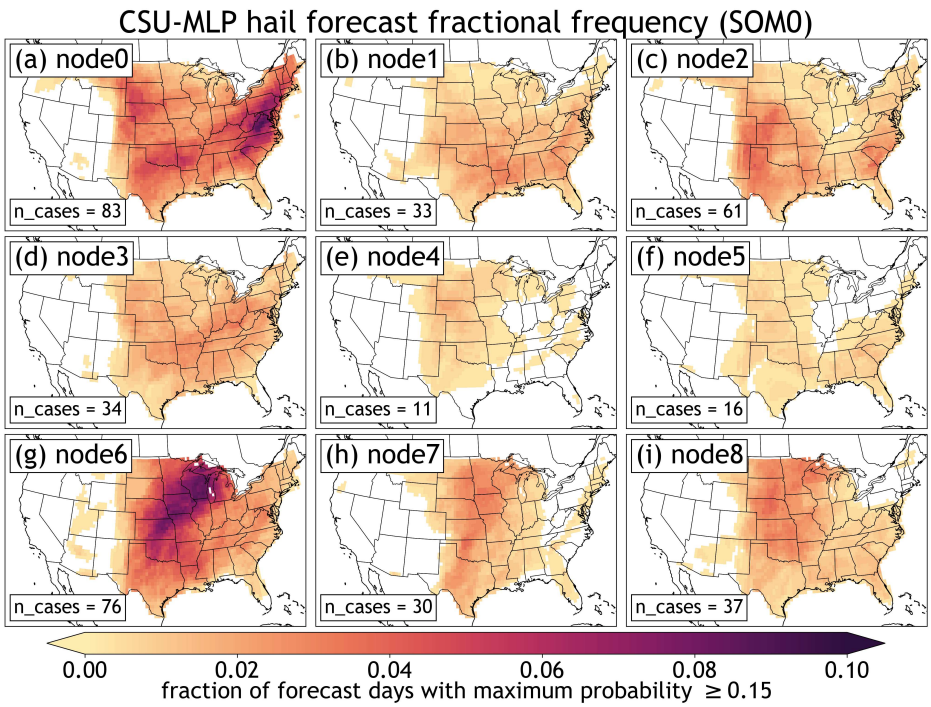

Figure 3.9: Fraction of non-null day-2 CSU-MLP hail forecasts out of total forecast days (381) at each node diagnosed at SOM0. A non-null forecast day is considered a forecast with a *maximum* hail probability of at least 15%; thus note that lower probabilities in the non-null cases are still considered here.

nodes (Fig. 3.8), whereas this is not the case for DJF. Nodes 4 and 5 are notably absent from the sets of wintertime cases. This result can likely be attributed to the methods used and seasonal variability in severe storm environments. Recall that while all two years of reanalysis data are used for the SOM development, only cases associated with CSU-MLP forecasts that have at least a “slight" risk are used for the node composites and subsequent analysis. Severe storm environments tend to be much more common in spring through fall than they are in the winter (and thus frequency of CSU-MLP severe probabilities also tend to fluctuate seasonally), so a disproportionate number of winter environments are removed from the dataset. Thus, events in the spring through fall appear to be described by the SOM regimes with more granularity than the winter cases.

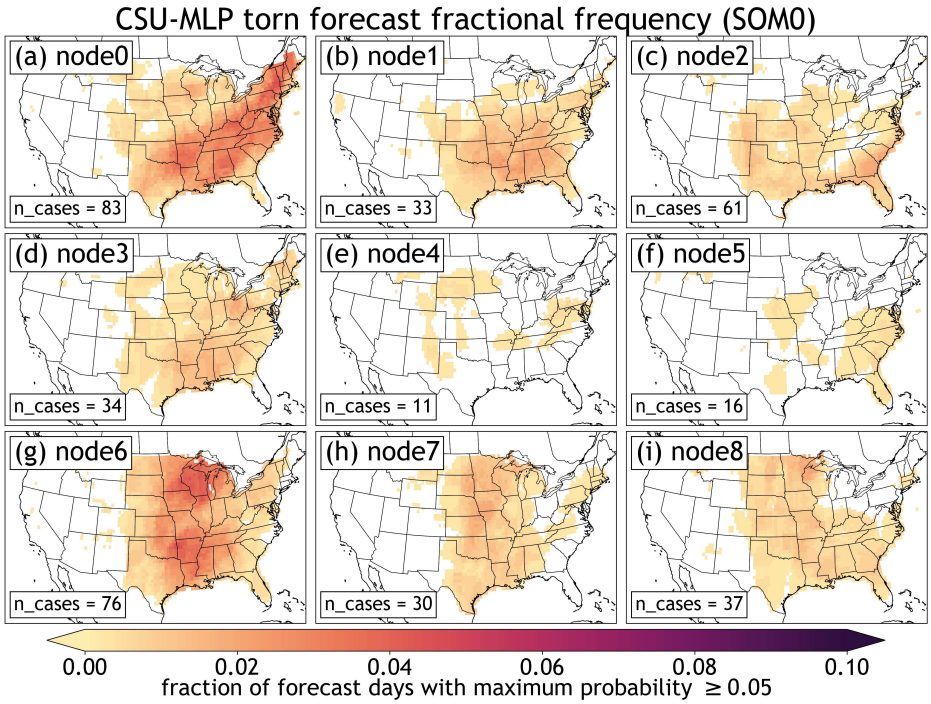

Figure 3.10: As in Fig. 3.9, but for day-2 CSU-MLP tornado forecasts. Note the maximum daily probability must only exceed 5% to be considered a null case here.

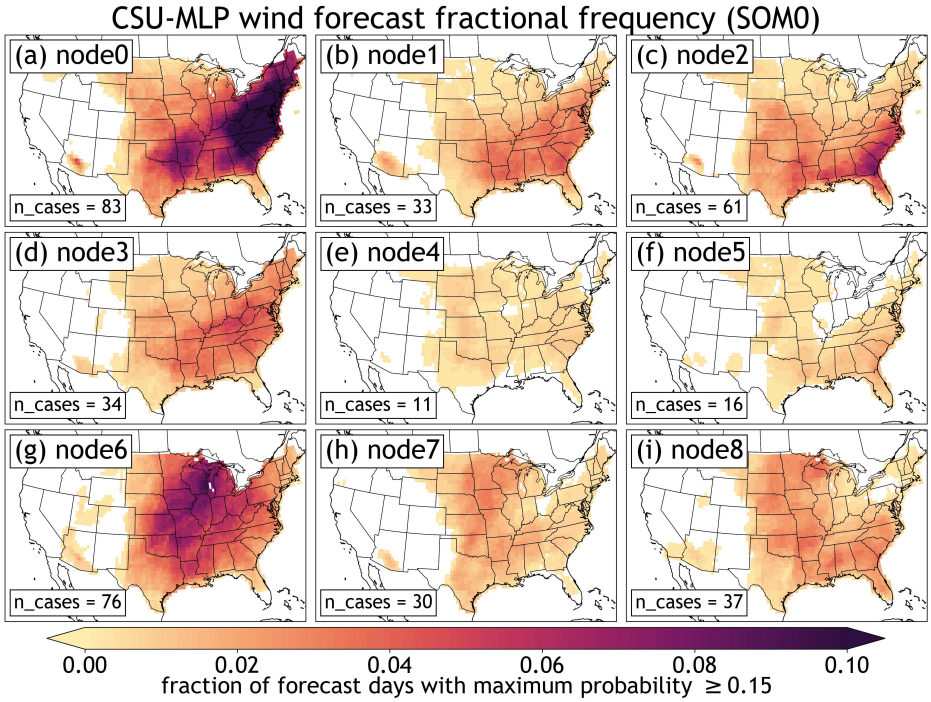

Figure 3.11: As in Fig. 3.9, but for day-2 CSU-MLP wind forecasts.

3.3.2 CSU-MLP forecast characteristics across regimes

With these SOM regimes introduced, characteristics of the CSU-MLP forecasts in the nodes can be discussed. Figs. 3.9- 3.11 show the fraction of day-2 CSU-MLP forecasts (for hail, tornadoes, and wind respectively) out of the total number of non-null forecast days (i.e., "slight" risk days, or forecasts with a daily maximum probability of at least 5% or greater for tornadoes or 15% or greater for hail) that fall into a given SOM0-defined regime[11]. Equivalent plots for the SOM1 regimes are not discussed here for the sake of brevity, but they can be found in Appendix A.

There is apparent spatial variability in the CSU-MLP hail, tornado, and wind forecasts. One notable feature is that a majority of CSU-MLP hail probabilities along the East Coast tend to be associated with the node0 regime (Fig. 3.9a). This observation holds true for the tornado (Fig. 3.10a) and especially the wind forecasts (Fig. 3.11) as well. This regime is characterized by relatively weak synoptic forcing compared to many of the other regimes according to the 500 hPa height anomalies (Fig. 3.2a), with composite PWAT anomalies 2 to 3 standard deviations above the mean, especially over the Northeast (Fig. 3.4a). Low-level (10m–850 hPa) shear values are also anomalously high across this region (Fig. 3.6). A large fraction of northern High Plains hail probabilities also are grouped into node0, though that is not the case for the tornado or wind probabilities. In this region, the node0 composites show near-normal to below-normal shear and instability, which contrasts the pattern along the East Coast (Figs. A.6a; 3.7a).

[11] Note that the contours for each forecast day include probabilities as small as 2% for tornadoes and as small as 5% for hail and wind (i.e., a "marginal" risk). So, shaded areas reflect the frequency of CSU-MLP forecasts with a *maximum* probability at least in the "slight" risk (not the overall frequency of 15% or greater probabilities).

Node 6 also describes a relatively large fraction of CSU-MLP tornado, hail, and wind forecasts (Figs. 3.9g; 3.10g; 3.11g). Of particular note is the large fraction of tornado, wind, and hail forecasts in the Midwest and tornado probabilities in the Deep South that fall into this node. Node 6 is characterized by a distinct swath of positive 500 hPa height anomalies over the western CONUS coupled with negative height anomalies to the east (Fig. 3.2g). These height anomalies are accompanied by large positive PWAT anomalies stretching from the Midwest to Gulf Coast, with particularly high values near the Great Lakes (greater than 3 standard deviations above the mean; Fig. 3.4). There are also anomalously high amounts of low-level (10m-850 hPa) and deep-layer (10m-500 hPa shear) present (Figs. 3.7g; 3.6g) in these locations.

Separate from nodes 0 and node 6, there are a few other characteristics about the forecast frequencies across the SOM0 nodes that are worth mentioning. For example, CSU-MLP wind probabilities are common across the southeastern CONUS, and they occur with notable frequency across a number of nodes in addition to nodes 0 and 6 (i.e., nodes 1, 2, 3 and 8; Fig. 3.11). Nodes 4, 7 and 8 seem to be dominated by High Plains hail probabilities as opposed to other regions (Fig. 3.9). Lastly, probabilities west of the Rockies are rare, but they seem to be well-distributed across the nodes (rather than confined to specific regimes). While the same could be said for a number of regions east of the Rockies, it is a little surprising given the infrequency of severe storm environments in western CONUS cases that they did not cluster on specific CAPE/shear regimes. Perhaps these cases were largely ignored by the SOM algorithm due to relatively weak signals in the SBCAPE and shear fields compared to other cases.

Table 3.3: BSS threshold for 75th percentile and 25th percentile cases for the 381 CSU-MLP and SPC day-2 forecasts of hail, tornadoes, and wind.

Forecast product	75th percentile BSS threshold (“best" cases)	25th percentile BSS threshold (“worst" cases)
CSU-MLP hail	0.051	-0.127
CSU-MLP tornado	0.005	-0.061
CSU-MLP wind	0.110	-0.017
SPC hail	0.057	-0.027
SPC tornado	0.009	-0.016
SPC wind	0.087	-0.003

3.3.3 CSU-MLP forecast skill across regimes

Table 3.3 shows the minimum threshold for the 25% most skillful (or “best") and maximum threshold for the bottom 25% least skillful (or “worst") day-2 CSU-MLP and SPC forecasts, based on daily BSS. Note that the “best" forecast thresholds are greatest for the wind forecasts (followed by the hail and then the tornado forecasts), meaning that those forecasts have the most stringent criteria for being considered a “best" forecast (relative to the hail and tornado forecasts). It also demonstrates that the wind forecasts tend to be the most skillful among the three hazards, which generally agrees with the findings in Hill et al. (2020) though is somewhat complicated due to practices surrounded estimated wind reports (e.g., Edwards et al., 2018).

Using the thresholds in Table 3.3, the “best" and “worst" forecasts in each of the SOM nodes can be identified. Figs. 3.12 and 3.13 show the relative percentage of CSU-MLP and SPC forecasts in each node that are considered a “best" or “worst" forecast. Bars that are above the 25% line indicate that forecasts in those nodes tend to be disproportionately skillful (panels (a) and (b)) or unskillful (panels (c) and (d)) compared to the other nodes. Note that this metric does not

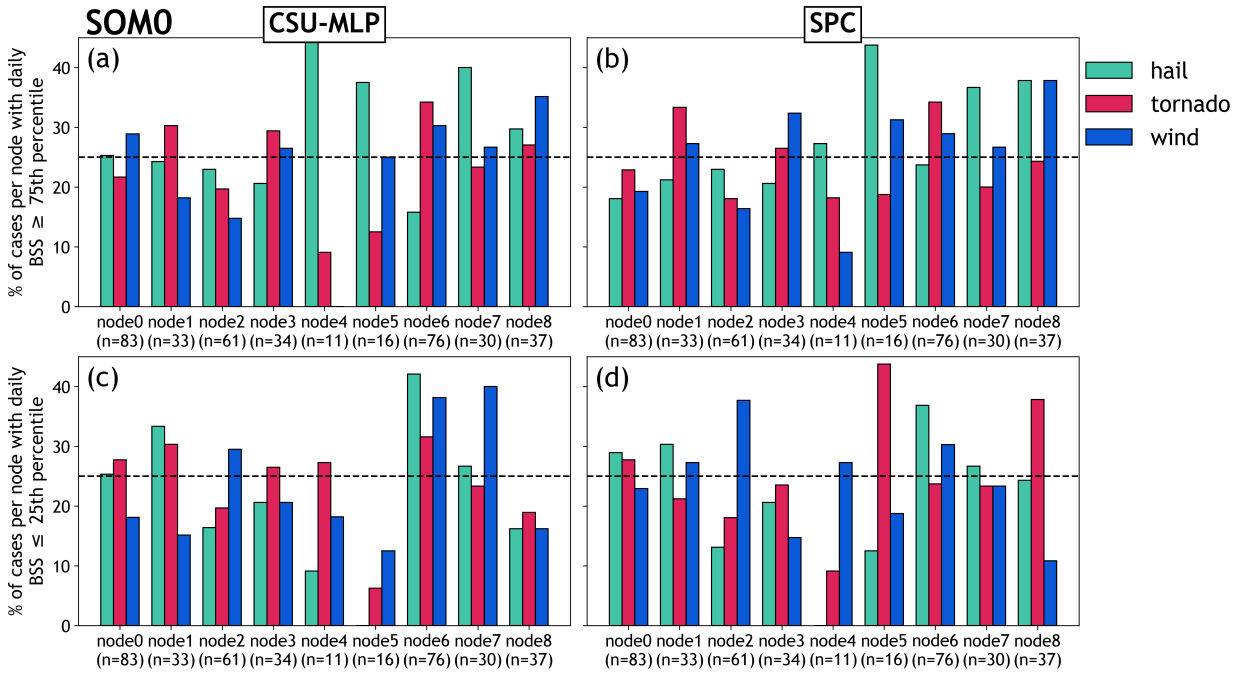

Figure 3.12: In the top panels, relative percentage of cases in each node with a daily BSS in the top 25% of all (a) day-2 CSU-MLP forecasts and (b) day-2 SPC outlooks, separated by SOM0-identified node. In the bottom panels, relative percentage of cases in each node with a daily BSS in the bottom 25% of all (c) day-2 CSU-MLP forecasts and (d) day-2 SPC outlooks, separated by node. BSS for best and worst cases are shown for day-2 hail (teal), tornado (pink), and wind (blue) forecasts. The black dashed line marks 25%.

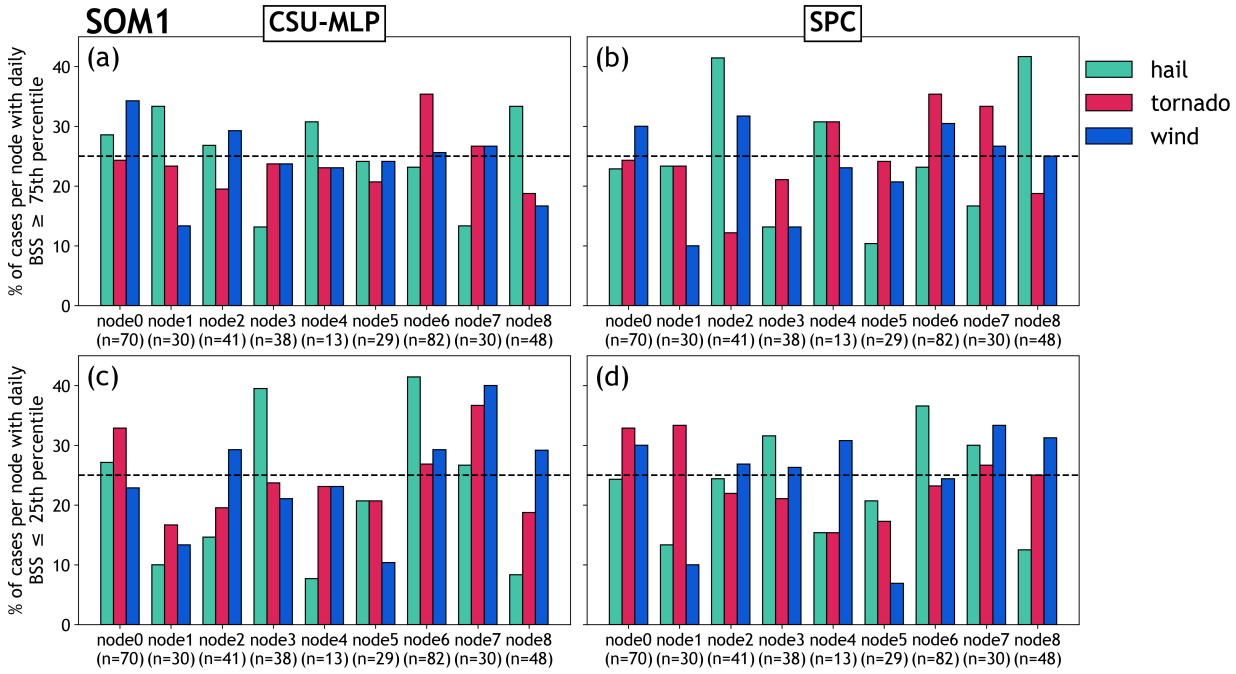

Figure 3.13: As in Fig. 3.12, but for the SOM1 node configuration.

necessarily describe which nodes have the largest (or smallest) BSS, but rather it describes which nodes are most often associated with the most and least skillful forecasts.

There is variability across the hazards in terms of the frequency that forecasts in the nodes tend to be skillful versus unskillful. In the CSU-MLP forecasts classified by SOM0, the most skillful hail forecasts are most often associated with nodes 4, 5, and 7 (Fig. 3.12a), whereas the least skillful forecasts tend to most frequently be associated with nodes 1 and 6 (Fig. 3.12c). SPC forecast skill largely mirrors this result, with the exception of node 8 also showing a disproportionate number of skillful forecasts compared to the other nodes (Fig. 3.12b). Nodes 6 and 7 exhibit similar z500 composite anomalies (i.e., troughing over the western CONUS; Fig. 3.2), but CSU-MLP hail probabilities in node 7 seem to be more confined to areas over the High Plains and Midwest, whereas the probabilities in node 6 are more widespread (Fig. 3.9). It is worth noting that both the CSU-MLP and SPC hail forecasts classified under the node 6 regime in SOM1 (which is synoptically similar to node 6 in SOM0), also have a tendency to have low skill (Fig. 3.13).

Like the CSU-MLP hail forecasts, tornado forecasts in the SOM0 node 1 and 6 regimes have a greater tendency to perform poorly compared to forecasts classified under other nodes (Fig. 3.12c). However, the tornado forecasts also have a greater propensity to perform well in nodes 1 and 6– both in the CSU-MLP (Fig. 3.12a) and SPC (Fig. 3.12b) forecasts, which was not the case for the hail forecasts. In other words, it seems that the tornado forecasts have both strong successes (forecasts in the “best" category) and big failures (forecasts in the “worst" category) in these types of regimes relatively often. Synoptically, these regimes are quite different from each other (Figs. 3.2b,g; 3.4b,g), and their forecast frequencies also differ spatially (Fig. 3.10b,g), so additional analysis would be needed to understand why these particular regimes have greater likelihoods of

having both good and poor forecast skill. It is worth noting that the SPC forecasts in nodes 1 and 6 *do not* have an increased tendency to be low-skill compared to forecasts in other nodes (Fig. 3.12d). The worst-performing SPC forecasts seem to be disproportionately associated with nodes 5 and 8.

In general, the best CSU-MLP wind forecasts are relatively well-distributed across the nodes. This pattern can be seen in both SOM0 (Fig. 3.12a) and SOM1 (Fig. 3.13a) and suggests that the skillfulness of wind forecasts may be less reliant on regime type (compared to the hail and tornado forecasts). The worst performing CSU-MLP wind forecasts tend to be associated with the regimes described by nodes 2, 6 and 7 in both SOM0 and SOM1 (Figs. 3.12c; 3.13c). SPC wind forecasts in node 2 of SOM0 also have a notably greater tendency to do poorly compared to forecasts associated with other nodes (Fig. 3.12d). Interestingly, nodes 2, 6 and 7 are all characterized by a prominent regions of positive and negative anomalies in the 500hPa height composites (Figs. 3.2c,g,h; 3.3c,g,h), suggesting that well-defined wave patterns in mid-level heights (that can at times provide strong synoptic forcing) may not necessarily benefit wind forecast skill.

Across the different SOM nodes, there is clear variability in the frequency that CSU-MLP probabilistic forecasts tend to exist among the best and worst predictions by the ML system. But what does this skill look like across each of the nodes? Fig. 3.14 illustrates the mean BSS of the best- and worst-performing CSU-MLP and SPC forecasts in each of the SOM0 nodes (analogous results for SOM1 are in Appendix A; Fig. A.21). In SOM0, the nodes with the most-skillful CSU-MLP forecasts (i.e., forecasts that have a BSS above the 75th percentile) for hail, tornadoes, and wind are nodes 3, 6 and 5 respectively (Fig. 3.14a). The nodes with the lowest mean BSS among the poorest-performing forecasts, are nodes 3, 7, and 1 for the hail, tornado and wind forecasts respectively (Fig. 3.14b). Indeed, the mean best and worst CSU-MLP hail forecasts are classified

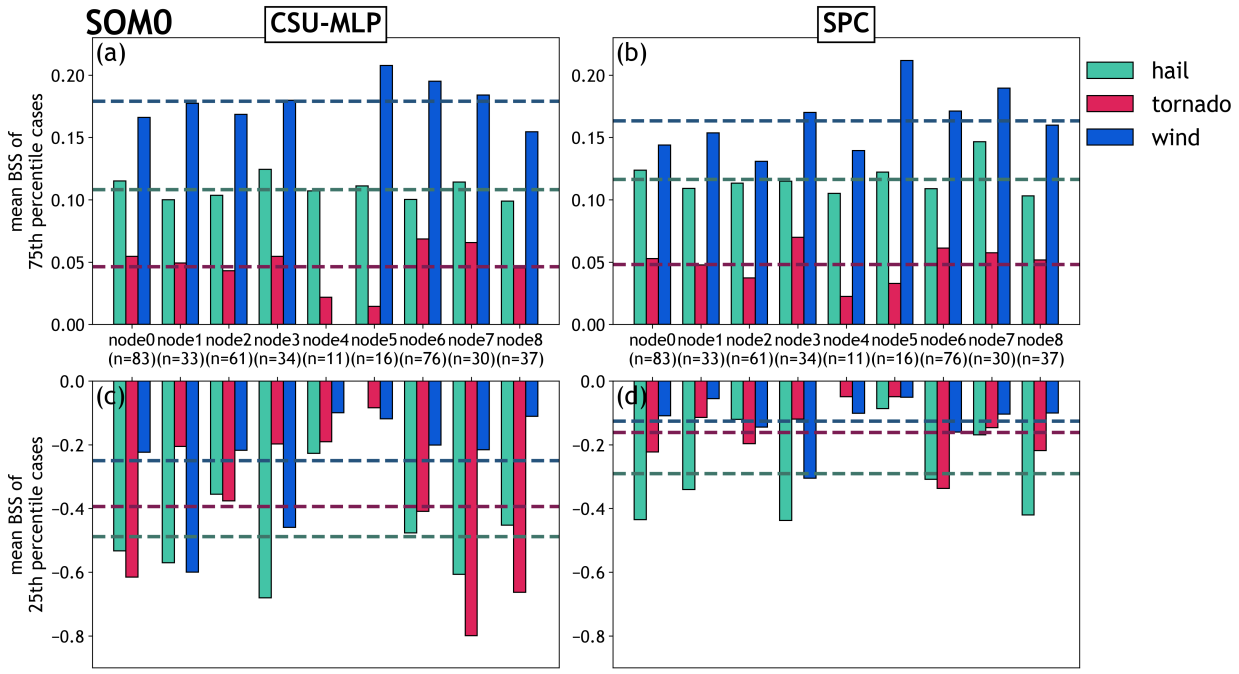

Figure 3.14: In the top panels, mean BSS for top 25% most-skilled (a) CSU-MLP day-2 forecast cases and (b) SPC day-2 outlooks, separated by SOM0-identified node. In the bottom panels, mean BSS for the bottom 25% least-skilled (c) CSU-MLP day-2 forecast cases and (d) SPC day-2 outlooks across the nodes. BSS for best and worst cases are shown for day-2 hail (teal), tornado (pink), and wind (blue) forecasts. Teal, pink, and blue dashed lines represent the mean BSS among all best (or worst) hail, tornado, and wind forecasts across all the nodes. The number listed by each of the SOM node labels represents the total number of non-null forecasts that are in that node.

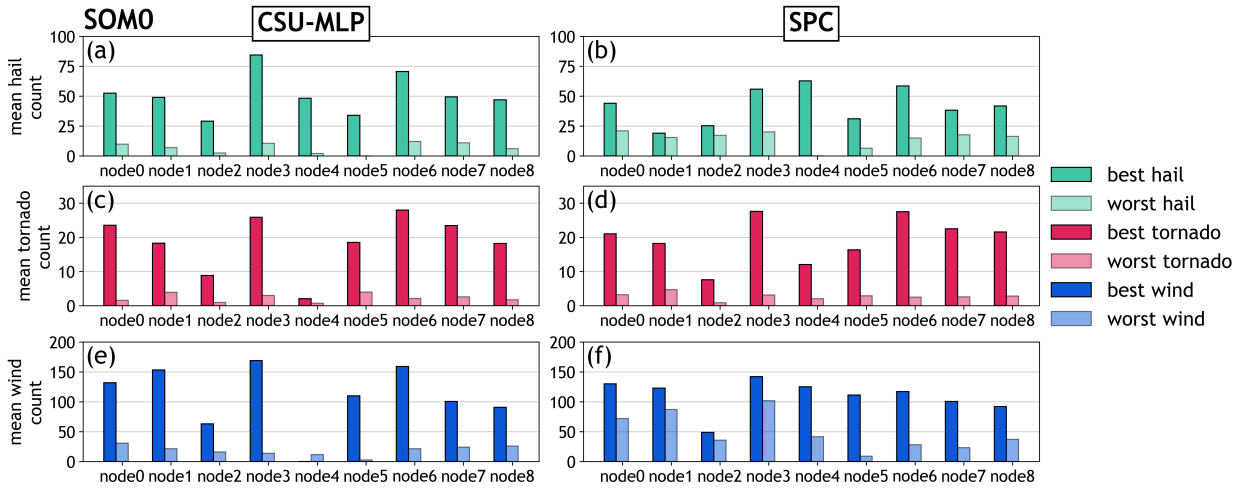

Figure 3.15: Mean number of CSU-MLP grid points with at least one (a), (b) hail, (c), (d) tornado, or (e), (f) wind report among best and worst CSU-MLP and SPC forecasts, separated by SOM0 nodes. Mean counts are shown for CSU-MLP forecasts in the left column and SPC forecasts in the right column.

under the same SOM node (note that SPC forecast skill is also poor in this node; Fig. 3.14d). However, node 5–the most skillful among the “best" wind forecasts–has one of the lowest mean BSS among the “worst" forecasts in that node, suggesting that wind overall forecast skill is relatively high compared to other nodes and that node 5-type wind events are quite predictable. SPC wind forecast skill is also very high in node 5, evidenced by the high mean BSS among both the best- and worst-performing forecasts in that regime (Fig. 3.14b,d). Meanwhile, the regimes with the highest mean BSS among the best-performing SPC forecasts are nodes 7 and 3 for the hail and tornado forecasts (respectively).

One approach towards understanding variability in forecast performance across the different SOM nodes is to study reports across each of the nodes. In Fig. 3.15, it is immediately apparent that the best-performing forecasts (both by CSU-MLP and SPC) tend to be associated with a much larger number of storm reports compared to the worst-performing forecasts. This observation is true across all hazard types and nodes, as well as for SOM1 (Fig. A.22). This result

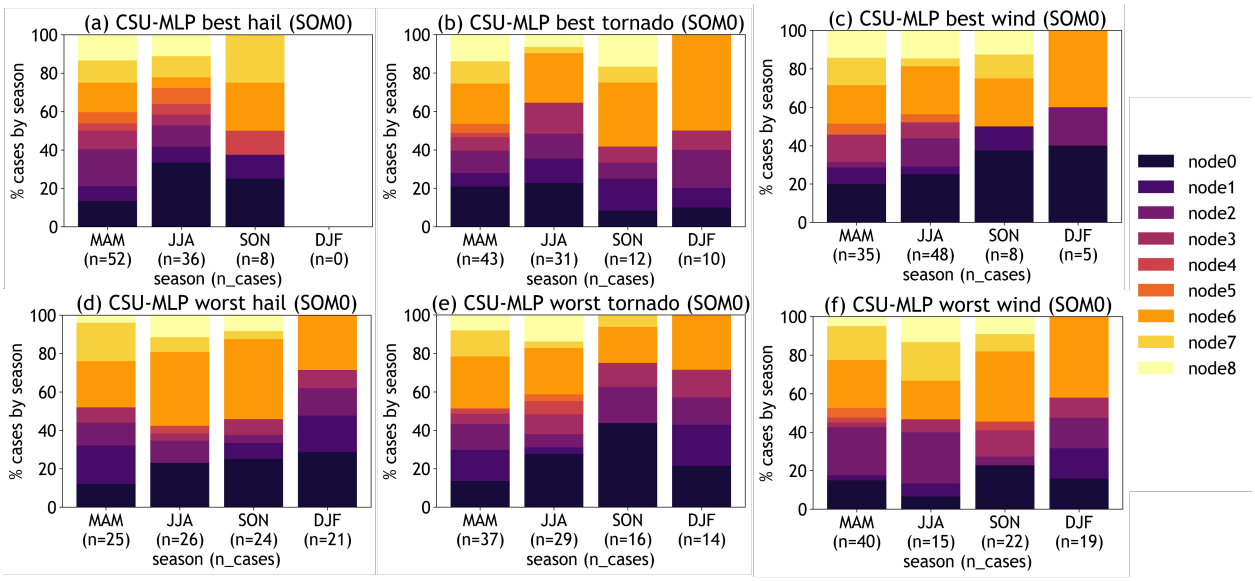

Figure 3.16: For each season, SOM0 node composition of best and worst (a),(d) hail, (b),(e) tornado, and (c),(f) wind forecasts that occurred during that season. The total number of best or worst forecast cases falling into each season for a given hazard are annotated along the x-axes.

suggests that CSU-MLP and SPC forecasts performance is high during severe weather outbreaks (when there are many reports) and lower when severe weather events are more isolated. The relatively few reports associated with the poor-performing forecasts can also be viewed in map form across the various nodes (Figs. A.23 - A.25). Among the individual nodes, the largest numbers of storm reports among the best-performing CSU-MLP forecasts seem to occur in nodes 3 and 6 (Fig. 3.15a,c,e), which happen to be the best-performing nodes for the hail and tornado forecasts respectively (Fig. 3.14).

To further understand characteristics of the most- and least-skillful CSU-MLP forecasts, the best and worst forecasts can also be examined seasonally across the SOM0 nodes. Fig. 3.16 illustrates the seasonal SOM0 node composition of best- (panels a-c) and worst- (panels d-f) performing CSU-MLP hail, tornado, and wind forecasts. One notable feature is the difference in the fraction of node 6 forecasts comprising the hail forecasts between the best- and worst-performing

cases. Among the best-performing hail forecasts (Fig. 3.16a), the forecasts associated with node 6 occupy a relatively small fraction of the MAM and JJA forecasts. However, among the worst-performing forecasts, node 6 accounts for nearly 40% of the overall worst-performing forecasts in JJA (Fig. 3.16d). This result suggests that JJA hail forecasts made under node 6 conditions are likely to be less skillful. The frequency of the node 0 forecasts in the best- and worst-performing tornado and wind forecasts also varies substantially. Among the best-performing tornado forecasts, node 0 represents a relatively small fraction of the best-performing forecasts across the seasons, especially in SON and DJF (Fig. 3.16b). Yet, node 0 tornado events account for very large fractions of the worst-performing tornado events in JJA and SON (Fig. 3.16e). However, nearly the opposite is true in the seasonal breakdown of the best- and worst-performing wind forecasts, where node 0 forecasts dominate the best-performing wind forecasts (especially in SON and DJF), and they comprise relatively small fractions of the worst-performing forecasts (Fig. 3.16c,f). To summarize, node regimes are more common at certain times of year than others, and there is seasonal variability in how often they tend to yield forecasts with high skill versus low skill.

3.4 Discussion: node characteristics of best- and worst-performing CSU-MLP forecasts

It is helpful to contextualize the CSU-MLP forecast skill across the SOM nodes in terms of the regime and forecast characteristics. As such, this section takes a closer look at the SOM0 regime characteristics associated with the best- and worst-performing hail, tornado, and wind forecasts. The specific nodes that are discussed here have the highest (or lowest) mean BSS among the 75th percentile and 25th percentile CSU-MLP forecasts that are in each SOM0 node. According to Fig. 3.14a,b, nodes 3, 6, and 5 have the highest mean BSS among the best-performing hail, tornado,

and wind forecasts (respectively), and nodes 3, 7 and 1 have the smallest mean BSS among the worst performing hail, tornado, and wind forecasts (respectively).

Both the best- and worst-performing CSU-MLP hail forecasts tend to occur in node 3 of SOM0 (Fig. 3.17). Hail forecasts in node 3 cover almost the entire eastern two-thirds of CONUS, and thus they do not account for a large fraction of hail forecasts in any particular area (Fig. 3.9d). However, reports associated with the best-performing node 3 hail forecasts tend to occur over the Plains, parts of the Deep South (including Louisiana, Mississippi, and Alabama), as well as in parts of the Mid-Atlantic (Fig. A.23d). Reports associated with worst-performing are sporadic and occur intermittently over the High Plains and Deep South. Node 3 hail forecasts constitute a similar fraction of forecasts across all seasons, with their biggest prevalence seeming to occur during the summer and winter months (Fig. 3.8a). The best-performing node 3 hail cases occur exclusively in MAM and JJA, while the worst-performing node 3 hail forecasts span all seasons (Fig. 3.16a,d). Node 3 is characterized by very little synoptic forcing based on the 500 hPa and 850 hPa height anomalies (Fig. 3.17b,c). There tends to be subtle troughing over Canada near the Great Lakes region in node 3 cases, evidenced by below-normal mean sea level pressure and enhanced low-level shear (Fig. 3.17d,f). Lastly, there are indications of a west-east moisture and instability gradient between the Plains and areas further east (Fig. 3.17a,i) and a cold frontal trough (Fig. 3.17d), which could perhaps act as a catalyst boundary for severe storm development.

The best-performing CSU-MLP tornado forecasts with the highest average BSS occur in SOM0's node 6 (Fig. 3.18). Node 6 tornado forecasts also span a large area of CONUS, but they comprise the largest fraction of forecasts over the over the Great Lakes region and the Deep South relative to other locations (Fig. 3.10g). The tornado reports associated with the best-performing node 6

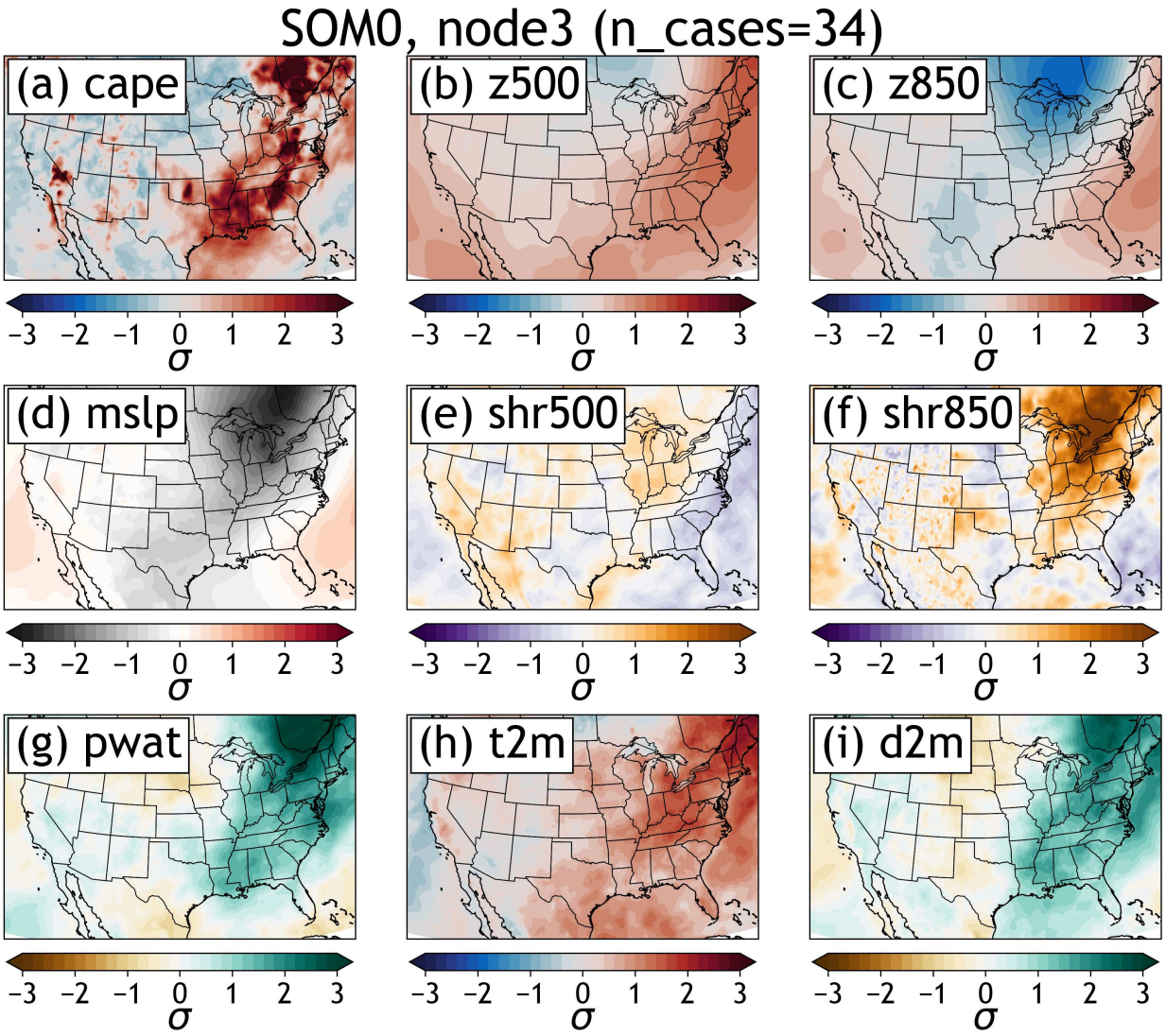

Figure 3.17: ERA-5 reanalysis composite anomalies for a variety of fields for the node 3 regime in SOM0. The CSU-MLP day-2 hail forecasts with both the best and worst skill tend to occur in node 3.

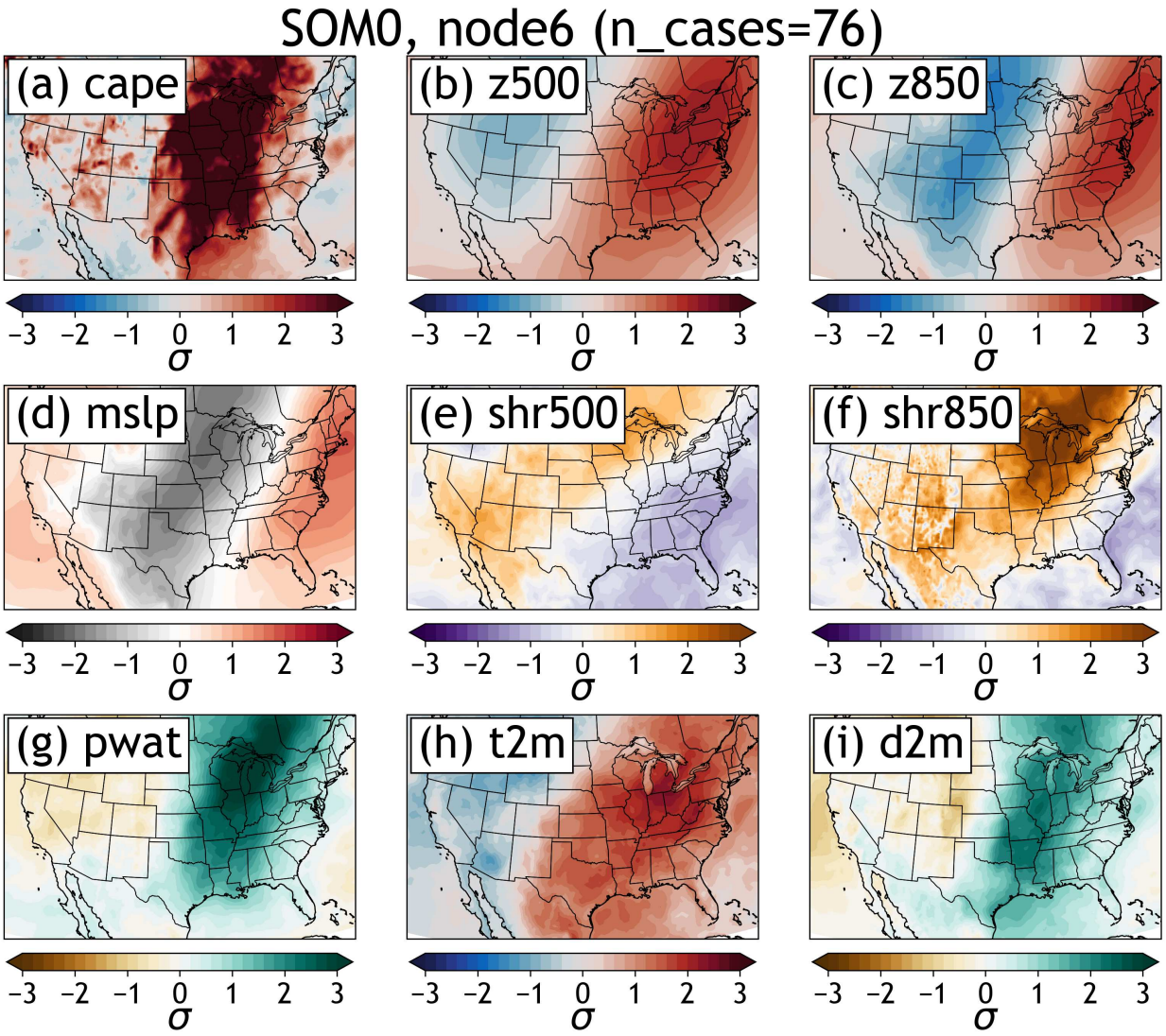

Figure 3.18: As in Fig. 3.17, but for node 6. The best-performing CSU-MLP day-2 tornado forecasts tend to be associated with node 6 regimes.

forecasts tend to be concentrated over these regions as well (Fig. A.24g), which makes sense given the enhanced positive skill of these forecasts. In general, node 6 forecasts comprise an appreciable fraction of forecasts in each season (Fig. 3.8a), and the most-skillful tornado forecasts in node 6 also span all seasons (Fig. 3.16b). However, the most-skillful node 6 CSU-MLP tornado forecasts comprise the largest fraction of the best-performing tornado forecasts during SON and DJF, accounting for 40% and 50% of cases respectively. Contrary to node 3, node 6 suggests prominent synoptic-scale forcing, with anomalous low 500 hPa heights over the Intermountain West and a 850 hPa negative height anomaly displaced slightly further east (Fig. 3.18b,c). There is anomalous shear present on the eastern sides of the trough anomalies (Fig. 3.18e,f) that is also accompanied by anomalously warm, moist, unstable air (Fig. 3.18a,g,h). This overall pattern is indicative of an ejecting longwave trough with moist, unstable air downstream and is a classic pattern associated with severe weather outbreaks.

The worst-performing CSU-MLP tornado forecasts tend to be associated with synoptic conditions described by node 7 (Fig. 3.19). In general, the tornado forecasts characterized by node 7 are mostly limited to the central CONUS (Fig. 3.10h). Among the tornado reports associated with the best- and worst-performing forecasts, most reports occur in tandem with the *best-performing* forecasts in that node; there are very few reports that occur with the worst-performing forecasts (Fig. A.24h). This poor performance accompanied by few reports suggests that the worst-performing tornado forecasts in node 7 occur either because the model overforecasts tornado probabilities under these conditions (and the magnitude and/or spatial extent of the probabilities is not backed up by reports), or the probabilities are displaced and the reports are fully missed by the forecasts. As for seasonality, node 7 forecasts only occur in spring, summer, and fall (Fig. 3.8a).

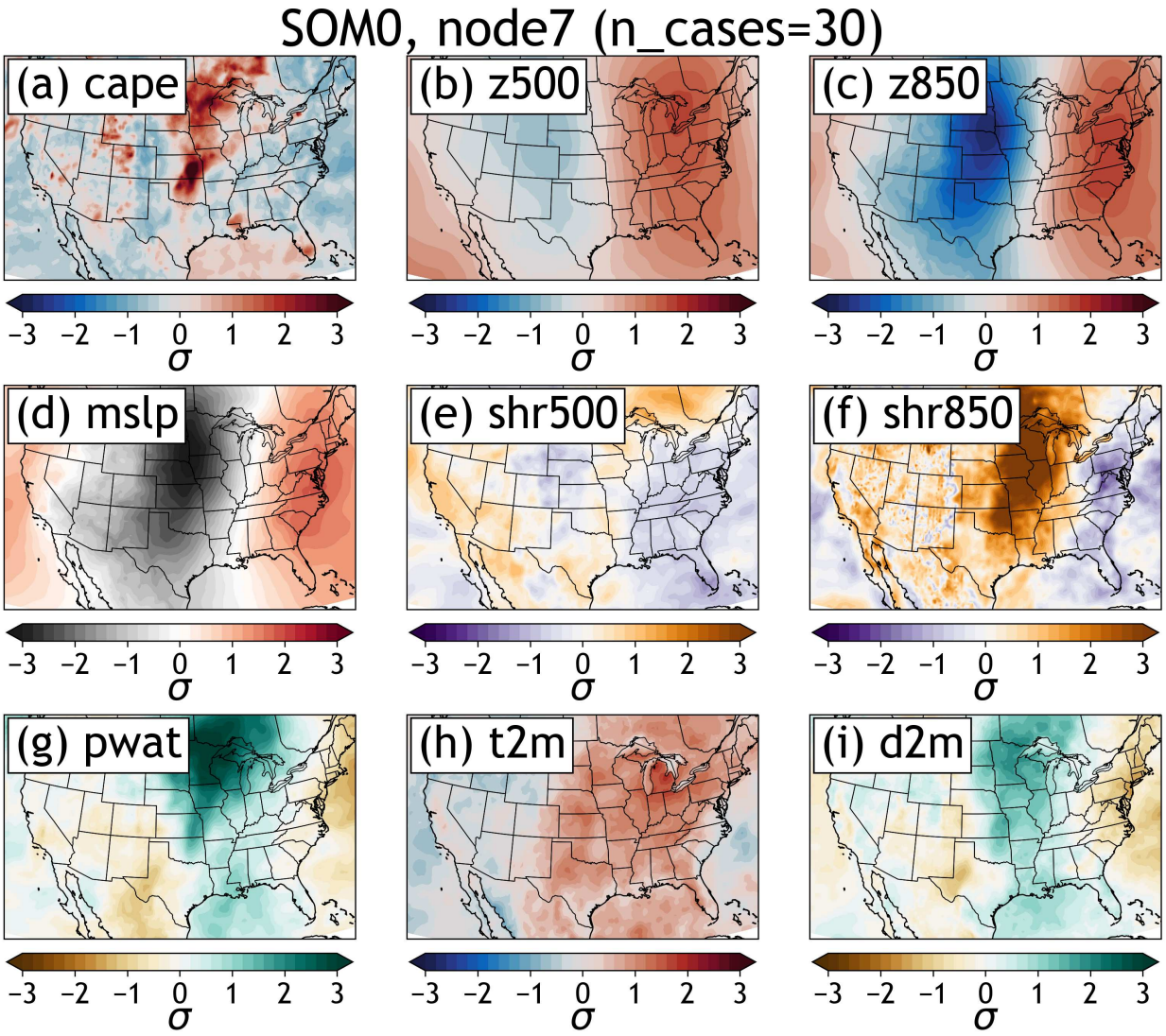

Figure 3.19: As in Fig. 3.17, but for node 7. The worst-performing CSU-MLP day-2 tornado forecasts tend to be associated with node 7 regimes.

The worst-performing node 7 CSU-MLP tornado forecasts comprise the largest portion of overall worst-performing tornado forecasts during MAM (Fig. 3.16e), though best-performing node 7 CSU-MLP tornado forecasts comprise a similar proportion of the best-performing MAM tornado forecasts. Node 7 composite anomalies show below-normal z500 over the western half of CONUS and the opposite over the eastern half of CONUS (Fig. 3.19b). This pattern is even more strongly-pronounced in the z850 anomalies (Fig. 3.19c). Other features include positive SHR850, PWAT, and 2-m dew point anomalies over the Midwest and Plains (Fig. 3.19f,g,i), suggesting the presence of northward moisture transport. Interestingly, despite the anomalously strong SHR850, the SHR500 is very weak in node 7 (Fig. 3.19e), suggesting that the strongest shear in these cases may be limited to the lower levels in this node. SHR500 has been shown to be an important variable to the CSU-MLP tornado forecasts (Mazurek et al., 2025), so perhaps the model struggles with this prediction task when SHR500 is weak.

Among the best-performing CSU-MLP wind forecasts, the forecasts with the highest mean BSS tended to occur in node 5 (Fig. 3.20). The wind reports associated with the best-performing wind forecasts occur almost exclusively across the Southeast (Fig. A.25f). CSU-MLP wind probabilities also occur in this vicinity as well as across the Midwest and Plains, suggesting that the above-average skill among the node 5 wind forecasts may be limited to forecasts over the Southeast (Fig. 3.11f). However, it is worth mentioning that there are relatively few node 5 forecasts (16 total, with 4 being classified as “best-performing"), so conclusions may be limited. Despite the limited number of cases, there is a clear pattern in the z500 and z850 composites, including anomalously low heights over the eastern third of CONUS (Fig. 3.20b,c). Composites of other

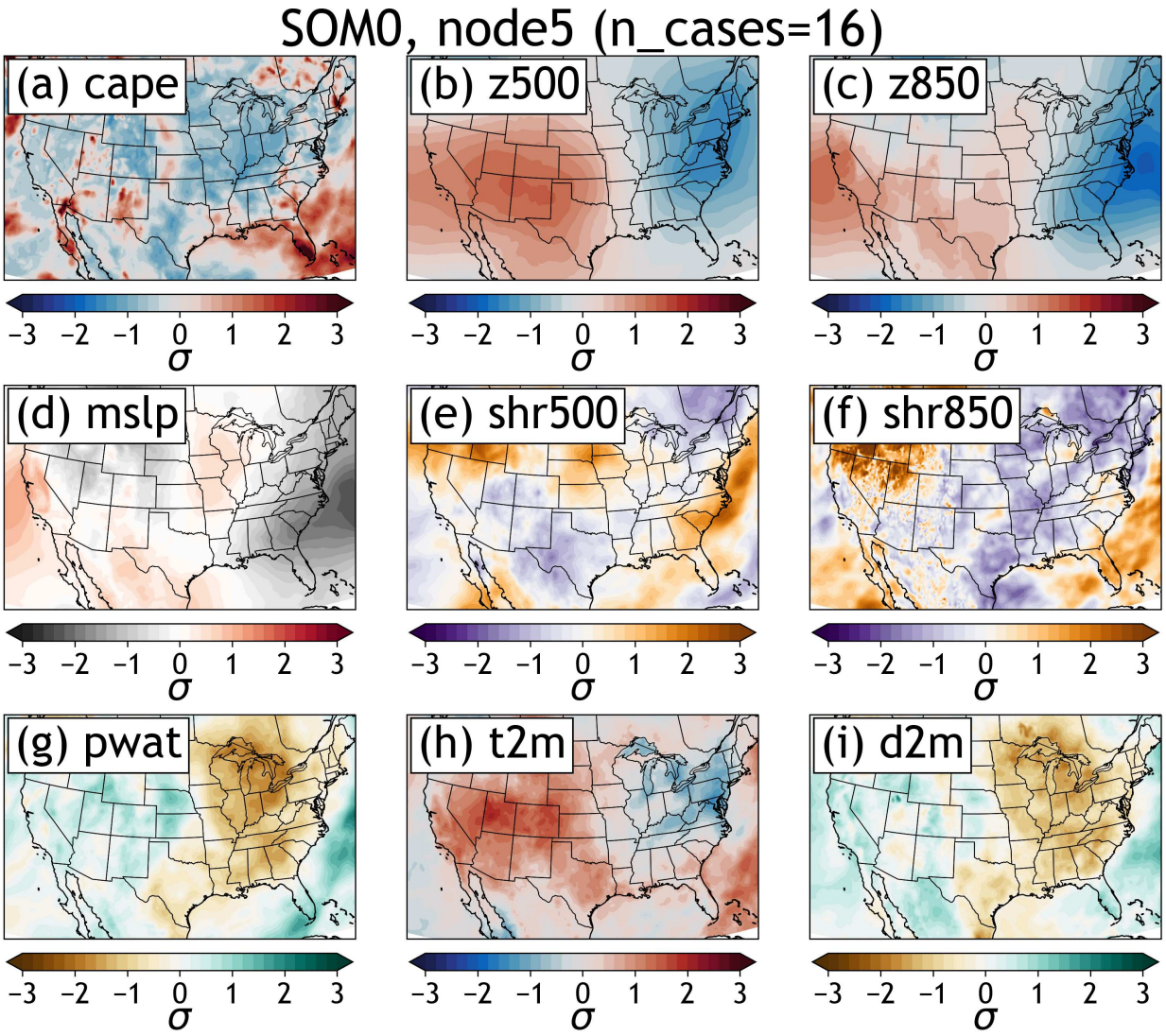

Figure 3.20: As in Fig. 3.17, but for node 5. The best-performing CSU-MLP day-2 wind forecasts tend to be associated with node 5 regimes.

fields offer fewer details, though it seems there may be enhanced deep-layer shear (SHR500) over the Southeast, near where the CSU-MLP wind forecasts and associated reports tend to occur.

Lastly, some of the worst-performing CSU-MLP wind forecasts tend to occur in node 1 of SOM0 (Fig. 3.21). CSU-MLP wind probabilities in node 1 occupy a notable fraction of total probabilities over the Southeast, Mid-Atlantic and Ohio Valley (Fig. 3.11b). Reports associated with the worst-performing node 1 forecasts mostly occur across the South, though there are a number of cases that also occur in the western CONUS (Fig. A.25b). Node 1 forecasts occupy the largest fraction of DJF forecast cases compared to other seasons (Fig. 3.8a), which is also the case among the worst-performing CSU-MLP wind forecasts specifically (Fig. 3.16f). Node 1 is characterized by anomalous CAPE, SHR850 and PWAT over Southeast CONUS (Fig. 3.21a,f,g).

3.5 Summary and Conclusion

This work utilizes self-organizing maps, or SOMs, to study variability in forecast skill of two years of day-2 probabilistic random forest-based severe weather forecasts for tornadoes, wind, and hail from the Colorado State University Machine Learning Probabilities (CSU-MLP) system. Key points from this research can be summarized as follows:

- The SOMs trained on surface-based CAPE and SHR850 (SOM0) and surface-based CAPE and SHR500 (SOM1) are both effective at diagnosing nine synoptically-distinct regimes. There is some overlap in node characteristics between the two SOMs, though each SOM also has nodes with unique characteristics.

- SHR850 composites overall have stronger, more definitive anomalies than the SHR500 composites in both SOMs, suggesting that SHR850 may be a better parameter to use for regime diagnostics.

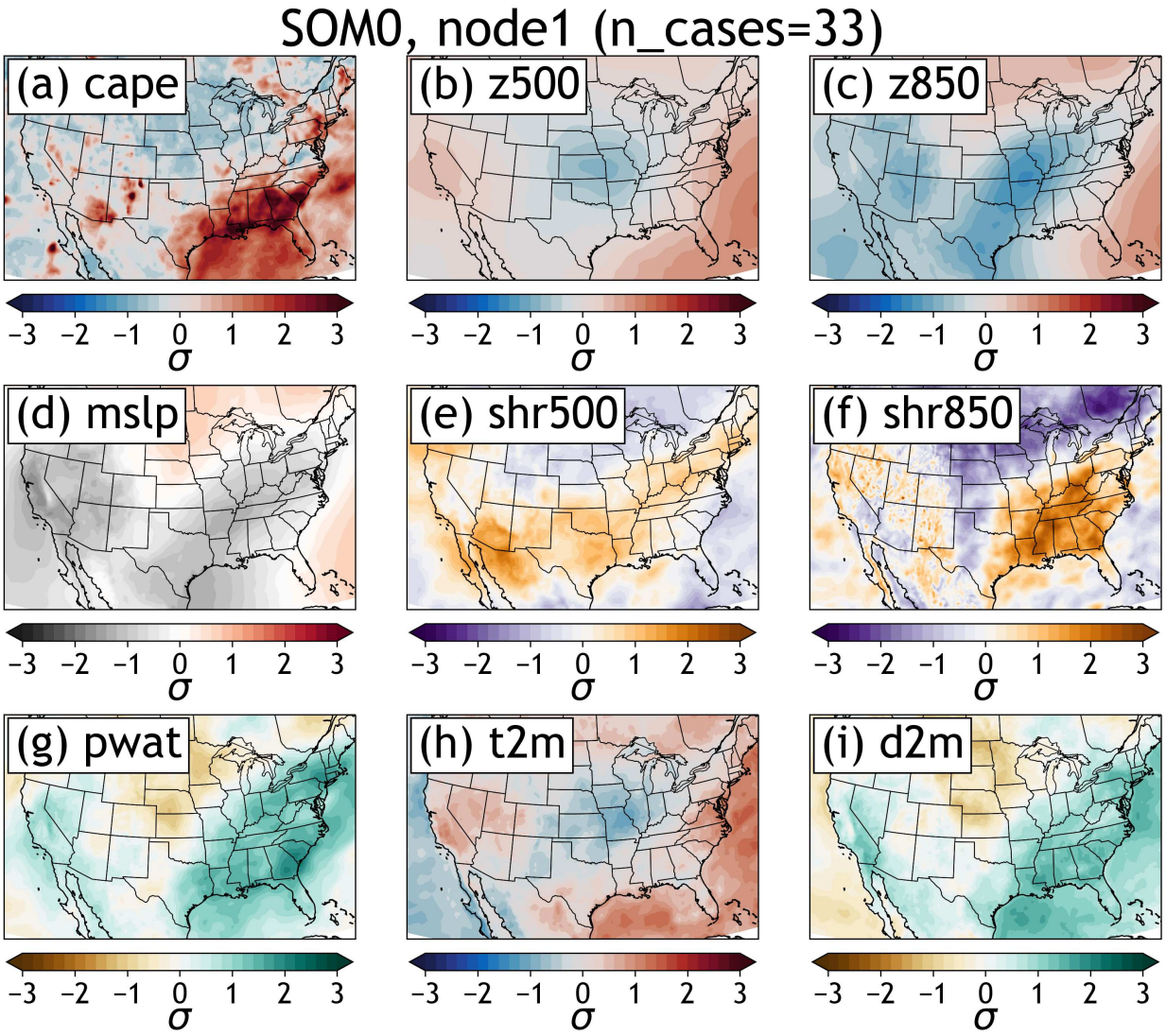

Figure 3.21: As in Fig. 3.17, but for node 1. The worst-performing CSU-MLP day-2 wind forecasts tend to be associated with node 1 regimes.

- CSU-MLP forecasts exhibit some spatial variability across the nodes, suggesting that some of the SOM-identified regimes tend to be correlated with severe weather events in certain locations.

- There is variability in the relative frequency of the most and least skillful forecasts across the nodes. This result at times varies between CSU-MLP forecasts and SPC guidance.

- Some nodes have a tendency to be associated with *both* disproportionately more most- *and* least-skillful forecasts compared to the other nodes, suggesting that forecast performance in some nodes can fluctuate substantially (despite similar environmental conditions).

- Some nodes capture both some of the best- and worst-performing forecasts among the whole dataset, which further supports the previous claim. The CSU-MLP hail forecast performance in SOM0's node 3 exemplifies this finding.

- Forecast skill in some nodes remains high even when the worst-performing forecasts are considered. The skillful CSU-MLP wind forecast performance across many of the SOM0 nodes, especially nodes 5 through 7 support this observation.

- The most skillful CSU-MLP and SPC forecasts tend to be associated with more storm reports than the least skillful forecasts. This result suggests that the worst-performing forecasts tend to stem from either missed isolated reports or false alarms (rather than incorrect contour placement during events with many reports, for instance).

More broadly, this work offers valuable insights on the CSU-MLP system's predictability across different environments. Understanding scenarios where the model tends to make skillful versus less skillful predictions could allow the products to be used more strategically. Thus, these

results may be useful to forecasters who actively use or would like to use CSU-MLP guidance in their day-to-day predictions. More broadly, by establishing relationships between synoptic patterns and forecast skill, this work classifies connections between large-scale environments and characteristics of subsequent severe weather events. These results could help inform future studies aiming to better understand how subtle differences in large-scale atmospheric patterns are correlated with severe storm outcomes.

In this study, SOMs are shown to be a useful tool here for examining ML forecast performance across different weather regimes. There are a number of avenues for future work that are inspired by this research. For example, experiments with different SOM configurations could be conducted, such as adding or substituting environmental fields used in SOM training (e.g., moisture variables). Different combinations of variables could also be used to train separate SOMs for examinations of various hazard types (e.g., a SOM trained on downdraft CAPE may prove useful for assessing skill of probabilistic CSU-MLP wind forecasts specifically). Best- and worst-performing CSU-MLP forecasts could also be used to stratify SOM inputs before training; this approach could yield more granular insights on the relationships between environmental characteristics and CSU-MLP forecast performance. Adding more CSU-MLP forecasts to the dataset could also clarify findings for a similar reason. Additionally, future work could aim to better correlate the timestamps of input fields with the onset of severe storms (rather than using data at a single timestamp). Doing so may illuminate stronger relationships between CSU-MLP forecast skill and timing of observed severe weather. SOMs could also easily be used to evaluate CSU-MLP forecasts at longer lead times (particularly in the medium range) or other ML-based prediction systems altogether. Lastly, while this work has focused on using SOMs to study CSU-MLP forecast skill, it could also be used to

study the skill of the GEFS itself. Understanding areas of strong/weak predictability in the GEFS could illuminate regimes under which there are deficiencies in the ensemble's ability to forecast the environment (which inevitably would impact the CSU-MLP system's forecasting capabilities).

Chapter 4

Analyzing Derived Convective Parameters from Deep Learning Weather Prediction Models

4.1 Introduction

Artificial intelligence (AI)-driven weather forecasting systems are rapidly being developed across numerous agencies, particularly over the last couple of years. Many of these models harness deep learning machine learning (ML) approaches (especially deep learning) to emulate numerical weather prediction (NWP) output. Such methods offer computationally-savvy, data-driven alternatives to traditional methods that rely on prognostic equations and parametrizations. As data-driven methods, these deep learning weather prediction (DLWP) models are diverse in their training approaches, including relying on observations (e.g., Vaughan et al., 2024), reanalysis (e.g., Pathak et al., 2022; Bi et al., 2023; Bonev et al., 2023; Chen et al., 2023; Lam et al., 2023; Schmude et al., 2024; Zhong et al., 2024), or some combination of reanalysis and NWP output (e..g, Bodnar et al., 2024; Lang et al., 2024; Pathak et al., 2024).

DLWP seeks to provide both speed and accuracy advantages over NWP models. The initial training of DLWP models is extremely computationally expensive, and due to their complexity relative to other ML methods, substantial energy resources are needed (Xu et al., 2021). However, once trained, DLWP models are extremely fast and efficient: Google's GraphCast (Lam et al., 2023), for example, can generate a 10-day global forecast in less than a minute–a task that would require substantially more time in a physics-based system. With increased computing capabilities accompanied by longer historical data records (exemplified by the recent extension of ERA-5 back

to 1940, for instance; Hersbach (2023)), DLWP model development is more attainable now than ever before. And in terms of accuracy, developers of these models have demonstrated that their forecasts are competitive with forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF)'s Integrated Forecasting System (IFS) (e.g., Bi et al., 2023; Lam et al., 2023; Lang et al., 2024)–considered to be the gold standard for operational NWP (Rasp et al., 2024). From a speed and accuracy perspective, it seems that DLWP models are ready to compete with physics-based weather models.

However, while evaluations of DLWP systems from model developers are optimistic, many additional comprehensive, operationally-focused evaluations are needed in order to identify their fitness for various forecasting challenges. These types of studies can elucidate model behaviors and provide application-based results that could make it easier for forecasters to increase their literacy on the output (e.g., Ebert-Uphoff and Hilburn, 2023). To support these efforts, a number of these DLWP systems have begun to be studied and leveraged for operational use (e.g., Ben Bouallègue et al., 2024) and specific prediction tasks that have relevance to forecasters. For example, Feldmann et al. (2024) computes most unstable convective available potential energy (MUCAPE) and shear parameters from a number of these models to understand how they model pre-convective environments. Their work finds that deriving MUCAPE and shear from DLWP forecasts does produce reasonable results across multiple global regions, though the skill of these parameters varies substantially across each modeling system. DeMaria et al. (2024) evaluates DLWP model forecast skill of tropical cyclones, showing that they are competitive against current NWP-based systems for predicting their tracks but poor at forecasting intensity. Olivetti and Messori (2024) explores yet another complex forecast problem by examining whether DLWP models are capable of predicting

extreme temperatures and winds near the surface. Their work shows that DLWP models can produce skillful forecasts for extremes, however their skill varies substantially on location, magnitude of the extremes, and other factors.

These works motivate the need to examine DLWP output for specific forecast problems, and this study aims to contribute to furthering this research. Like Feldmann et al. (2024), this work examines convective environments and parameters in output from several DLWP models. However, this research differentiates itself from that study in a few ways, such as the initial conditions, study period, and types of convective parameters used in the analysis. Additionally, this study is framed around determining the fitness of DLWP for day-to-day forecasting tasks, and emphasizes forecast metrics to a smaller degree.

In this study, 22 months of forecasts from three DLWP systems, Pangu-Weather (Bi et al., 2023), GraphCast Operational (Lam et al., 2023), and FourCastNetv2-small (Bonev et al., 2023), are examined. These forecasts are run courtesy of the Cooperative Institute for Research in the Atmosphere (CIRA) using initial conditions from the Global Forecasting System or GFS (Radford et al., 2025). A combination of native output variables and derived convection-pertinent variables (specifically, precipitable water, convective available potential energy, and deep-layer wind shear) are compared to ERA-5 reanalysis and output from an operational physics-based model (i.e., the GFS). Differences between the forecasts and reanalysis are studied seasonally and across different convective events. This approach therefore provides both a big picture view of what derived convective parameters generally look like from these three DLWP models, as well as investigates how operationally-informative they may be for day-to-day forecasting. Thus, *the overarching goal of this work is to examine differences in the derived convective parameters and environments between*

the DLWP model forecasts, ERA-5 reanalysis, and GFS forecasts, both broadly and for specific severe convective events. The hope is that this research will help catalyze future investigations into the potential utility of DLWP model output for operational forecasting.

This chapter continues as follows. In section 2, the DLWP model architectures, forecast data, and analysis strategies are introduced. Section 3 highlights results for two years of forecasts as well as select severe weather cases. Discussion, conclusion, and avenues for future work are included in section 4.

4.2 Data and methods

4.2.1 Deep learning weather prediction models

The three DLWP models used here are similar in that all three are trained on ERA-5 reanalysis (Hersbach et al., 2020) and are capable of producing global weather forecasts out to 10 days. However, there are differences among their ML architectures.

FourCastNetv2 is a second generation version of its predecessor model, FourCastNet (Pathak et al., 2022), which is regarded as the first DLWP model able to generate global forecasts that were competitive with NWP. The original version of FourCastNet, developed by NVIDIA, is a vision transformer model that relies on an Adaptive Fourier Neural Operator (AFNO; Guibas et al., 2022)-type architecture to make its forecasts (Pathak et al., 2022). However, this architecture led to unstable performance when used with spherical coordinates, so a Spherical Fourier Neural Operator (SFNO) approach was designed and used to develop FourCastNetv2, which led to more stable and accurate results (Bonev et al., 2023). Pangu-Weather was developed at Huawei Cloud soon after the original FourCastNet. It distinguishes itself from predecessor DLWP in two ways. First, its architecture is characterized by a novel three-dimensional Earth-specific vision transformer that

allows for three dimensional data to be directly input to the neural network (i.e., dimensionality reduction is not needed). Second, they employ hierarchical temporal aggregation, which reduces the number of time iterations needed to make a forecast and has speed and accuracy benefits, particularly at longer lead times (Bi et al., 2023). GraphCast is Google's reanalysis-trained DLWP model, and it relies on Graph Neural Networks or GNNs (Lam et al., 2023). The GNNs in GraphCast are defined by nodes and edges, where the nodes contain information about the state of the atmosphere and the edges describe the relationships between the adjacent nodes (which presents itself as a graph-like structure). In addition to its architecture, GraphCast further distinguishes itself from Pangu-Weather and FourCastNetv2 in its temporal feature assembly approach by considering data from two timesteps prior to the forecast valid time (rather than only one timestamp) as inputs. The operational version of GraphCast, which is the version studied here, is fine-tuned from its parent model with several years of output from the ECMWF's deterministic IFS model in its training, meaning that it is not solely trained on reanlaysis. Since only the operational version of GraphCast will be used in this study, it will be referred to throughout as GraphCast.

4.2.2 Generating forecast data

Open-source code provided by developers of all three DLWP models used in this study allows for public users to download the already-trained models to their own machines. By removing computationally-intensive training requirements, the average user can run their own forecasts with relative ease. Researchers at the Cooperative Institute for Research in the Atmosphere (CIRA) are one group that has harnessed these resources to generate forecast output (Radford et al., 2025).

Instead of initializing forecasts using ERA-5 reanalysis (which all three models are trained on), CIRA DLWP forecasts are forced using initialized conditions from the Global Forecast System (GFS). Using inputs from an operational numerical weather prediction (NWP) model allows for predictions to be made in real-time, which offers benefits to real-world forecasting scenarios that require timely model output. At the timing of writing this dissertation, CIRA forecasts are actively being run for the three DLWP models examined in this study (as well as others) two to four times per day (i.e., every 6 to 12 hours), using GFS initial conditions. Additionally, retrospective forecasts have been processed using GFS initializations back to at least 1 January 2022, with archived forecasts from Pangu-Weather and FourCastNetv2 going back even further[12]. Real-time forecasts are publicly available[13], and retrospective forecasts can be accessed via Amazon Web Services[14].

An overview of the native variables output in the CIRA DLWP forecasts is shown in Table 4.1. Note that some variables differ from model-to-model, as each of them are trained to predict a slightly different suite of variables. Vertical pressure levels are kept consistent to 13 across the three models, and forecasts have a quarter-degree spatial resolution that matches both the ERA-5 and GFS grids.

As seen in Table 4.1, the forecast variables are fundamental but limited in use for studying convective environments. As such, a number of derived parameters that have relevance to convective forecasting were generated from the CIRA DLWP forecasts. These variables are listed in

[12]GraphCast forecasts prior to 1 January 2022 were not generated, as the GraphCast Operational training dataset includes data through the end of 2021, and thus forecast inputs prior to 2022 would not be independent from the training data.

[13]https://aiweather.cira.colostate.edu/

[14]https://noaa-oar-mlwp-data.s3.amazonaws.com/index.html

Table 4.1: Native variables in archived CIRA DLWP forecasts (Radford et al., 2025). Variables that are not included across all three models are bolded.

Model	Vertical levels (hPa)	Pressure-level variables	Single-level variables
GraphCast Operational	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, and 50	geopotential height, temperature, u- and v-wind, **specific humidity**, **vertical velocity**	10m u- and v-wind, 2-m temperature, MSLP, **6-h precipitation**
Pangu-Weather	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, and 50	geopotential height, temperature, u- and v-wind, **specific humidity**	10m u- and v-wind, 2-m temperature, MSLP
FourCastNetv2-small	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, and 50	geopotential height, temperature, u- and v-wind, **relative humidity**	10m u- and v- wind, **100m u- and v-wind**, 2-m temperature, MSLP, **surface pressure**, **total column water vapor**

Table 4.2. Given the limited vertical resolution of the DLWP output, it should be noted that some of these parameters use very few vertical levels in their calculations. This limited resolution obviously limits the quality of the output and will be discussed in greater detail later. These derived parameters are not currently generated from the forecast output in real-time, but archived data are accessible via the aforementioned Amazon Web Services bucket.

4.2.3 Analysis techniques

In this study, daily forecasts initialized between 1 January 2022 to 31 October 2023 are examined, yielding 22 months of data. While records of the native variables extend beyond that, there remains a gap in the derived parameters between 1 November 2023 to 31 December 2023, and so the study is restricted to just shy of two years. This limitation is not overly concerning, given that severe weather activity is much lower during the cool season versus the warm season. The analysis

Table 4.2: Derived convective parameters available in the CIRA DLWP archive. Variables examined in this study are bolded.

Derived convective parameters
convective available potential energy (**surface-based**, mixed-layer, most-unstable)
convective inhibition (surface-based, mixed-layer, and most-unstable)
lifted index
planetary boundary layer height
precipitable water
storm relative helicity (0-1 km, 0-3 km)
Bunkers right-mover
vertical wind shear (850 hPa-200 hPa, surface-850 hPa, **surface-500 hPa**)

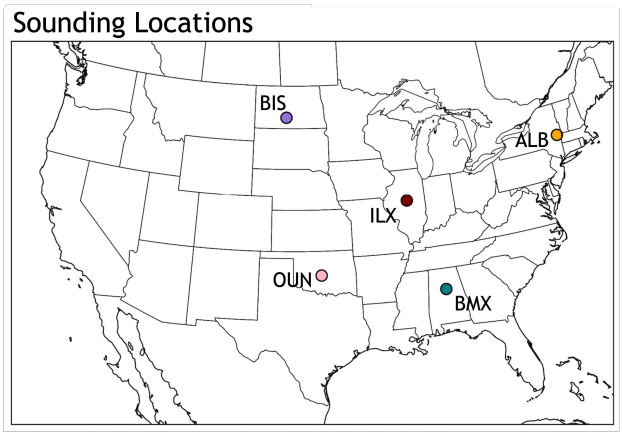

Figure 4.1: Selected point sites for the study, which are strategically co-located with upper-air stations: Bismarck, ND (BIS), Norman, OK (OUN), Lincoln, IL (ILX), Birmingham, AL (BMX), and Albany, NY (ALB).

is restricted within a set of boundaries that roughly encompass CONUS (i.e., 60°W to 135°W and 20°N to 55°N) and forecasts valid at 0000 UTC are primarily focused on throughout this work (in an effort to best-align the characteristics of the environment with the peak timing of severe storms over CONUS). Both native and derived variables from the DLWP models are compared to ERA-5 reanalysis (Hersbach et al., 2020) and operational GFS forecasts in various capacities.

Forecasts are analyzed in three ways across this study. First, differences between the DLWP derived parameters and ERA-5 derived parameters are studied across the full 22-month dataset. This manuscript evaluates these differences in surface-based convective available potential energy

Table 4.3: Number of "convectively favorable" forecasts at each site of interest. Convectively favorable forecast days are defined by days when an SPC enhanced risk or greater intersects the County Warning Area of the National Weather Service office that is co-located with the given site.

Location	Number of forecasts
ALB	6
BIS	8
BMX	23
ILX	15
OUN	38

(SBCAPE)[15], precipitable water (PWAT), and vertical wind shear between the surface and 500hPa layer. These variables are chosen as proxies for instability, moisture, and shear, which are three critical ingredients for severe convection. In this analysis, as well as throughout the manuscript, the ERA-5 fields are computed only with the 13 vertical levels that are in the DLWP output in an effort to provide a more fair comparison between the two datasets. For some analyses, the ERA-5 land-sea mask is applied, and only points containing land are included. These distinctions are noted in the figure captions where appropriate. The objective of this analysis is to understand how the DLWP derived parameters generally differ from ERA-5 parameters during different times of year as well as spatially.

The second analysis in this study examines forecast data from five geographically-diverse sites over CONUS (Fig. 4.1). In addition to their geographic locations and propensity for severe thunderstorms, these sites were specifically selected because they are co-located with upper-air sites (which has particular relevance to the third analysis), and they are also co-located with National Weather Service Weather Forecast Offices (NWS WFOs). This latter point was important for

[15]SBCAPE has limitations, particularly in that it fails to capture nocturnal instability when the boundary layer cools. However, given that this analysis studies output valid at 0000 UTC, these concerns should be mostly mitigated.

the next part of the methods, which seeks to limit this part of the analysis only to convectively-favorable days. “Convectively-favorable days" were determined using the Storm Prediction Center (SPC) convective outlooks as a proxy: days are considered to be “convectively-favorable" when a day-1 enhanced risk or greater (NOAA Storm Prediction Center, 2023b) intersects the county warning area associated with the site. These cases were identified using the IEM automated data plotter[16]. The enhanced categorical risk was chosen in an effort to strike a balance between only examining “higher-end" convective environments (rather than more marginal cases) while also not ruling out too many cases such that the sample size is small. Over the 22 months of forecasts, the method resulted in 90 cases across the 5 sites, with notable variability primarily due to differences in severe storms climatology (Table 4.3). Note that there are some forecast days that are considered a qualifying event at multiple sites, but given the distances between the sites and assumed environmental differences between them, the cases are considered separately.

Using this dataset of convectively-favorable days, vertical profiles of the atmosphere (specifically temperature and dew point[17]) are examined for each of the sites. These vertical profiles are compared both to ERA-5 as well as GFS output across varying lead times. GFS vertical levels are restricted to 13 in the same way as ERA-5 for this analysis. The goal of this assessment is to examine how the DLWP vertical profiles might depart from reanalysis and operational guidance for events favorable for severe convection.

In the third analysis, DLWP forecasts for two case studies, chosen for their large differences in synoptic characteristics, geographic location, and intensity, are presented. Forecasts from the

[16]https://mesonet.agron.iastate.edu/plotting/auto/?q=200

[17]Dew points are derived from either the specific humidity or relative humidity in the DLWP models.

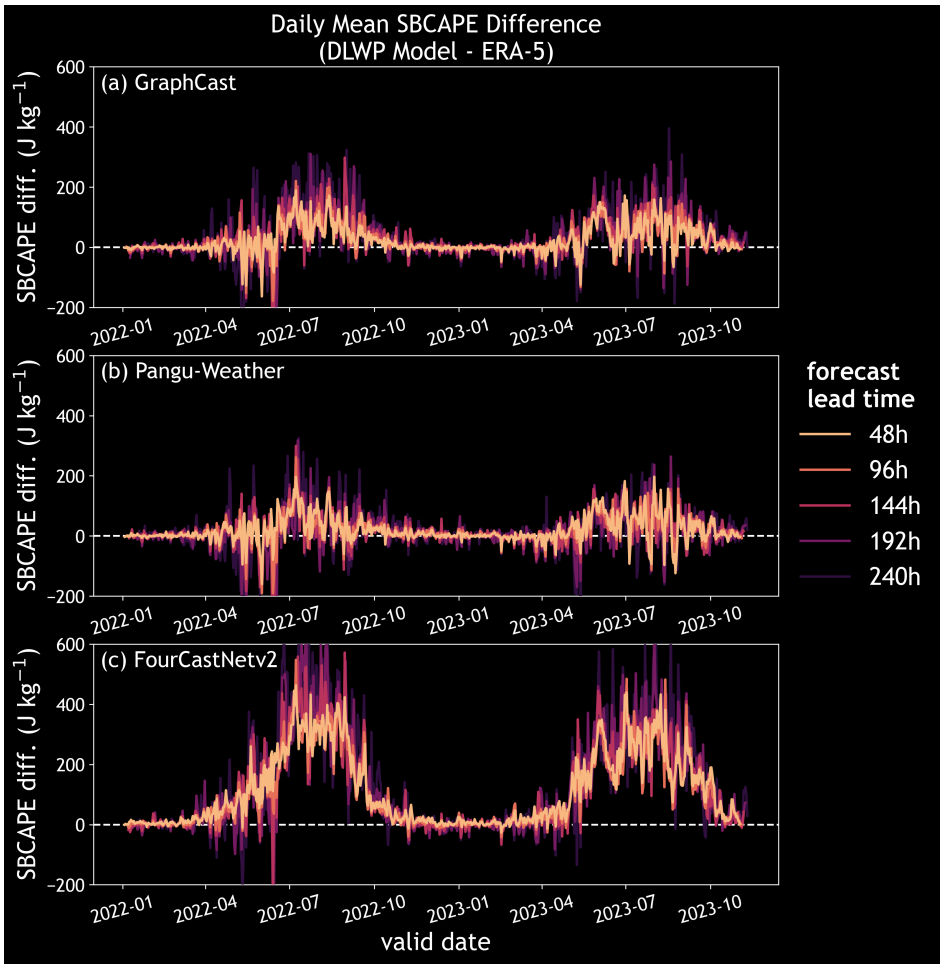

Figure 4.2: Daily mean difference in surface-based convective available potential energy (SBCAPE) between ERA-5 reanalysis and (a) GraphCast, (b) Pangu-Weather, and (c) FourCastNetv2-small for forecasts initialized at 0000 UTC between 1 January 2022 and 31 October 2023. Data are plotted according to valid time. Only grid points that contain land within the “CONUS" bounds defined in the methods are considered. For each panel, lines are colored according to the forecast lead time listed in the legend.

DLWP models are compared to operational GFS guidance (with all levels retained) as well as ERA-5. Convective parameters, soundings, and hodographs are compared among the three datasets. Sounding and hodograph data are analyzed at the aforementioned point locations where appropriate, and observed soundings from the sites are used as another comparison point.

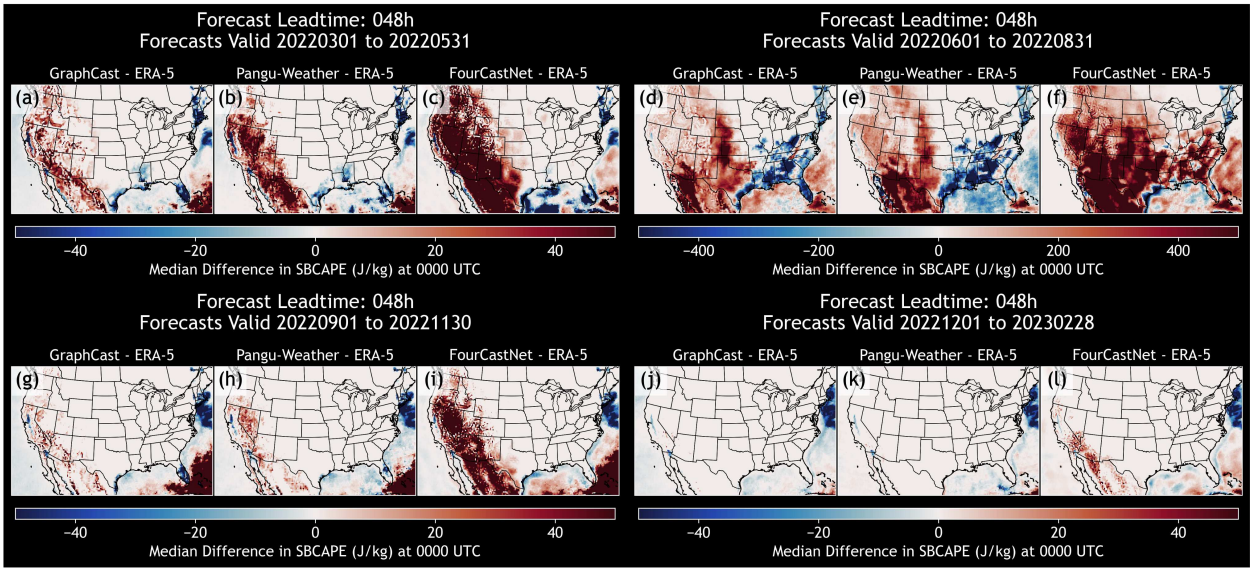

Figure 4.3: Median differences in derived surface-based CAPE (SBCAPE) between the three DLWP models and ERA-5 reanalysis for (a)-(c) MAM 2022, (d)-(f) JJA 2022, (g)-(i) SON 2022, and (j)-(l) DJF 2022-2023. DLWP forecasts are issued at a 48-h lead time, and forecasts and reanalysis are valid at 0000 UTC. Note the JJA use a different color bar than the other three seasons.

4.3 Results

4.3.1 Seasonal characteristics of derived convective parameters

Beginning with examining daily mean differences in values of SBCAPE (Fig. 4.2), there is clear seasonality in the degree to which the DLWP forecasts differ from ERA-5. In all three models, the daily mean differences are largest in the late spring through early fall, with a peak around August. Differences are small in the winter season, likely because surface-based instability is very small or non-existent over CONUS at that time of year. The differences in GraphCast and Pangu-Weather are similar (Fig. 4.2a,b), and it appears that both DLWP models tend to generally forecast too much SBCAPE relative to ERA-5. FourCastNetv2 also has this tendency, but to a more extreme degree (Fig. 4.2c). SBCAPE differences between ERA-5 and the DLWP models generally get smaller with shorter lead times across the three models.

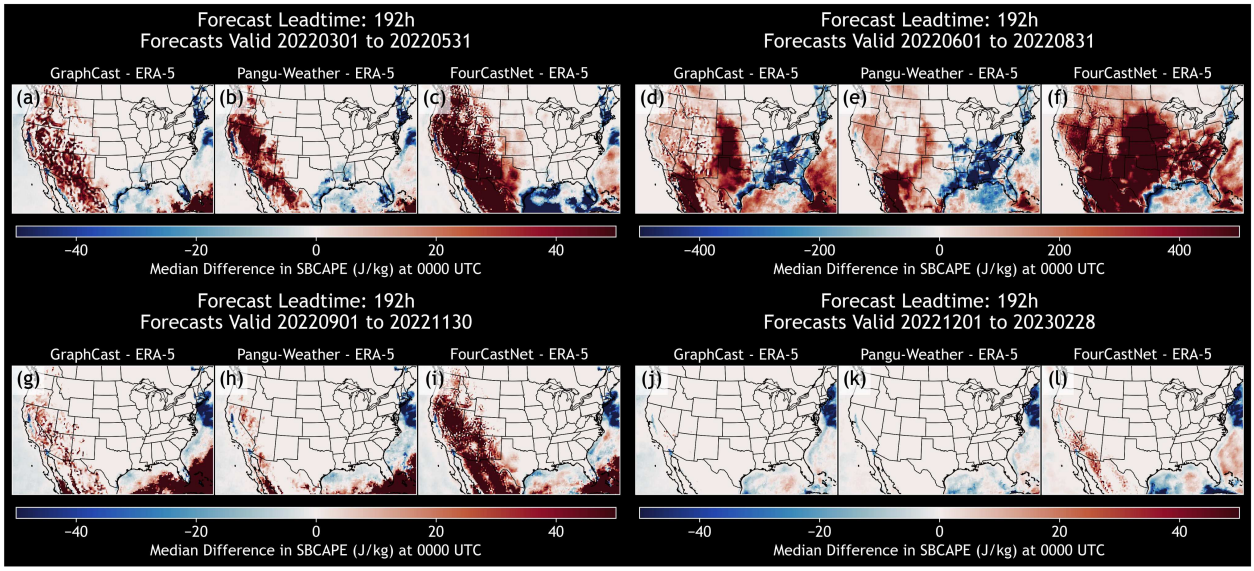

Figure 4.4: As in Fig. 4.4, but for SBCAPE derived from the 192-h forecasts.

Examining median SBCAPE differences between the DLWP and ERA-5 in map view across different seasons can provide additional insights (Figs. 4.3; 4.4). As anticipated, differences are very small over land areas during DJF (Fig. 4.3j,k,l), while differences are very large in JJA (Fig. 4.3d,e,f). However, in the summer months, the magnitude and sign of the SBCAPE differences varies substantially across CONUS. Both GraphCast and Pangu-Weather show a tendency for the DLWP models to produce less SBCAPE than ERA-5 over the Southeast and Ohio Valley, while they tend to produce more over the Great Plains (Fig. 4.3d,e). FourCastNetv2 SBCAPE values, on the other hand, are larger than ERA-5 almost everywhere across CONUS during the summer months (Fig. 4.3d,e), which explains the differences in the daily means between those forecasts and the other two DLWP models (Fig. 4.2). Additionally, the derived SBCAPE from the DLWP over the Rockies is very noisy and tends to be larger than ERA-5 values there, particularly at longer lead times (e.g., c.f. Figs. 4.3a,b,c; 4.4a,b,c). The same number of levels are used to compute SBCAPE in the DLWP models and ERA-5 reanalysis here, so these differences cannot

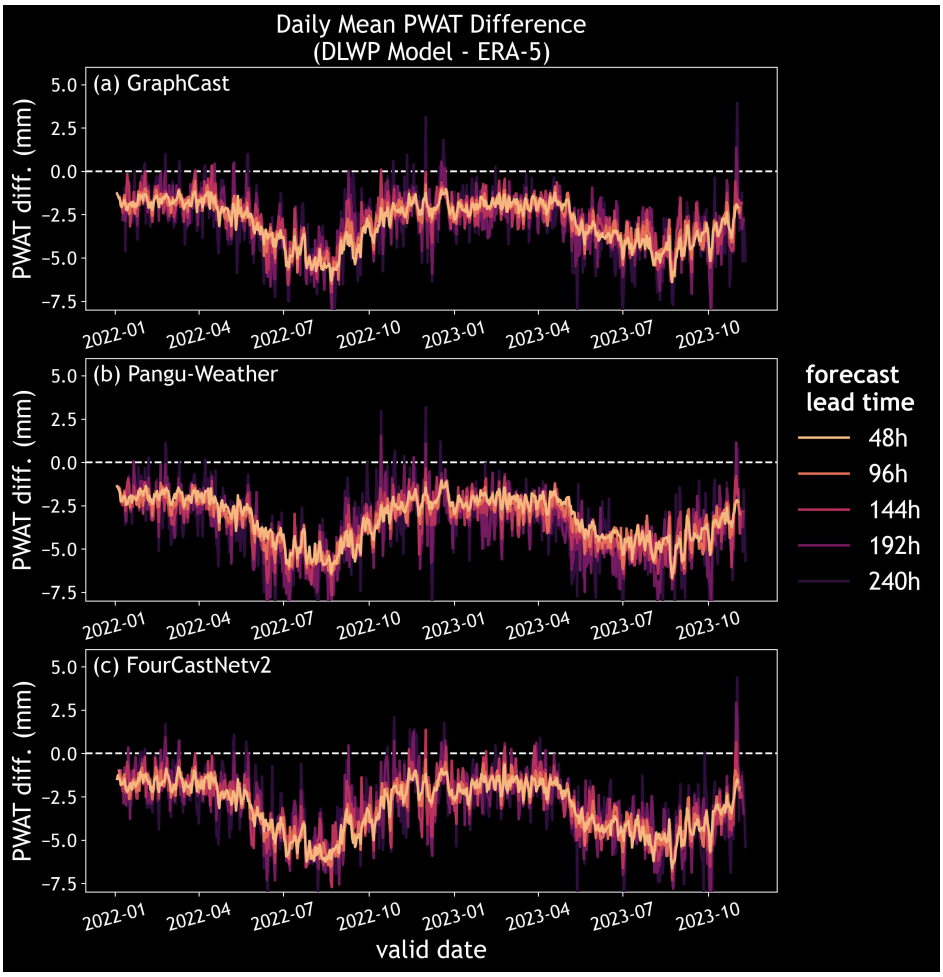

Figure 4.5: As in Fig. 4.2, but for precipitable water (PWAT).

be simply be attributed to the low vertical resolution in the DLWP models. Regardless, these figures show that the differences in SBCAPE between the forecasts and ERA-5 can be substantial, at times more than 500 J kg^{-1} during the summer months. If ERA-5 is taken as “truth", it is likely DLWP SBCAPE fails to properly characterize the convective environment.

Like SBCAPE, mean daily differences in PWAT between ERA-5 and the DLWP models are largest during the late spring to early fall across all three models (Fig. 4.5), with PWAT in the DLWP models being generally less than ERA-5. PWAT remains lower in the DLWP models in the cool season as well, though the daily mean differences are only on the order of a couple of

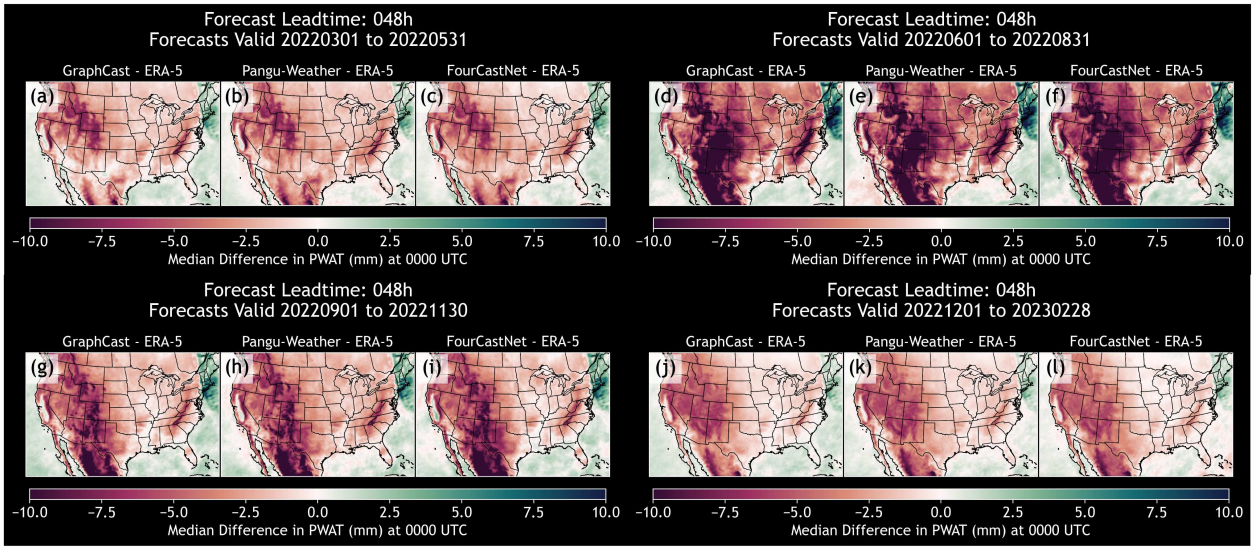

Figure 4.6: Median differences in derived precipitable water (PWAT) between the three DLWP models and ERA-5 reanalysis for (a)-(c) MAM 2022, (d)-(f) JJA 2022, (g)-(i) SON 2022, and (j)-(l) DJF 2022-2023. DLWP forecasts are issued at a 48-h lead time, and forecasts and reanalysis are valid at 0000 UTC.

mm. There are not obvious differences between the models both in these daily means, nor in the seasonal medians (Fig. 4.6). In fact, there are very few differences in the three models with regards to the magnitude of the PWAT differences across various regions and seasons. Like with SBCAPE, all models show the largest differences between ERA-5 and the DLWP models over the western CONUS. Notable differences can also be seen over the Appalachians, suggesting that even though the levels are consistent across reanalysis and the models, elevation seems to have some effect on the PWAT calculations across the two datasets. As is seen in Fig. 4.5, PWAT differences become larger with longer lead times, but the magnitudes of these differences are small and yield very few changes in the seasonal medians (not shown).

Daily mean SHR500 differences between the DLWP models and ERA-5 appear to show opposite seasonality compared to SBCAPE and PWAT. Fig. 4.7 illustrates that the largest shear differences between the DLWP forecasts and ERA-5 occur in mid-fall to mid-spring; this pattern is less obvious in the short-term forecasts but more obvious in the long-term ones. Given that

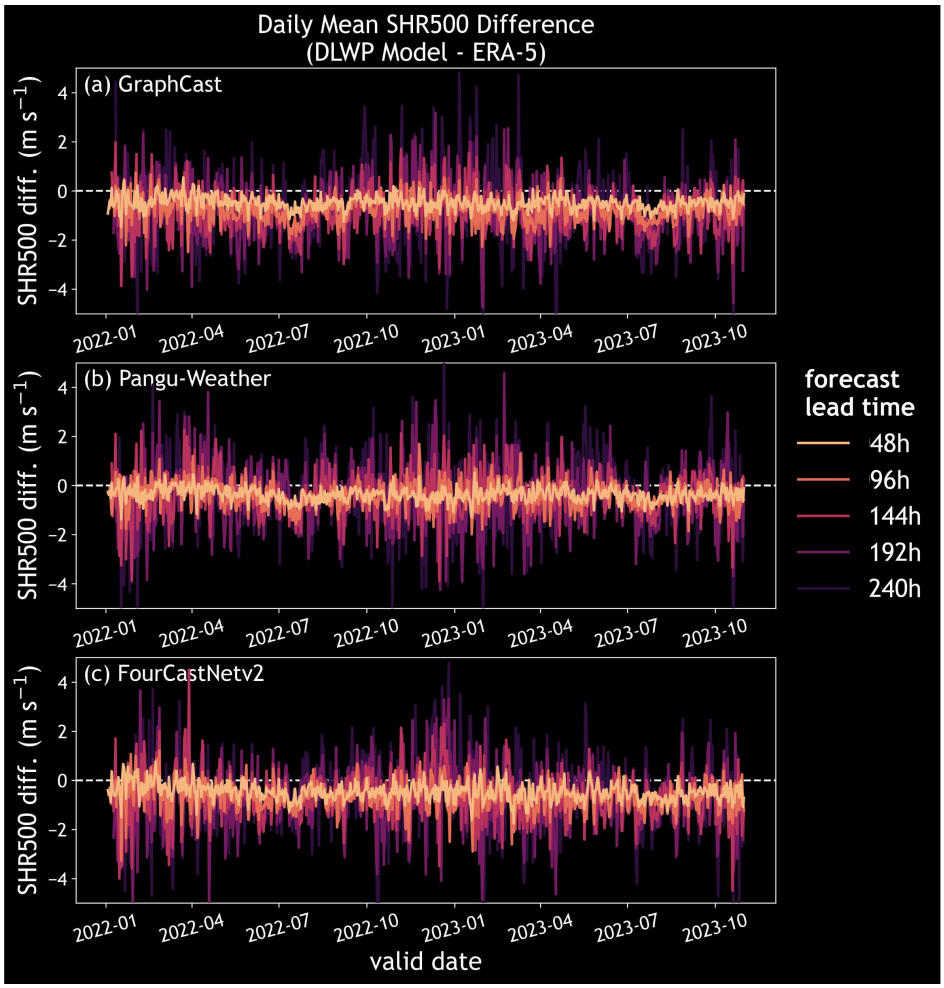

Figure 4.7: As in Fig. 4.2, but for surface to 500 hPa vertical wind shear (SHR500).

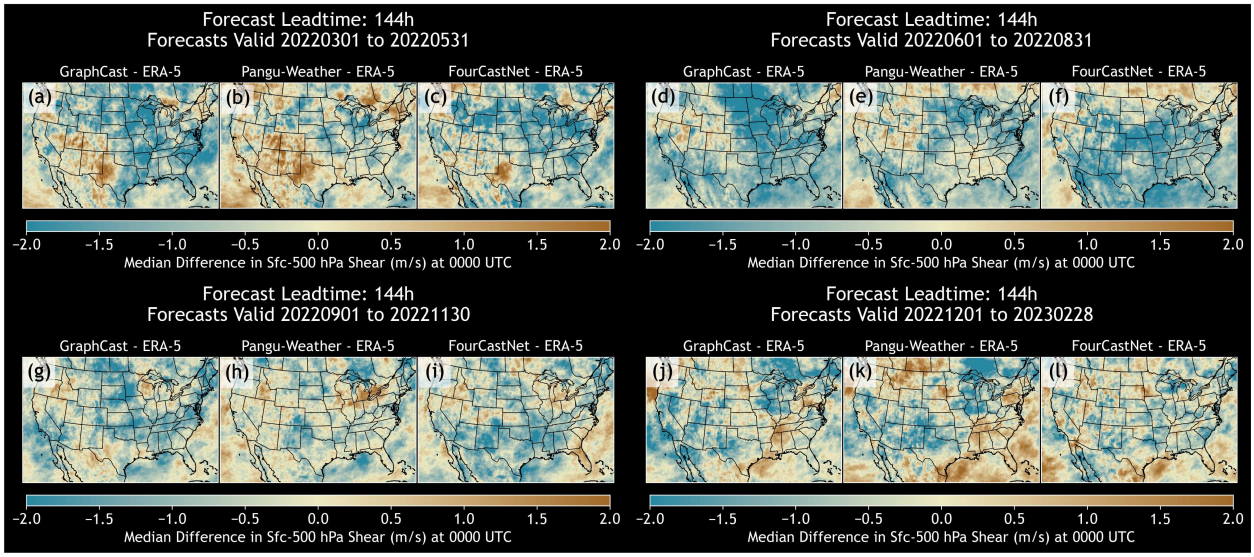

Figure 4.8: Median differences in derived 10m-500 hPa vertical wind shear (SHR500) between the three DLWP models and ERA-5 reanalysis for (a)-(c) MAM 2022, (d)-(f) JJA 2022, (g)-(i) SON 2022, and (j)-(l) DJF 2022-2023. DLWP forecasts are issued at a 144-h lead time, and forecasts and reanalysis are valid at 0000 UTC.

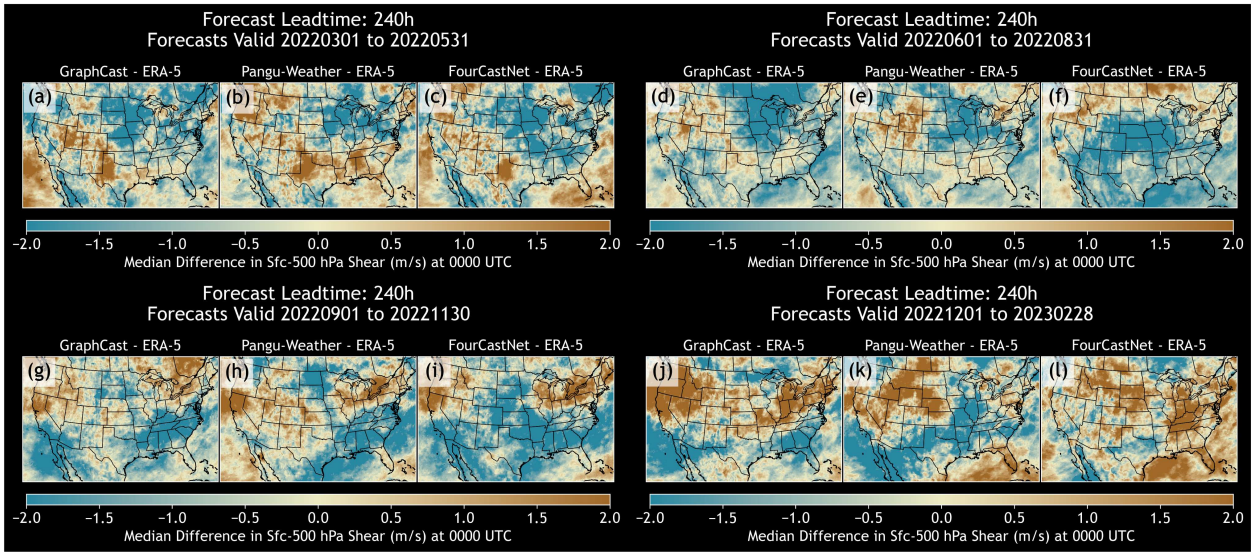

Figure 4.9: As in Fig. 4.8, but for SHR500 derived from the 240-h forecasts.

dynamics are generally weak in the warm season, this result is not overly surprising. For forecasts valid at shorter lead times, because the shear differences are so small, there are not obvious or persistent seasonal or regional differences in the SHR500 fields that stand out among the three models. However, when forecasts at longer lead times are evaluated, some subtle patterns begin to emerge. For example, examining the median seasonal differences in derived SHR500 between 6-day (Fig. 4.8) and 10-day (Fig. 4.9) DLWP forecasts and ERA-5 shows that in MAM and JJA, the DLWP models tend to have less shear than ERA-5 over the Midwest. This pattern is most obvious in GraphCast (Figs. 4.8a,d; 4.9a,d), but it can also be seen to some extent in Pangu-Weather (Figs. 4.8b,e; 4.9b,e). In the winter, when SHR500 differences tend to be greatest between the two datasets, the 10-day forecasts all show greater median values over the northwestern CONUS compared to ERA-5, particularly in GraphCast and Pangu-Weather (Fig. 4.9j,k). This pattern is not evident in the 6-day forecasts, but it can be seen in the 8-day forecasts (not shown). In the fall, Pangu-Weather and FourCastNetv2 seem to have greater SHR500 values compared to ERA-5 (Figs. 4.8h,i; 4.9h,i). Still, these patterns are much less clear than those in PWAT and SBCAPE, which could be attributed to the fact that shear is tied solely to kinematics (which the DLWP models have shown to be rather successful at capturing in past work) as opposed to thermodynamics (which the DLWP models have been shown to struggle with), so differences between the model output and reanalysis may more subtle and difficult to interpret in these cases. Still, other analyses may be better suited to assess SHR500 differences in greater detail.

4.3.2 Vertical profiles of temperature and dew point during severe weather events

In the previous section, it was shown that the PWAT in the DLWP models tended to be generally lower than ERA-5, and examining dew point profiles can help illustrate why these differences

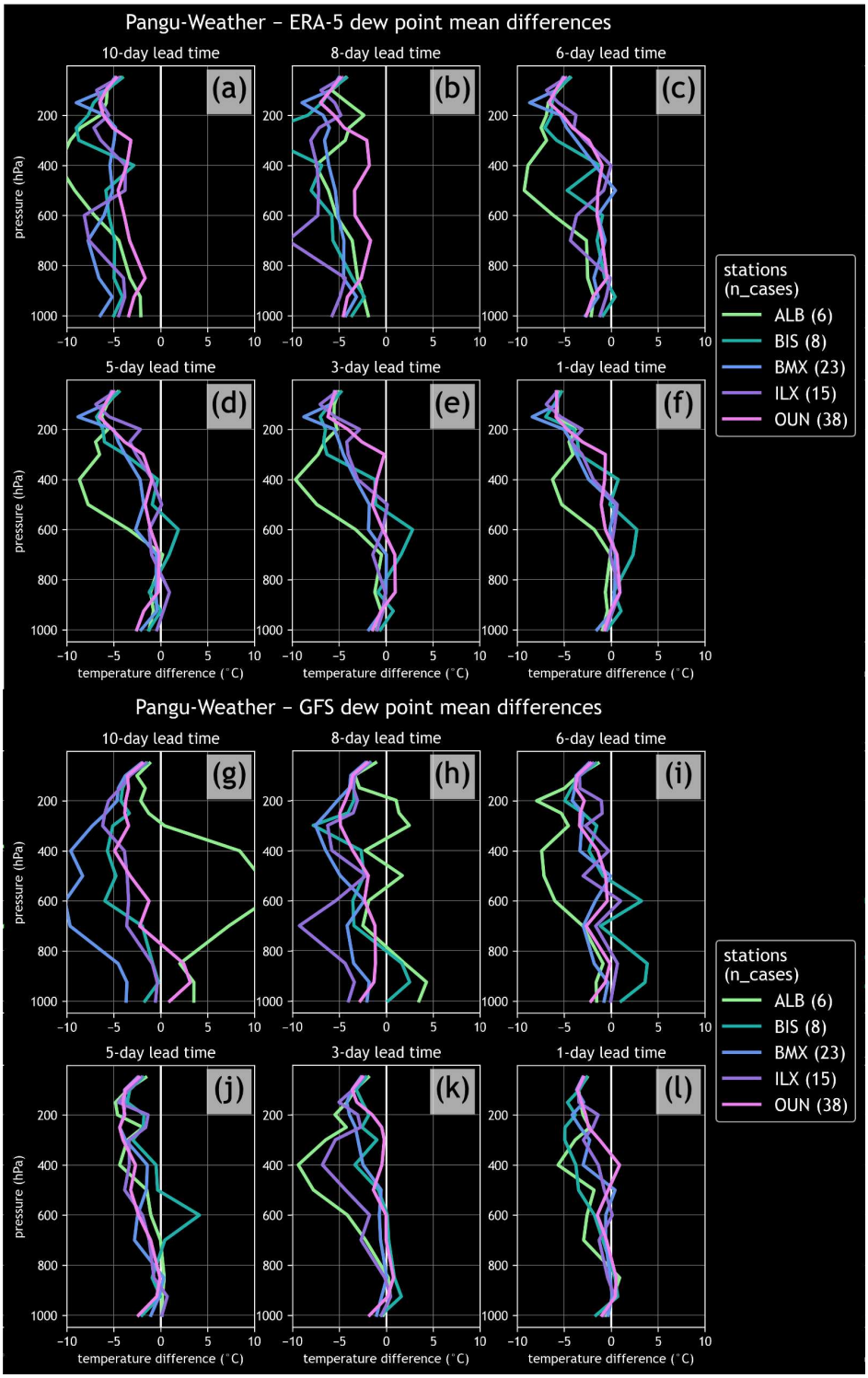

Figure 4.10: Mean differences in vertical profiles of dew point between (a)-(f) Pangu-Weather forecasts and ERA-5 and (g)-(l) Pangu-Weather forecasts and GFS forecasts. Differences are shown for forecasts valid at 0000 UTC on “convectively-favorable" forecast days (see methods for what constitutes such days). Mean differences are show for forecasts at 10-, 8-, 6-, 5-, 3- and 1-day lead times. Each line is colored according to the station they represent, which is listed in the legend along with the number of cases that contributed to the mean.

may exist. Beginning with Pangu-Weather, it is clear that across all 5 sounding sites that the DLWP forecasts tend to be much drier compared to ERA-5 reanalysis at longer lead times (e.g., Fig. 4.10a,b). As lead time decreases, the overall mean difference does get smaller, especially in the lower-levels (e.g., Fig. 4.10e,f). Dew point differences aloft remain several degrees Celsius lower in Pangu-Weather compared to ERA-5 at pressure levels above approximately 400 hPa. Given the already limited moisture at these levels of the atmosphere, this tendency is not overly concerning (at least in the context of large-scale storm environments–these differences would perhaps be of greater concern in finer-scale forecasts). Among the stations themselves, the mean dew point difference at OUN seems to be more closely aligned with ERA-5 (on average) compared to the other stations, though this could be related to the large number of samples included from that site.

When compared to the GFS, Pangu-Weather forecasts have less moisture throughout the vertical profiles in some, but not all cases. For example, at 10- and 8-day lead times, Pangu-Weather forecasts are generally drier compared to the GFS throughout the mid- and upper-levels at most stations, but the average differences in dew points near the surface tend to vary from station to station (Fig. 4.10g-h). This result implies that both Pangu-Weather and GFS are drier than ERA-5, but Pangu-Weather tends to be the driest overall. At shorter lead times, these differences become smaller, particularly in the lower atmosphere where the majority of moisture exists (e.g., Fig. 4.10k,l). It should be noted, however, that dew points near the surface remain slightly lower on average in Pangu-Weather than the GFS even at 1-day lead times (and this is also true in the ERA-5 data). While these differences are subtle, they could have notable implications in the derived convective parameter computations (e.g., SBCAPE and PWAT).

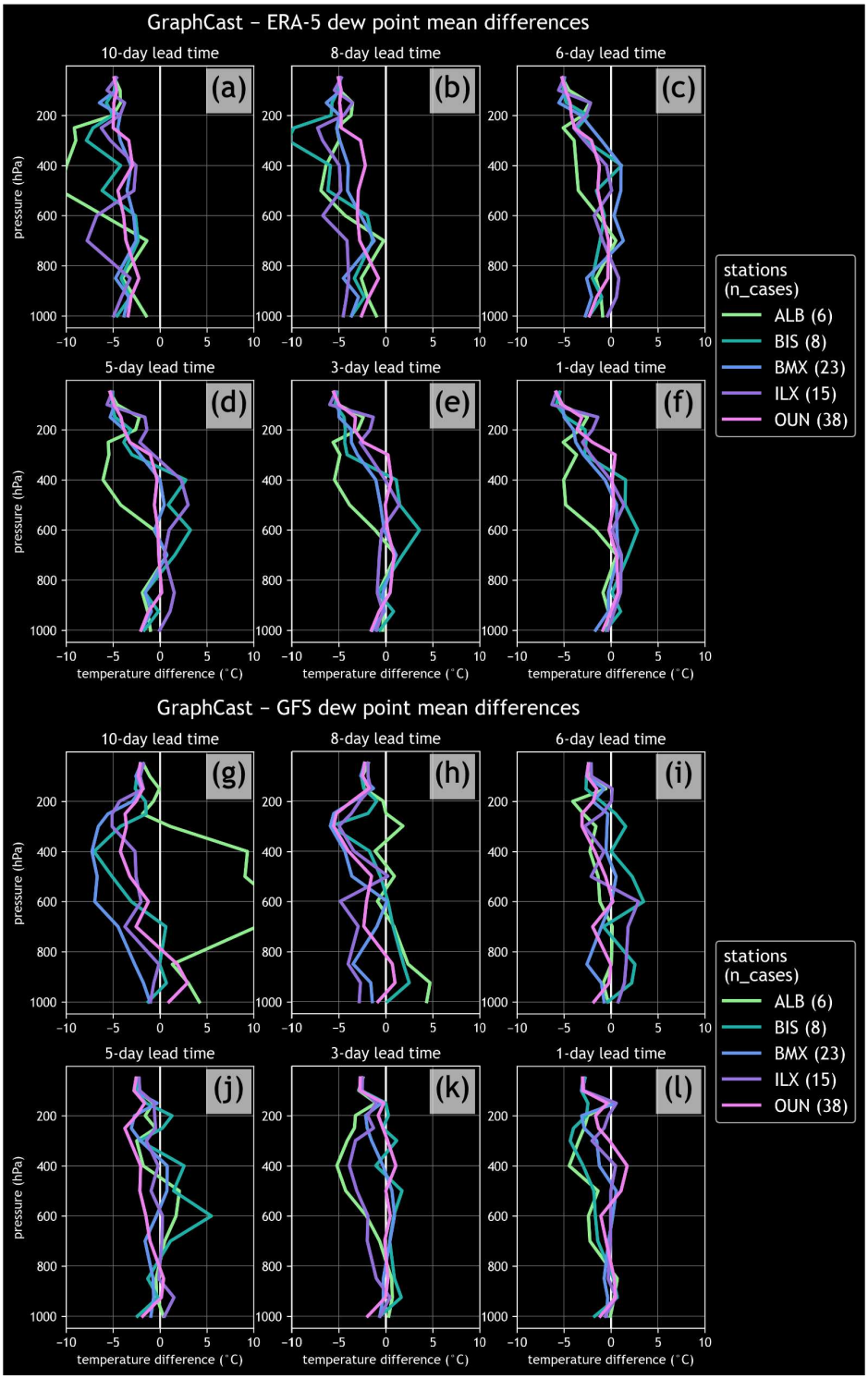

Figure 4.11: As in Fig. 4.10, but for GraphCast.

Like Pangu-Weather, GraphCast also shows a tendency for dew point temperatures to be lower than ERA-5 at longer lead times, and the differences become smaller at shorter lead times (Fig. 4.11a-f). In general, the mean dew point profile differences relative to ERA-5 are very similar between the Pangu-Weather and GraphCast forecasts over the various lead times. Thus, it is not too surprising that the profile differences between GraphCast and the GFS (Fig. 4.11g-l) also show many similarities to the dew point profile differences in Pangu-Weather versus the GFS.

The mean dew point difference profiles between FourCastNetv2 and ERA-5 also show that, like GraphCast and Pangu-Weather, FourCastNetv2 tends to have less moisture in the low- to mid-troposphere than the reanalysis (Fig. 4.12a-f). However, unlike Pangu-Weather and GraphCast, the FourCastNetv2 forecasts do not exhibit the drier dew points in the upper atmosphere (above approximately 400 hPa) like the other DLWP models. When compared to the GFS, the FourCastNetv2 profiles show *more* moisture in the upper levels, which was not seen in the GraphCast or Pangu-Weather comparisons with the GFS (e.g., Fig. 4.12k,l). Those differences remain even within a 1-day lead time. It is worth noting that FourCastNetv2 outputs relative humidity as a native variable, while Pangu-Weather and GraphCast output specific humidity. Perhaps because the models are trained to predict moisture slightly differently, there could be impacts on the dew points that are derived from these moisture variables as a result. However, Feldmann et al. (2024) suggest that moisture deficiencies in the models may be independent of thermodynamic moisture conversions, as their work showed that such errors exist in the native moisture variable outputs (before such conversions are made).

While the mean profiles can provide some useful insights on the overall behavior of the DLWP model dew points throughout the column, examining individual forecasts of dew point profiles can

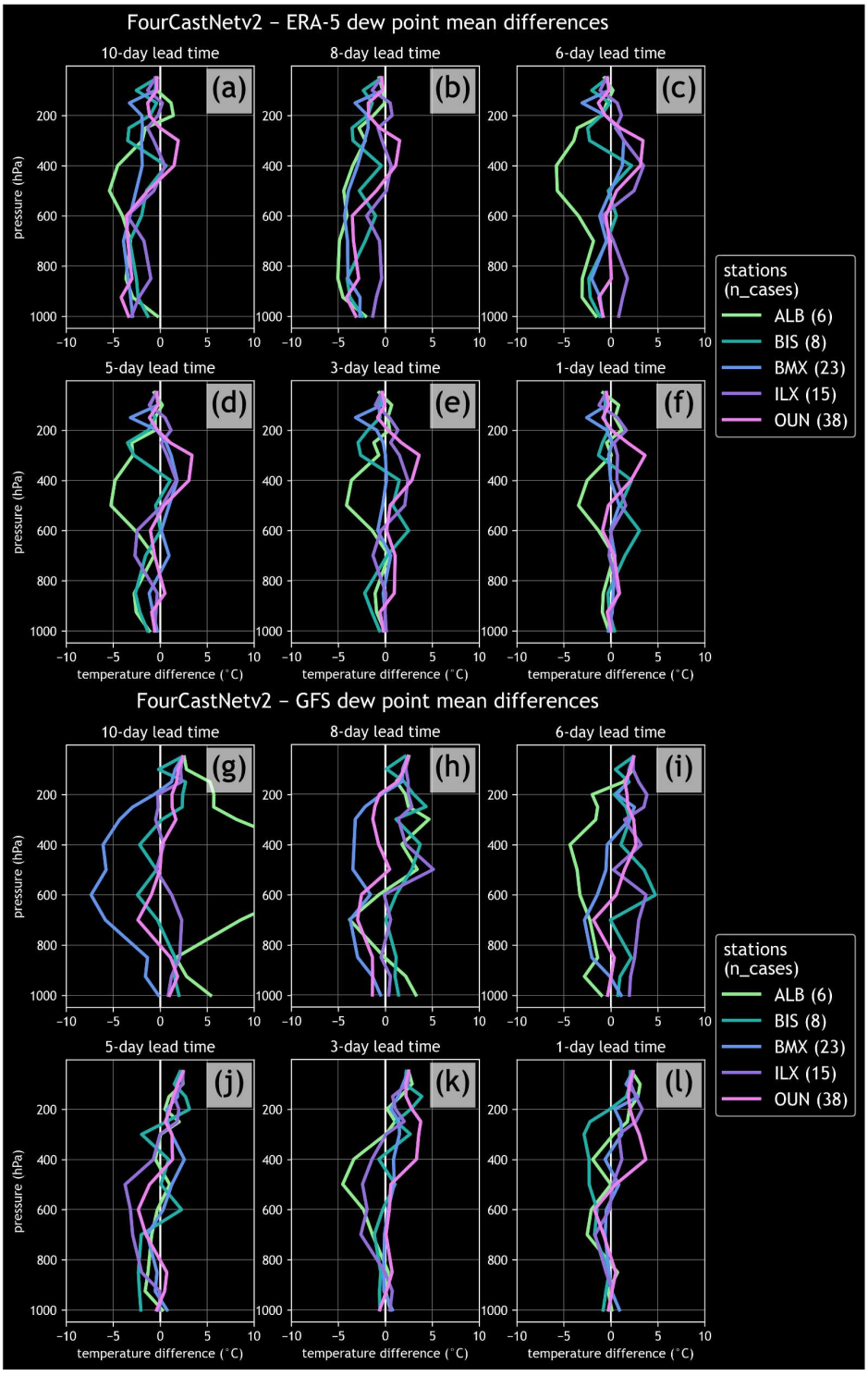

Figure 4.12: As in Fig. 4.10, but for FourCastNetv2.

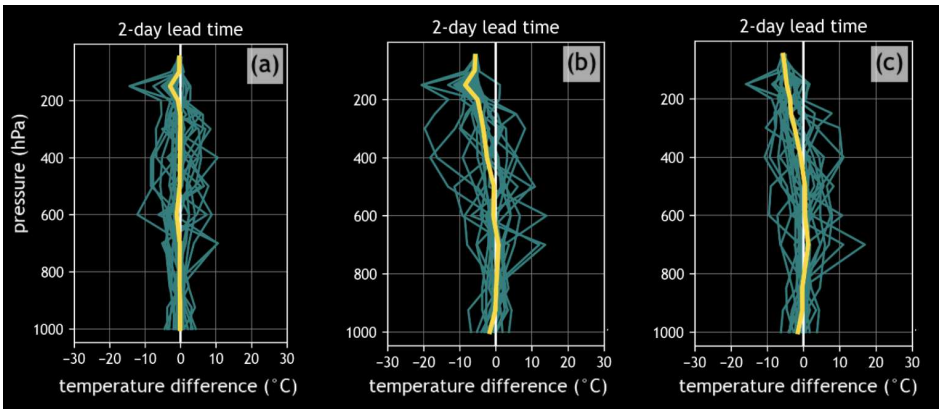

Figure 4.13: Differences in vertical profiles of dew point between day-2 (a) FourCastNetv2, (b) GraphCast, and (c) Pangu-Weather forecasts and ERA-5 for the 23 convectively favorable cases at Birmingham, AL. The teal lines represent the profile differences for each case, and the yellow line represents the mean difference.

provide additional details. Fig. 4.13 shows dew point differences between (a) FourCastNetv2, (b) GraphCast, and (c) Pangu-Weather and ERA-5 for the Birmingham, AL convective cases at 2-day lead times. FourCastNetv2 seems to exhibit the least amount of variability in its differences with ERA-5 compared to the other models (Fig. 4.13a), but all three forecasts display substantial case-to-case variability, even at this relatively short lead time. Thus, it is important to emphasize that the DLWP moisture tendencies will vary across individual forecasts.

Temperature profile differences between the DLWP models and ERA-5 show much less variability between them than the dew point profiles (Fig. 4.14). Around 50 to 200 hPa, model temperatures tend to be higher than ERA-5, especially in Pangu-Weather (e.g., Fig. 4.14a) and GraphCast (e.g., Fig. 4.14d). Below these levels, down to roughly 800 hPa, temperatures tend to be lower than ERA-5 in the models. These differences become smaller for shorter lead times (Fig. 4.14c,f,i), but differences on the order of several degrees Celsius between the models and reanalysis remain, particularly near the tropopause. Near-surface temperature differences compared to ERA-5 vary across the different models and sites, though by the 3-day lead time, most profiles

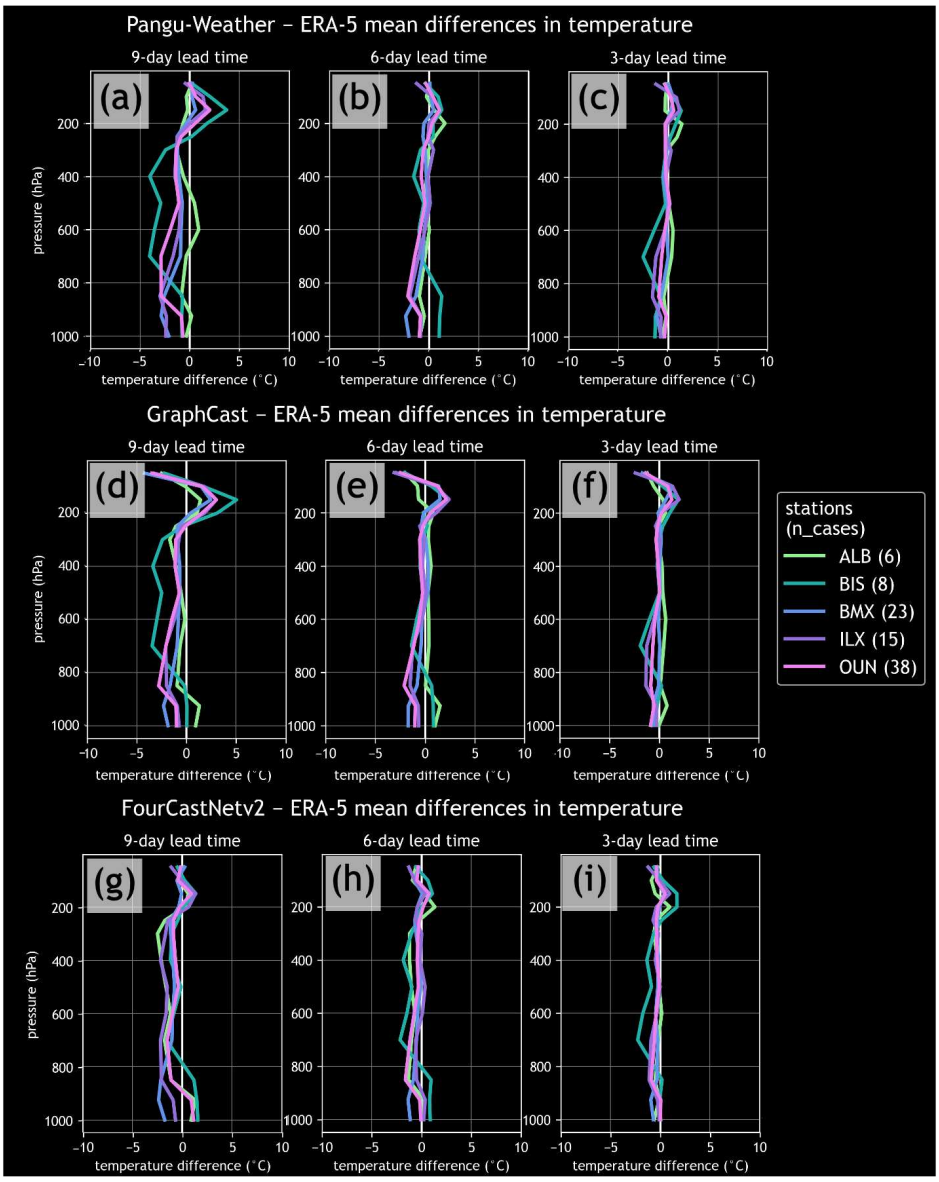

Figure 4.14: Mean differences in vertical profiles of temperature at 9-, 6-, and 3-day lead times between (a)-(c) Pangu-Weather, (d)-(f) GraphCast, and (g)-(i) FourCastNetv2 and ERA-5 reanalysis. Differences are shown for forecasts valid at 0000 UTC on "convectively-favorable" forecast days (see methods for what constitutes such days). Each line is colored according to the station they represent, which is listed in the legend along with the number of cases that contributed to the mean.

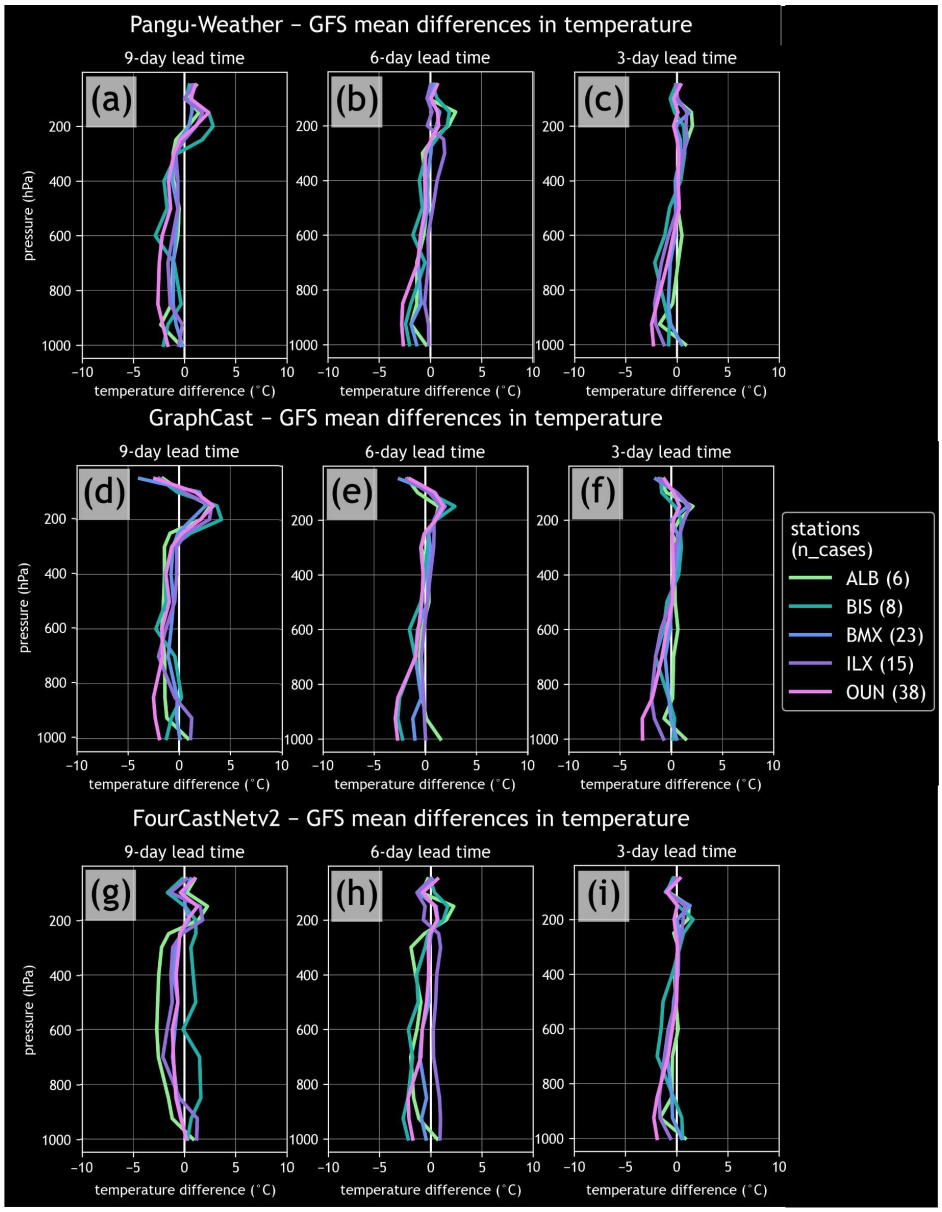

Figure 4.15: As in Fig. 4.14, but for differences between the DLWP models and GFS forecasts.

show a greater propensity for having lower near-surface temperatures than ERA-5. For some sites, discrepancies in mean temperature difference between the DLWP models and ERA-5 are at most a couple degrees Celsius. Broader evaluations of these DLWP systems have shown skillful performance for temperatures at the surface and aloft (Bi et al., 2023; Bonev et al., 2023; Lam et al., 2023), and while the temperature differences are not totally unreasonable in the cases presented here, these seemingly subtle differences can have subsequent implications (such as for computing derived parameters, as will be discussed in greater detail later).

Results are similar when comparing the DLWP modeled temperatures to the GFS predicted temperatures (Fig. 4.15). Temperatures above 200 hPa tend to be higher than the GFS and lower throughout much of the rest of the atmospheric column. Across most of the lead times and sites, temperatures near the surface also seem to be lower than the GFS, though there are some exceptions (e.g., ALB seems to be warmer near the surface compared to other models; e.g., Fig. 4.15e). Temperature differences throughout the atmospheric column will inherently impact derived parameter calculations, particularly for instability parameters such as SBCAPE.

4.3.3 Case studies

To examine the DLWP model output in a practical, operational manner, it is helpful to retrospectively examine forecasts for individual cases. The two cases presented here examine two severe weather events that were selected due to differences in their environmental characteristics, seasonal timing, intensity, and location. The first case, also referred to as case 1 or the strongly-forced case, took place in spring 2023 and affected the Midwest and Deep South. The second case (case 2 or the weakly forced case), occurred in summer 2023 and impacted the Ohio Valley and

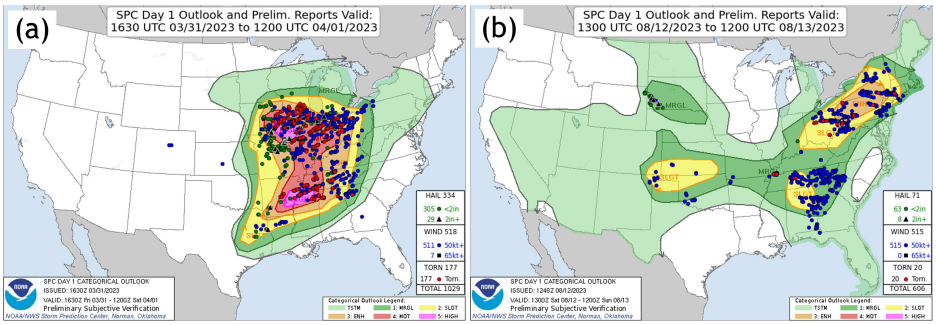

Figure 4.16: Day 1 SPC convective outlooks issued (a) 1630 UTC on 31 March 2023 and (b) 1300 UTC on 12 August 2023. Both outlooks are overlaid with tornado (red), hail (green) and blue (wind) reports. Significant severe wind and hail reports are labeled with squares and triangles respectively. The cases in panels (a) and (b) will be referred to as case 1 (or the "strongly-forced" case) and 2 (or the "weakly-forced" case) respectively.

Northeast. A brief meteorological summary is presented for each case prior to discussing their respective forecasts[18].

Case 1: 31 March - 1 April 2023

Case 1 was a rare SPC-defined high risk day, with reports ultimately stretching from the Great Lakes region to eastern Texas (Fig. 4.16a). In the upper levels, a deep longwave trough with a strong jet streak was centered near eastern Oklahoma and southern Nebraska and promoted divergence aloft (not shown). In the mid- and lower-levels, the trough was displaced further eastward (e.g., Fig. 4.17), which enhanced deep-layer vertical shear and created southerly flow that helped advect moisture from the Gulf of Mexico. At the surface, a low-pressure system centered over Nebraska quickly intensified as it moved into Iowa by the afternoon, which helped generate additional shear in the lower levels (not shown). This classic severe weather regime supported the development of both supercellular and linear convection throughout the outbreak area, ultimately leading to several hundred tornado, severe wind, and hail reports.

[18]The SPC Mesoscale Analysis Archive was used to summarize the cases (https://www.spc.noaa.gov/exper/ma_archive/).

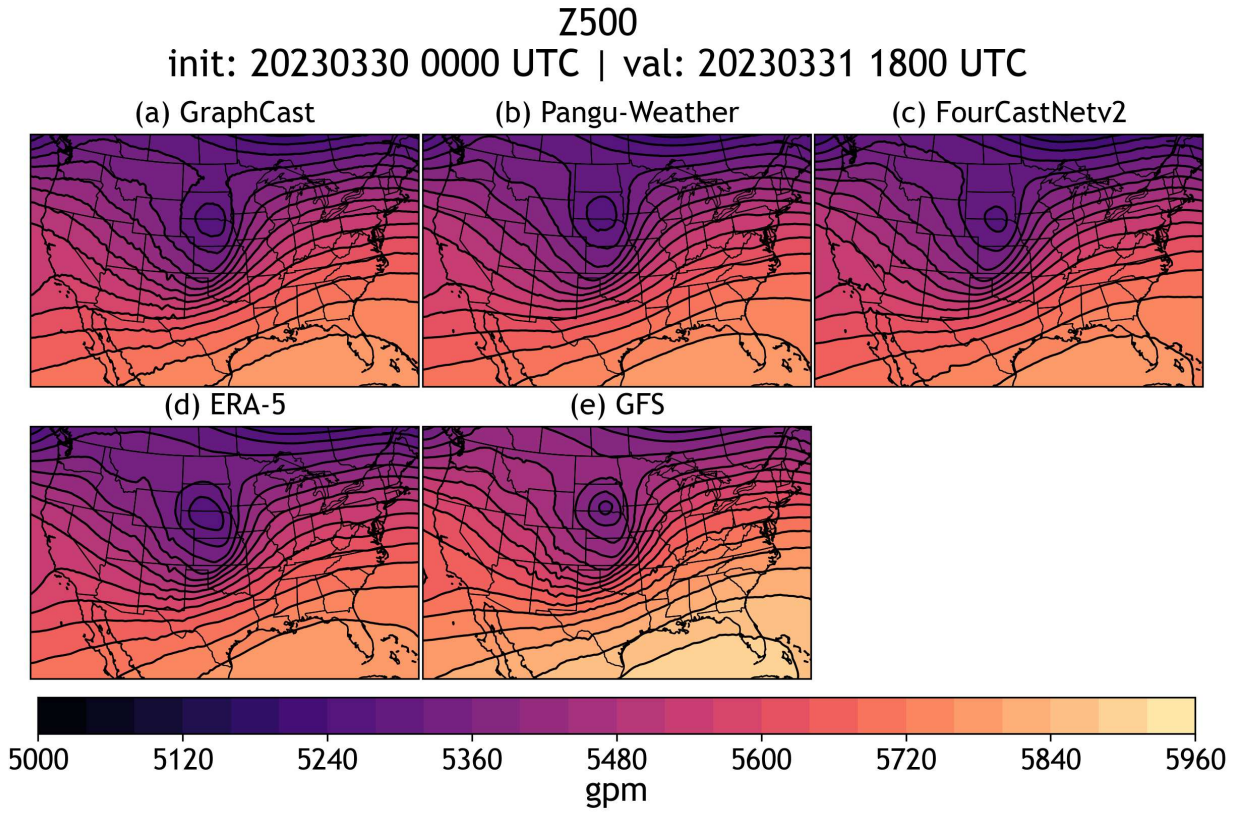

Figure 4.17: 42-h 500 hPa geopotential height forecasts valid for 1800 UTC on 31 March 2023 from (a) GraphCast, (b) Pangu-Weather, (c) FourCastNetv2, and (e) GFS, as well as (d) ERA-5 reanalysis valid at that time.

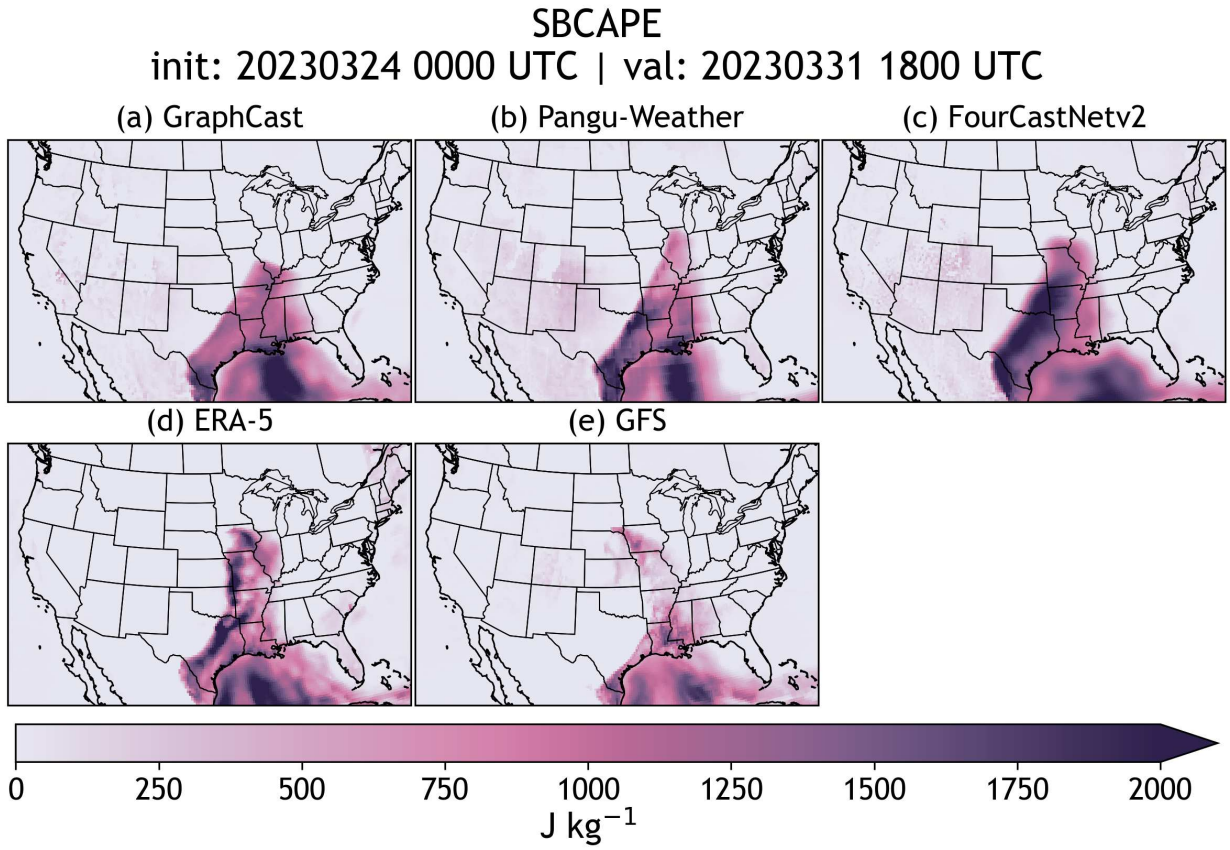

Figure 4.18: 186-h surface-based CAPE forecasts valid for 1800 UTC on 31 March 2023 from (a) GraphCast, (b) Pangu-Weather, (c) FourCastNetv2, and (e) GFS, as well as (d) ERA-5 reanalysis valid at that time.

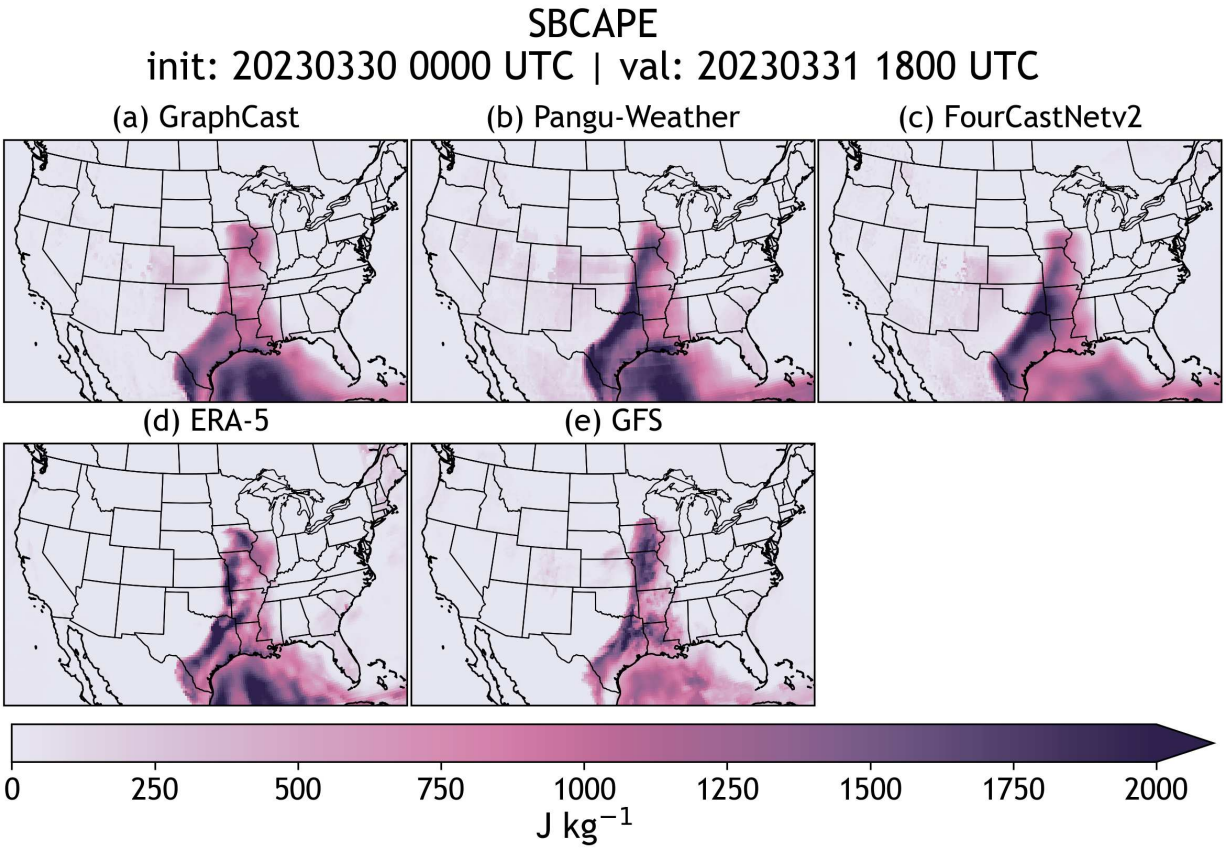

Figure 4.19: As in Fig. 4.18, but for the 42-h forecasts.

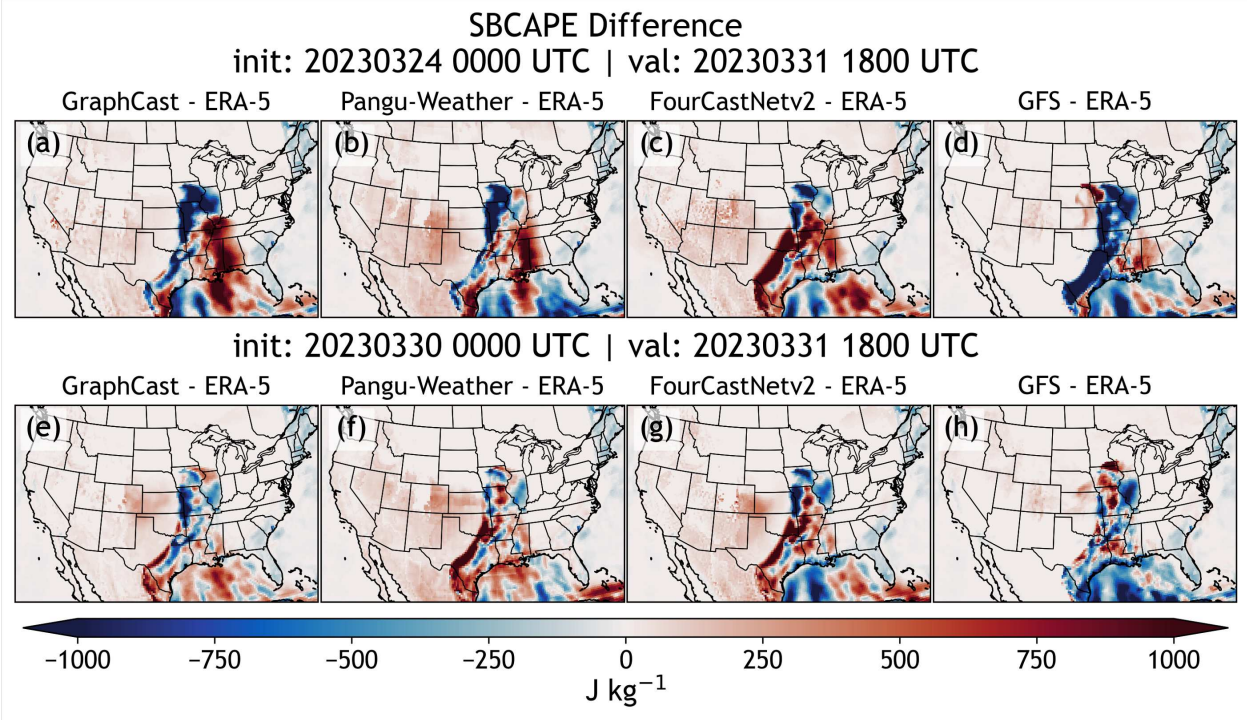

Figure 4.20: Differences between forecasted SBCAPE in (a), (e) GraphCast, (b), (f) Pangu-Weather, (c), (g) FourCastNetv2, and (d), (h) GFS and ERA-5 reanalysis. The top panels shows differences between the DLWP models and ERA-5 for forecasts initialized 24 March 2023 at 0000 UTC (as in Fig. 4.18), and the bottom panels show these differences for forecasts initialized 30 March 2023 at 0000 UTC (as in Fig. 4.19).

Evaluating the evolution of SBCAPE at different forecast times helps illustrate how well the models predicted the placing, timing, and spatial extent of the instability during the strongly-forced case. Comparing forecasts from the DLWP models and GFS to ERA-5 at lead times of approximately 8 (Fig. 4.18) and 2 days (Fig. 4.19), all three DLWP models illustrate a plausible SBCAPE footprint at the longer lead time that is fine-tuned by the day-2 forecasts. In the 184-h forecasts, GraphCast, Pangu-Weather, and FourCastNetv2 show a ribbon of moderate to high SBCAPE stretching from the Gulf of Mexico into the Midwest, which generally agrees with the ERA-5 reanalysis (Fig. 4.18d). There are some subtle differences among the three models at this lead time, such as that GraphCast limits the northern extent of the strongest instability compared to the other models (Fig. 4.18a), and FourCastNetv2 has larger amounts of SBCAPE (Fig. 4.18c). Compared to the GFS, the three DLWP seem to have a better handle on the magnitude of instability present, as the GFS only shows modest amounts in comparison (Fig. 4.18e). By the 42-h forecasts, however, the models are visually in much better agreement with ERA-5, though it should be noted that the GFS offers much more granularity than the DLWP models do (Fig. 4.19).

Another way to evaluate the SBCAPE from the DLWP models for these lead times is to examine their differences with ERA-5 (Fig. 4.20). For the 186-h forecasts, all three DLWP models show a tendency to limit the northward extent of the instability relative to ERA-5, and they also show too much instability to the east (Fig. 4.20a,b,c). Some of these issues may arise in subtle forecast timing differences, however. For instance, GraphCast and Pangu-Weather show smaller amounts of instability on the western side of the region of increased instability relative to ERA-5 (Fig. 4.20a,b), while FourCastNetv2 shows greater instability in this area (Fig. 4.20c). These

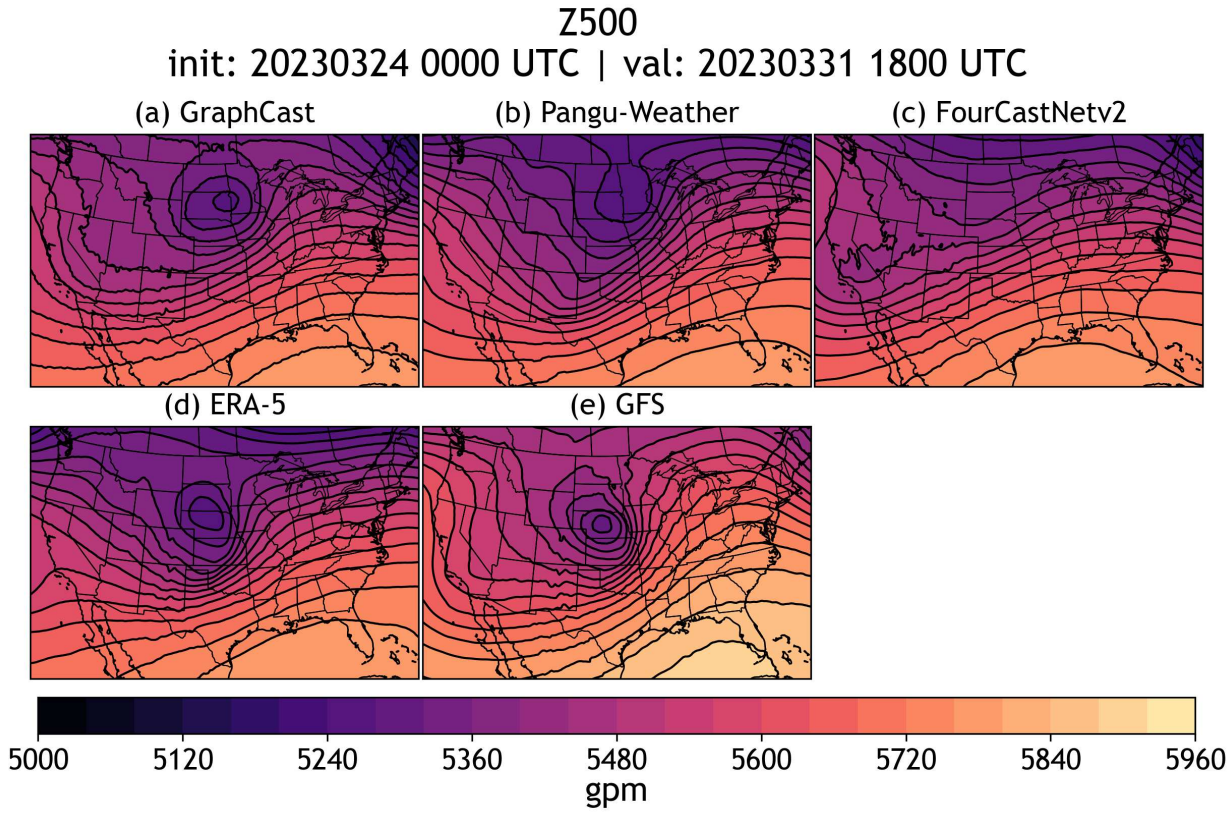

Figure 4.21: As in Fig. 4.17, but for forecasts initialized 24 March 2023.

differences suggest that for this particular forecast, GraphCast and Pangu-Weather move the synoptic system too quickly (i.e., locations to the west stabilize too quickly) and FourCastNetv2 may be too slow (i.e., locations to the west remain unstable for too long). The 500 hPa height forecasts at this lead time seem to confirm this theory: GraphCast and Pangu-Weather eject the trough more quickly than ERA-5 (Fig. 4.21a,b,d), and FourCastNetv2 has the trough lagging slightly behind ERA-5 (Fig. 4.21c,d). For the 42-h forecasts, despite the better placement of the greatest instability (Fig. 4.19a,b,c,) and stronger agreement in the placement of the trough (Fig. 4.17a,b,c), large differences in the magnitude of SBCAPE between ERA-5 and the DLWP models remain (Fig. 4.20d,e,f). Results are noisy in these plots, but in general it seems that GraphCast has a greater tendency to have too little SBCAPE relative to ERA-5, FourCastNetv2 tends to have too much (but also does not bring the instability far enough north), and Pangu-Weather shows mixed results.

To examine forecasts of the convective environment further, skew-T log-P diagrams can be used to visualize temperature and dew point forecasts compared to reanalysis and observations (Figs. 4.22; 4.23). Beginning with the forecasted soundings valid for 1800 UTC for 31 March at the grid point nearest Lincoln IL, it is clear that all the models (including the GFS) struggle to capture moisture relative to ERA-5 and observations, illustrated by their uncertainty across different forecast lead times (Fig. 4.22). The DLWP models show more smoothness over time in their dew point fields compared to the GFS (Fig. 4.22d), due in part to the fact that the GFS has more vertical levels in comparison. Still, while the DLWP dew point forecasts show uncertainty, they represent the environment reasonably well compared to ERA-5 and observations (even getting closer to the observed dew point profile in the mid-upper levels than ERA-5 suggests). This variability in the

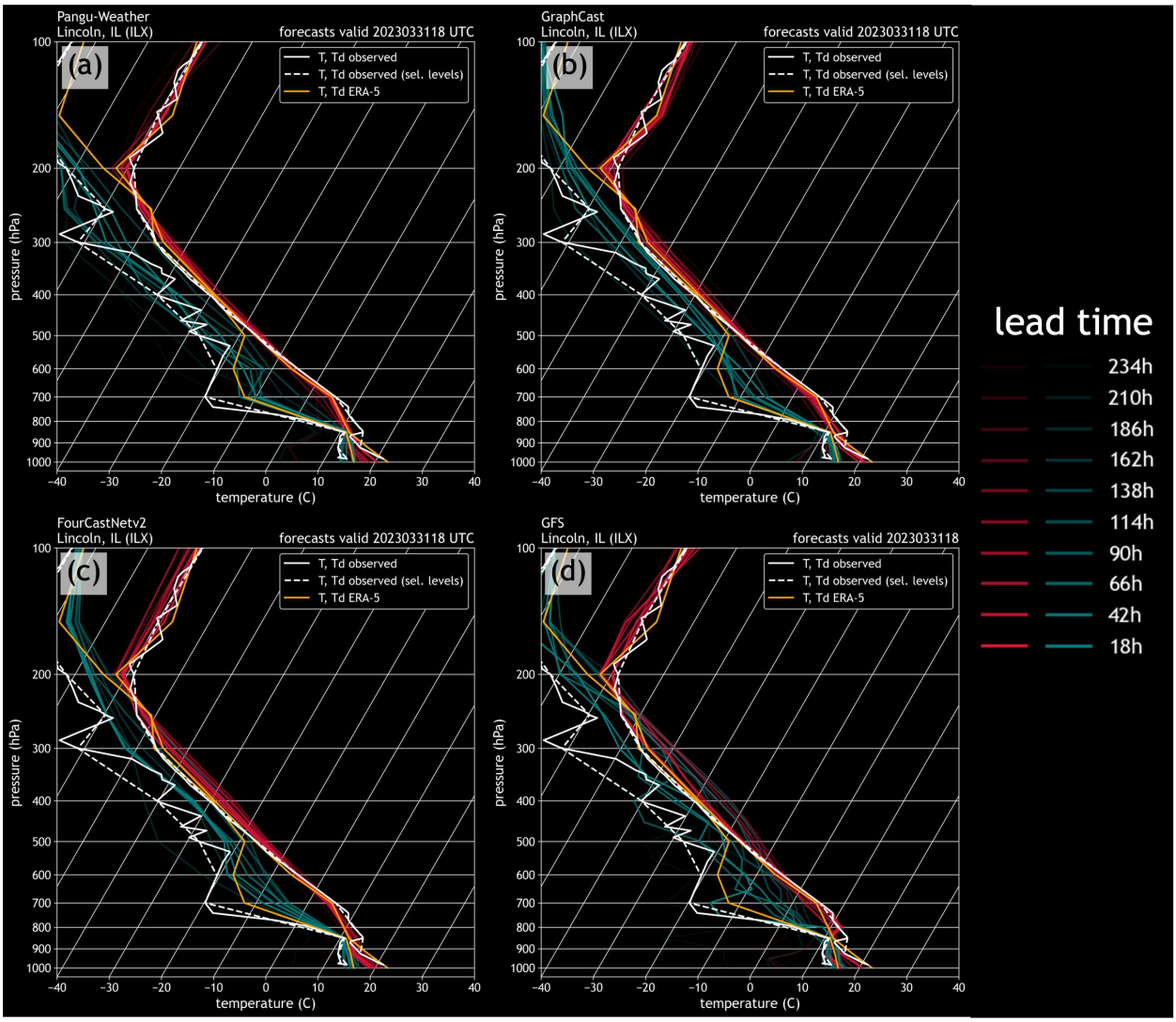

Figure 4.22: Time-evolution of forecast soundings from (a) Pangu-Weather, (b) GraphCast, (c) FourCastNetv2, and (d) the GFS valid at 1800 UTC 31 March 2023 at the model grid point nearest Lincoln, IL. Red and teal lines show the temperature and dew point profiles (respectively) for forecasts initialized every 24 hours beginning at a 234-h lead time (approximately 10 days) to an 18-h lead time; lines darken with decreasing lead time. The model forecasts are overlaid with the ERA-5 reanalysis using only 13 pressure levels (orange), as well as the observed ILX sounding with all levels (white solid) only 13 vertical pressure levels (white dashed).

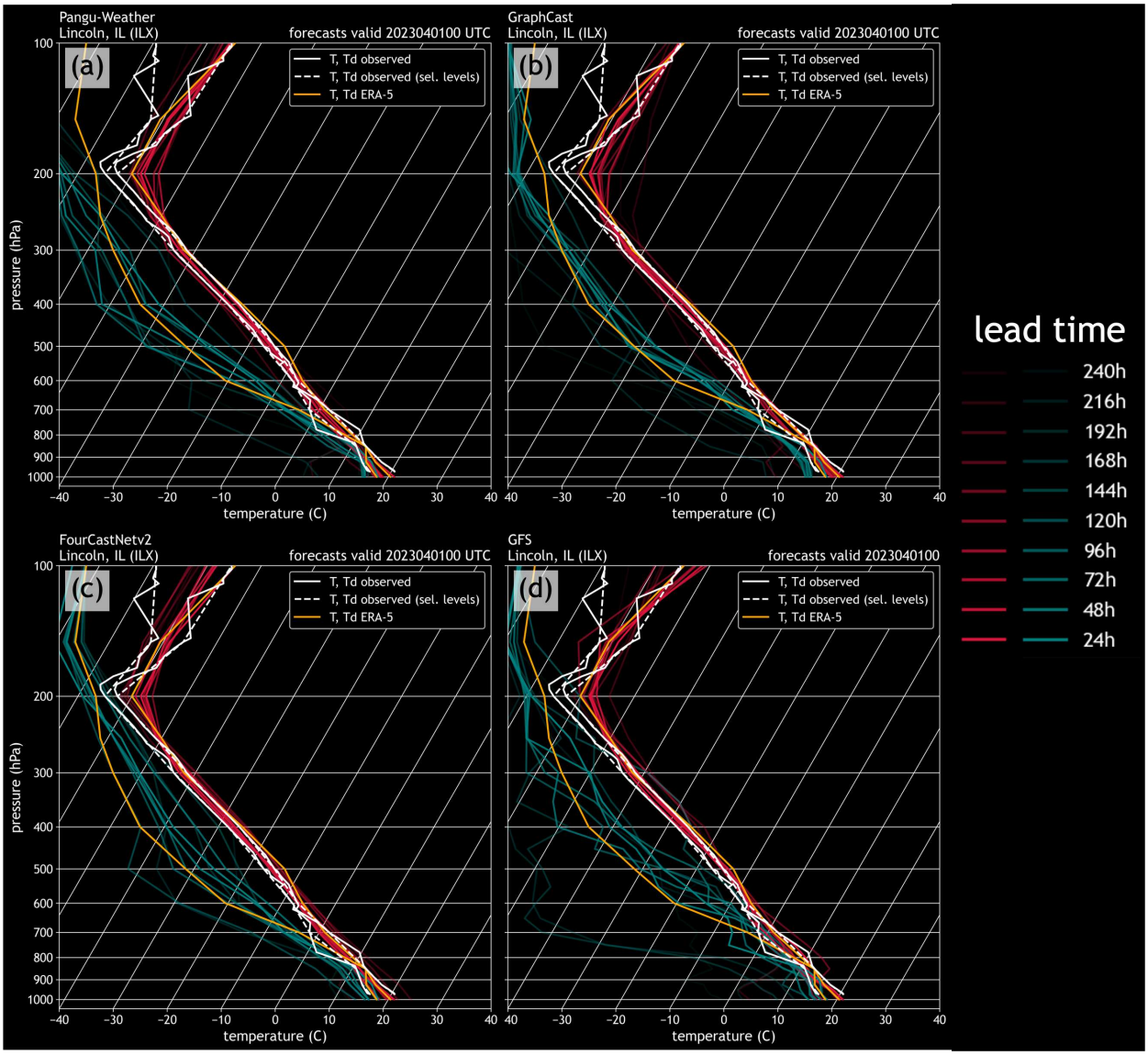

Figure 4.23: As in Fig. 4.22, but forecasts, reanalysis, and soundings are valid for 1 April 2023 at 0000 UTC, and forecasts are initialized every 24 hours beginning at a 240-h lead time up to a 24-h lead time.

dew point forecasts over time is also evident in forecasts 6 hours later (Fig. 4.23), though GraphCast seems to match the ERA-5 reanalysis profile most closely. Note that the dew point profile in the observed sounding is much different from both the forecasts and the reanalysis. It shows a well-saturated atmosphere, which resulted from of a thunderstorm moving through Lincoln just prior to the 0000 UTC launch (not shown). While the observed profile is different from the forecast data, it is to be expected that present global resolution models, both ML-based and NWP, will inherently miss these small details in forecasts. Further, it raises questions on how the vertical profiles from the ML-based forecasts might appear in and around observed precipitation given that (with the exception of GraphCast) they do not currently predict precipitation.

With respect to temperature, the DLWP models all demonstrate accuracy in their predictions, even at advanced lead times (Figs. 4.22a,b,c; 4.23a,b,c). However, there is one major caveat here: note that in the observed sounding at 1800 UTC, there is a low-level inversion (Fig. 4.22). None of the DLWP models model this stable layer, but the GFS (while not perfect) is able to. It is suspected that the primary reason for this deficiency in the DLWP models is due to the fact that they have too few vertical levels to be able to properly capture such details. The ERA-5 reanalysis, which only includes 13 vertical levels here, also does not show a stable layer and thus supports this theory, however additional experimentation with the GFS vertical levels would be needed to determine whether the DLWP have compounding issues or if missing stable layers is solely a matter of limited vertical resolution.

Skew-T diagrams can also be examined for Birmingham, AL, which describes the forecasted and observed environment in the southern part of the risk area (Fig. 4.24). Forecasted dew points in the DLWP models show similar degrees of uncertainty as the forecasts for Lincoln, IL, and much

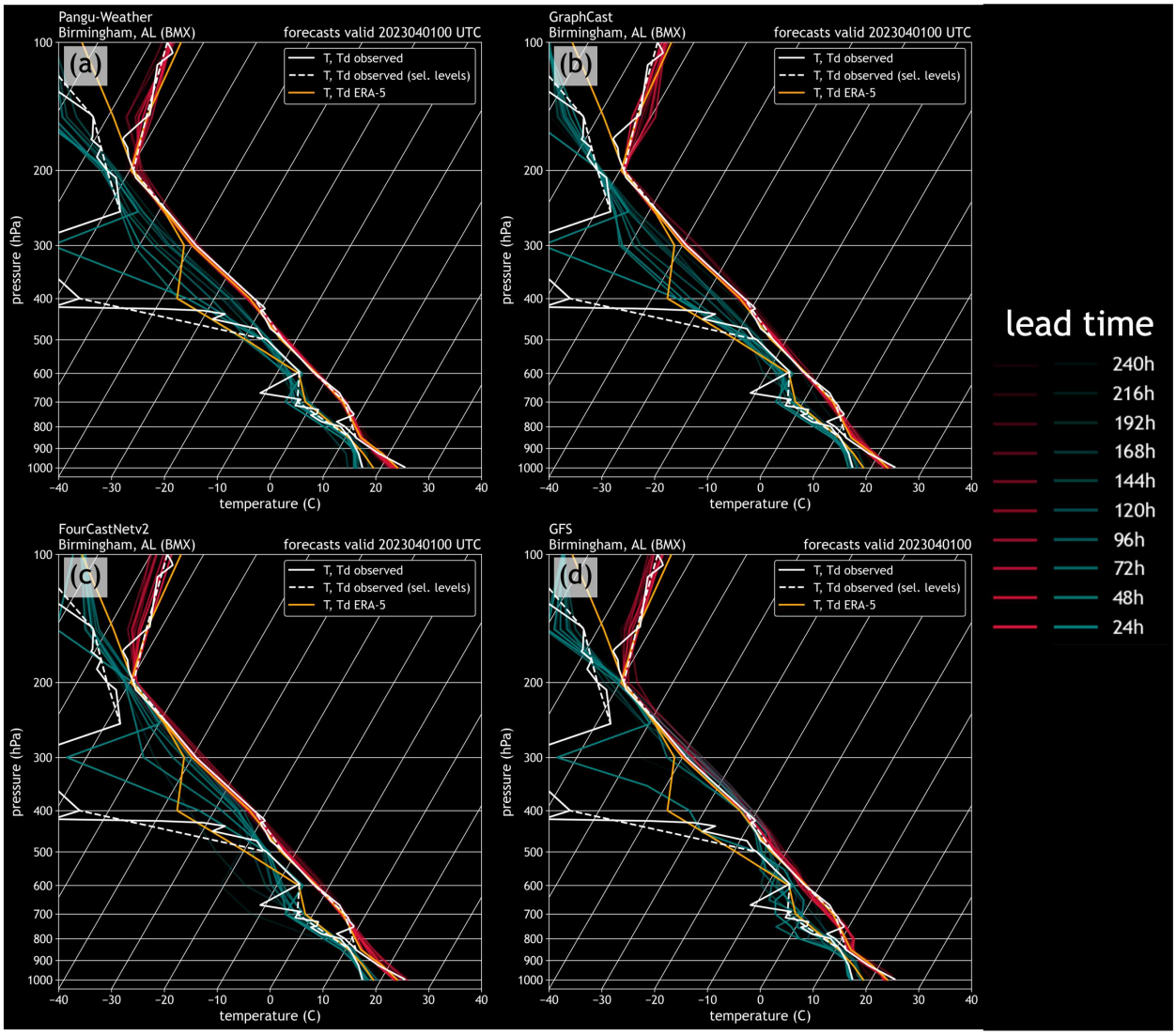

Figure 4.24: As in Fig. 4.23, but forecasts and reanalysis are valid at the grid point nearest Birmingham, AL. The 0000 UTC observed sounding from BMX is overlaid.

of this uncertainty occurs in the mid- and upper-levels. Forecasted dew points in the lower levels align well with the observed sounding in the three DLWP models. Interestingly, the GFS shows a saturated sounding in the majority of its forecasted profiles (Fig. 4.24d), which disagrees with the observed sounding as well as the DLWP forecasted moisture profiles. Temperature forecasts are in good agreement across the forecasts from all models, however the DLWP models fail to capture an observed stable layer in this example as well (Fig. 4.24a,b,c). To be fair, the GFS also largely does not predict the temperature inversion, but it does show more stability near 850 hPa (where the inversion occurred) compared to the DLWP models.

Comparing forecasted versus observed hodographs can help illustrate how well the models are capturing wind shear during the event compared to observations and reanalysis (Figs. 4.25; 4.26). Among the DLWP forecasts, the predicted shear profile does not change substantially in terms of magnitude or direction over the 10 days of forecasts, evidenced by both the 1800 UTC and 0000 UTC valid times (Figs. 4.25a,b,c; 4.26a,b,c). Compared to ERA-5, the shear magnitude and direction is well-predicted in both hodographs. However, compared to the observations, it appears that the DLWP models significantly underdo the magnitude of the wind shear, as well as miss some directional characteristics. For example, at 0000 UTC, the wind vectors in the DLWP models at 925 hPa (i.e., the second vertical level from the surface) are almost southwesterly (Fig. 4.26), while it is southerly in the observed sounding (and nearly southerly in many of the GFS runs). Further, the wind speed is nearly more than double in the observations at this level compared to the wind speeds in the DLWP models. These differences would undoubtedly have significant implications for derived shear-related parameters. The GFS seems able to capture the magnitude (at least in the mid-levels) slightly better than the DLWP models (e.g., Fig. 4.26d), but it too underdoes the shear

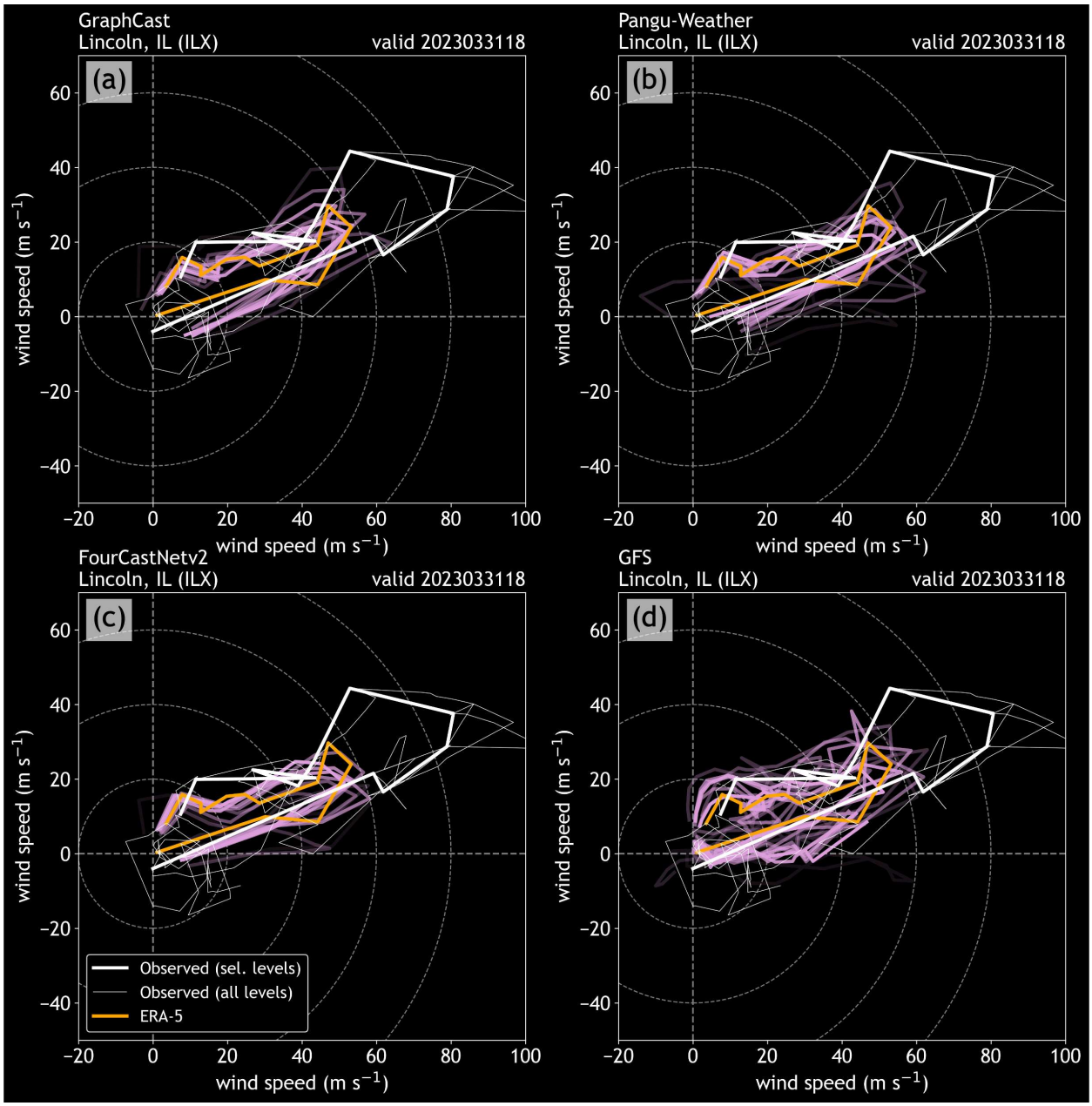

Figure 4.25: Time-evolution of forecast hodographs from (a) Pangu-Weather, (b) GraphCast, (c) FourCastNetv2, and (d) the GFS valid at 1800 UTC 31 March 2023 at the model grid point nearest Lincoln, IL. Pink lines show the wind hodographs for forecasts initialized every 24 hours beginning at a 234-h lead time (approximately 10 days) to an 18-h lead time; lines darken with decreasing lead time. The model forecasts are overlaid with the ERA-5 reanalysis using only 13 pressure levels (orange), as well as the observed ILX hodograph with all levels (thin white line) only 13 vertical pressure levels (bold white line).

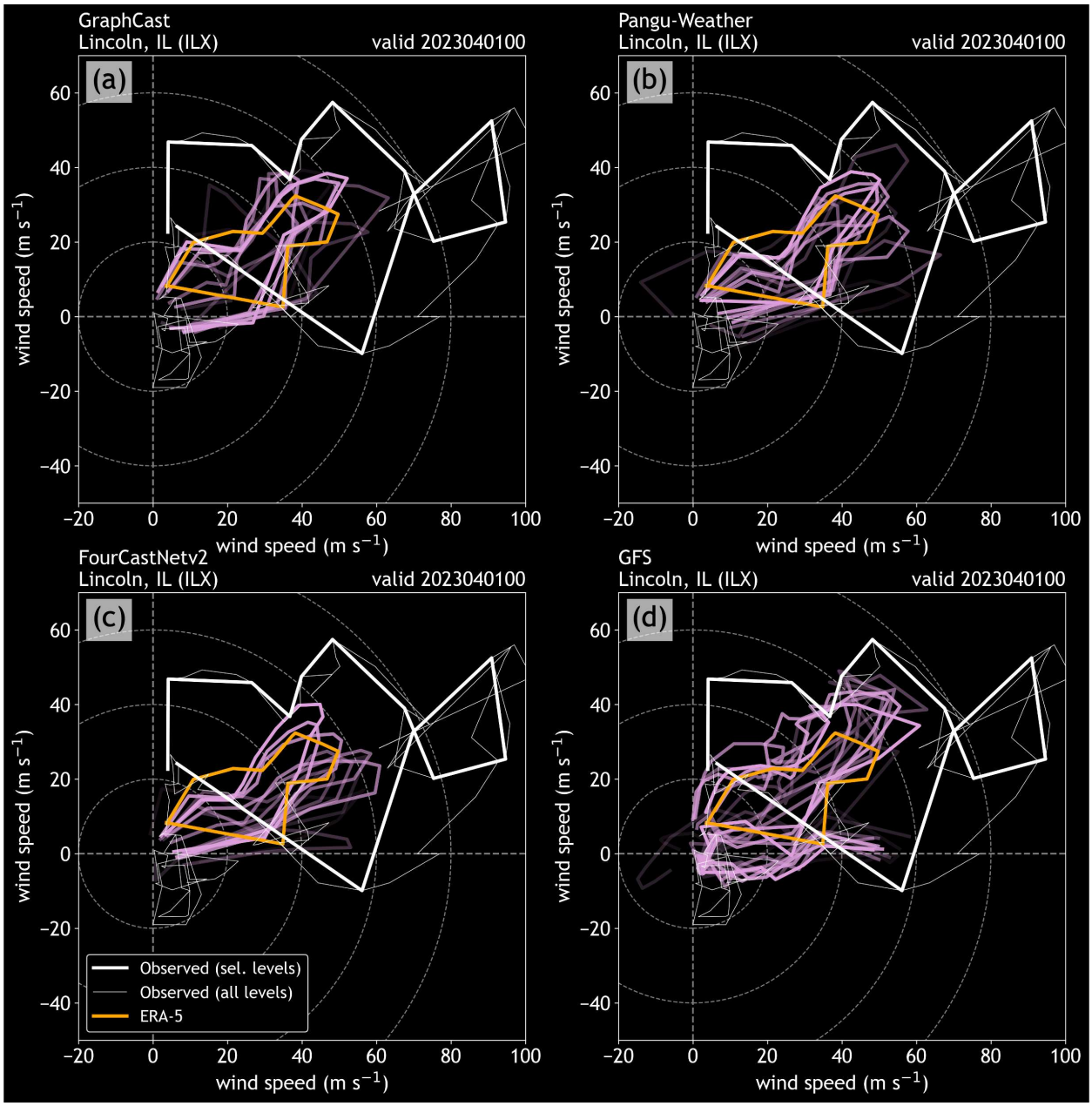

Figure 4.26: As in Fig. 4.25, but forecasts, reanalysis, and soundings are valid for 1 April 2023 at 0000 UTC, and forecasts are initialized every 24 hours beginning at a 240-h lead time up to a 24-h lead time

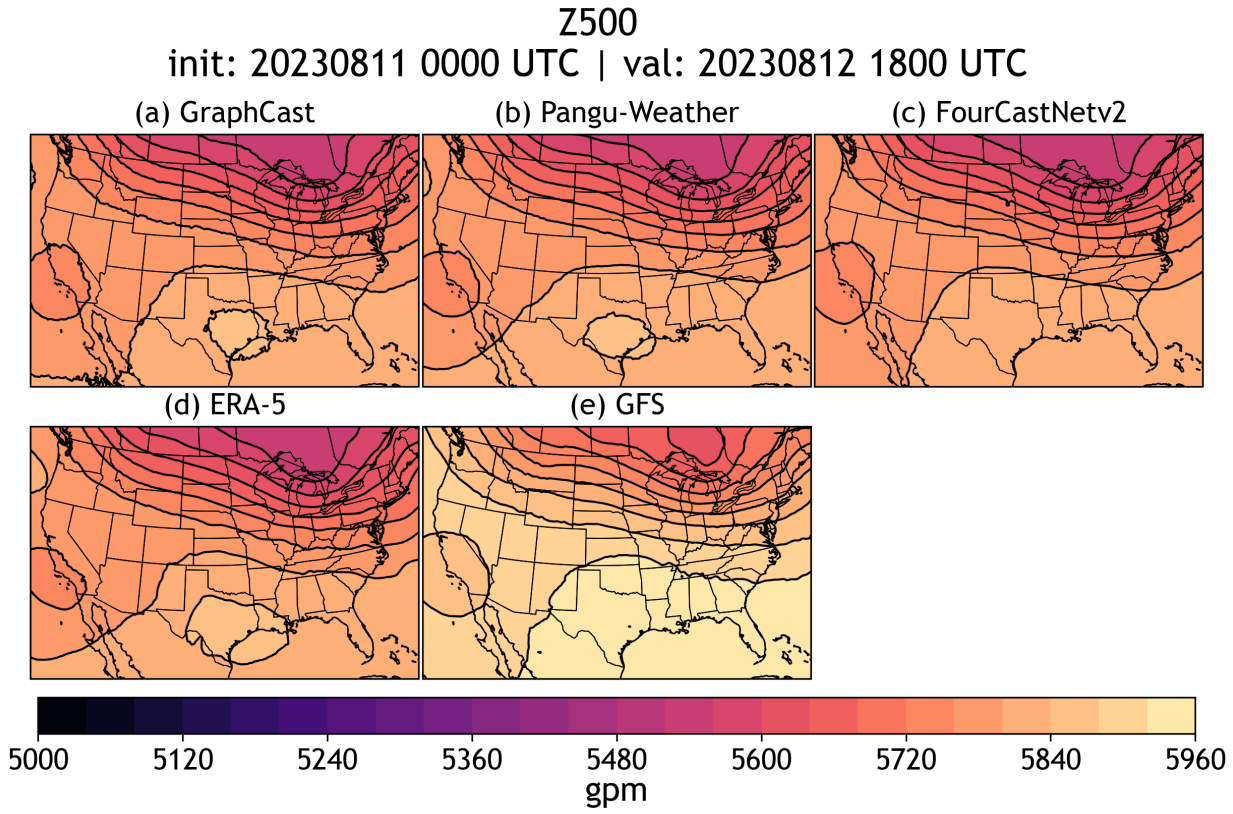

Figure 4.27: 42-h 500 hPa geopotential height forecasts valid for 1800 UTC on 12 August 2023 from (a) GraphCast, (b) Pangu-Weather, (c) FourCastNetv2, and (e) GFS, as well as (d) ERA-5 reanalysis valid at that time.

magnitude substantially. These results are consistent in the Birmingham, AL soundings for this case as well (not shown).

Case 2: 12-13 August 2023

The second case, which took place 12 to 13 August 2023, was a lower-end severe weather outbreak relative to case 1. On this day, there were actually four general areas where severe weather reports ultimately occurred, including western South Dakota, southwestern Kansas into the Texas Panhandle, the Tennessee Valley and Southern Appalachians, and the Ohio Valley into the Northeast (Fig. 4.16b). This case study will focus on this latter area of interest, where the SPC day-1

convective outlook highlighted an enhanced risk of severe storms that day. Surface maps preceding the event showed an occluding surface low located near Lake Ontario on the morning of 12 August (not shown). A warm front associated with the system gradually pushed northeastward across New York throughout the day, while the system's cold front moved across the Ohio Valley (not shown). Aloft, a shortwave trough became increasingly negatively tilted early in the day over the Great Lakes region (Fig. 4.27), which enhanced southwesterly flow and ushered in moisture and helped destabilize the atmosphere. New convection began to initiate midday across Ohio, growing into a relatively disorganized multicellular cluster that produced severe wind and tornado reports soon after (not shown). Meanwhile, downstream of this small thunderstorm complex, isolated storms developed ahead within the warm sector and produced damaging winds and a few tornadoes across New York, Pennsylvania, and other nearby areas (not shown). Beginning with SBCAPE forecasts made approximately 8 days from the start of the event (Fig. 4.28), there is much more variability in the SBCAPE forecasts among the three DLWP models compared to the 8-day forecasts for case 1. Relative to ERA-5, GraphCast shows too little SBCAPE over the Midwest and Ohio Valley (Fig. 4.28a). Pangu-Weather shows too much SBCAPE compared to ERA-5 in this area, while results are variable for FourCastNetv2 (Fig. 4.28b,c). The GFS shows little SBCAPE in the region of interest (Fig. 4.28). This lack of agreement among these forecasts suggest that SBCAPE predictability by the DLWP may be even more difficult in weaker-forced scenarios like this case, as opposed to stronger-forced cases like case 1, though additional case studies are needed to confirm this theory. At the day-2 lead time, consensus is stronger among the DLWP models (Fig. 4.29). Pangu-Weather has too much SBCAPE over Ohio and western Pennsylvania and New York (Fig. 4.29b), but the other two DLWP models are more closely aligned with

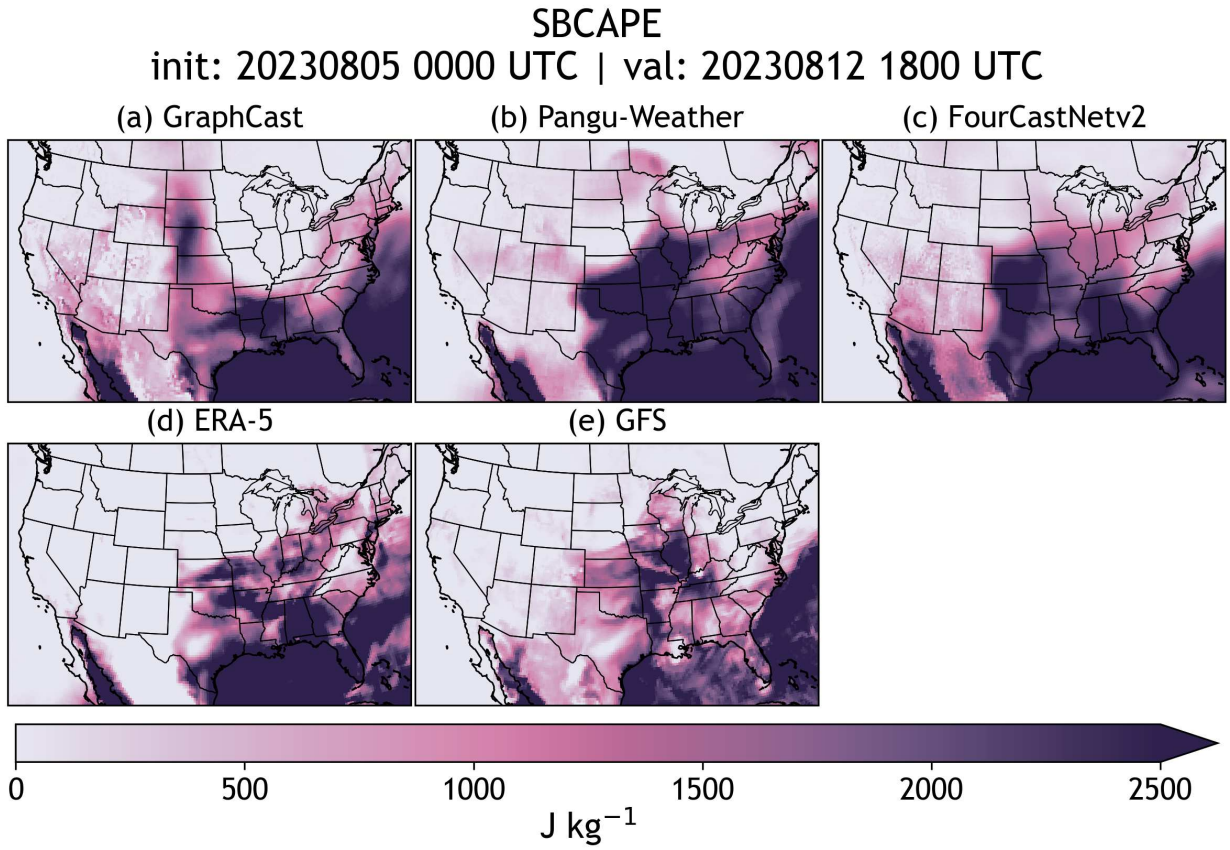

Figure 4.28: 186-h surface-based CAPE forecasts valid for 1800 UTC on 12 August 2023 from (a) GraphCast, (b) Pangu-Weather, (c) FourCastNetv2, and (e) GFS, as well as (d) ERA-5 reanalysis valid at that time.

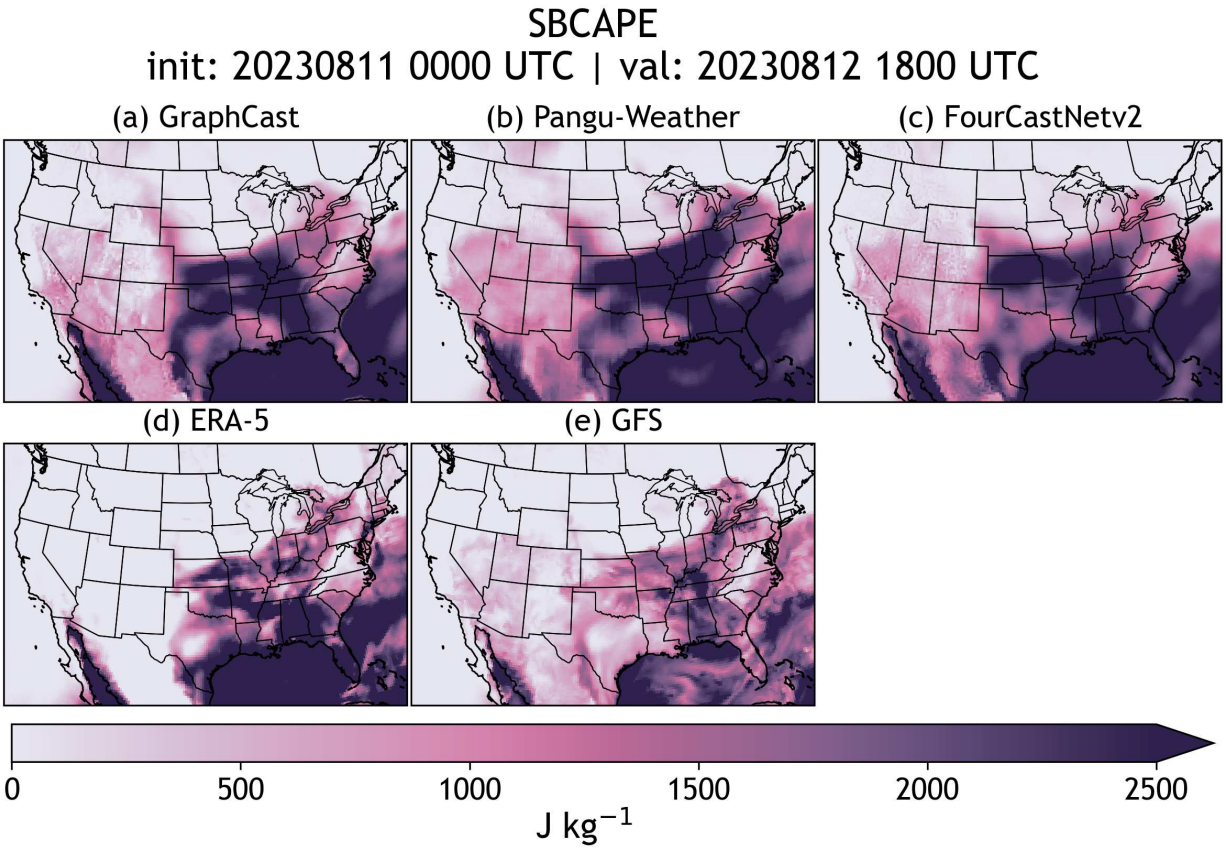

Figure 4.29: As in Fig. 4.29, but for the 42-h forecasts.

ERA-5 (Fig. 4.29a,c). While these forecasts generally place instability in the correct locations of where storms are likely, none of the DLWP models offer the level of detail provided by the GFS (Fig. 4.29e). Plus, the DLWP forecasts are still off by several hundreds of J kg^{-1} of SBCAPE even at a two-day lead time (not shown), so their output should be taken with caution.

The forecast soundings valid near Albany, NY at 0000 UTC for the case 2 period (Fig. 4.30) display similar findings to the case 1 results. The temperature profiles show better run-to-run agreement compared to the dew point profiles, but they also show more disagreement compared to the temperature profiles in case 1. Interestingly, in the lowest levels, the DLWP models are drier relative to ERA-5, but they are in closer agreement with the moisture profile in the observed sounding that was launched three hours prior (at 2100 UTC). This pattern is also evident in the mid- to upper-levels. It is worth noting that while there was not an inversion present on the observed soundings, some of the GFS forecasts show one (Fig. 4.30d), yet none of the DLWP models do, an observation that again may speak to limitations of their vertical resolution.

In the hodographs near Albany (Fig. 4.31), the DLWP models again show strong skill in capturing the shape of the hodograph in the ERA-5 reanalysis, including at advanced lead times. In this particular case, they perform much better than the GFS forecasts (Fig. 4.31d), which show a much different wind shear profile in some of the model runs. Still, like the first case, the magnitude of the shear in the DLWP forecasts is substantially underdone compared to the shear in the observed sounding. This discrepancy is particularly obvious in the lowest levels, where the wind speed at 925 hPa is at least double in the observed sounding compared to the other models. Shear in the lowest levels is crucial for tornado formation (e.g., Brooks et al., 2003; Thompson et al., 2003,

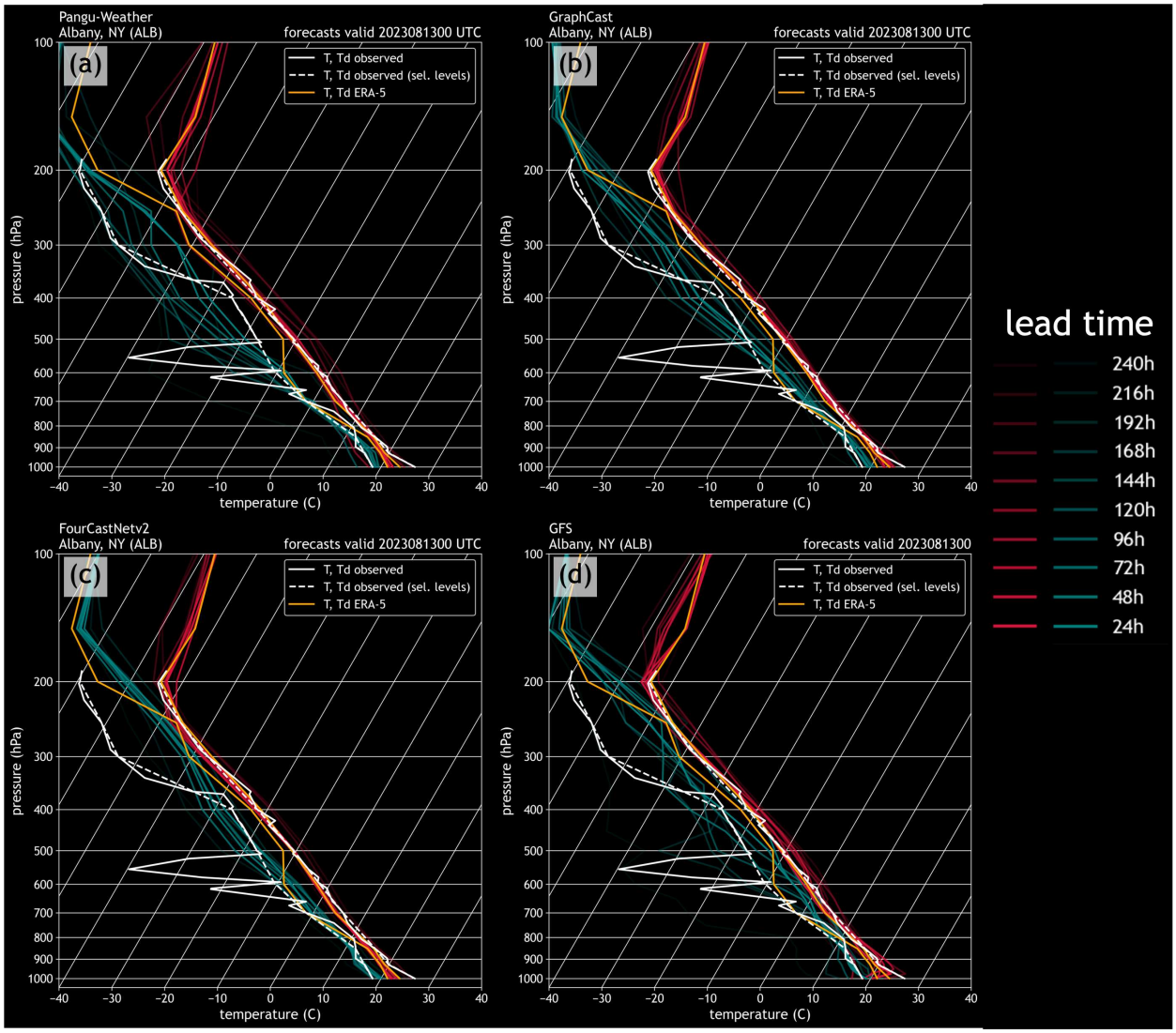

Figure 4.30: As in Fig. 4.23, but for case 2. Forecasts, reanalysis, and soundings are valid for 13 August 2023 at 0000 UTC at the point nearest Albany, NY. Note the overlaid sounding from ALB is valid 3 hours earlier at 2100 UTC on 12 August 2023, as this was the only available observed sounding at that site for this case.

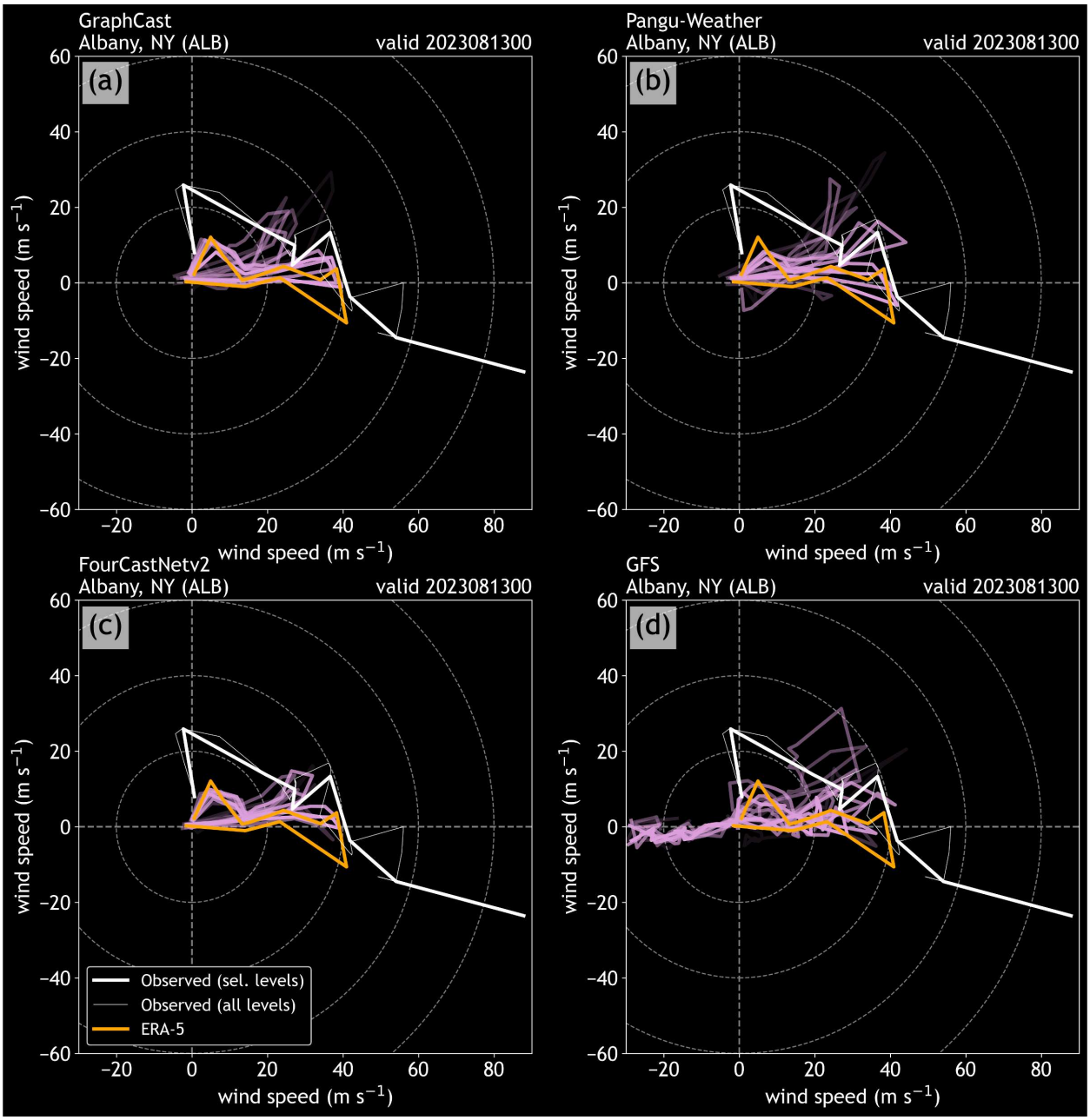

Figure 4.31: As in Fig. 4.26, for case 2. Forecasts, reanalysis, and soundings are valid 13 August 2023 at 0000 UTC. Note the overlaid hodograph from ALB is valid 3 hours earlier at 2100 UTC on 12 August 2023, as this was the only available observed sounding at that site for this case.

2012), so the fact that the models (and ERA-5) differ from observations to this extent in this case (as well as case 1) is noteworthy and has operational significance.

4.4 Discussion

To bring everything together, this study showed that vertical moisture profiles tended to be drier in the DLWP models relative to ERA-5, which helps explain why they also tended to have lower PWAT than ERA-5. DLWP temperature profiles also tended to be slightly cooler–especially in the mid-levels and at longer lead times. While the magnitude of these differences between the models and reanalysis may be on the order of only a few degrees Celsius, these differences could certainly have impacts on the derived SBCAPE profiles. Lower temperatures in the mid-levels would favor steeper low-level lapse rates and thus lead to larger SBCAPE values (with all else being equal), so these subtle differences could be contributing to the generally larger SBCAPE values in the DLWP models. However, the tendency for the DLWP models to have lower dew points and lower temperatures near the surface would counter this point. Additionally, the apparent inability of the DLWP models to identify inversions (due to their limited vertical resolution) has obvious implications on the SBCAPE calculations. By limiting ERA-5 to 13 vertical levels, the reanalysis also fails to account for inversions in its SBCAPE calculation, so inversions would not impact the analysis shown here. However, using the full ERA-5 profile would inevitably change the results. This point is discussed in greater detail below.

There are some similarities and differences between the findings in this study and those in Feldmann et al. (2024). For example, their work also examined a strongly-forced severe weather event as a case study, and like this study, they showed that the DLWP models were capable of predicting large CAPE values at long lead times (i.e., a week) over the location where severe storms

occurred. They also noted that the operational NWP model captured much finer details than the DLWP models. In their broader CAPE analysis (their work used most-unstable or MUCAPE), they also showed that FourCastNetv2 tended to predict the largest amounts of CAPE compared to the other models and that the models performed best when CAPE was small. However, they also found that FourCastNetv2 was less accurate at predicting humidity compared to GraphCast and Pangu-Weather. Examining dew point and PWAT in this work, there did not seem to be obvious deficiencies in that model that were not also present in GraphCast and Pangu-Weather. However, these results focused more on scenario-based evaluations rather than broader forecast metrics, so these results do not necessarily contradict their findings, as direct comparisons are limited. Regardless, both works showed that the DLWP struggled with characterizing moisture, and there is room for developmental improvements on this front.

It seems that using 13 vertical ERA-5 levels to compute the derived parameters (in the same way they are computed for the DLWP models) has impacts on the results presented here. Namely, preliminary research in this study used the raw PWAT and SBCAPE values from ERA-5 as the reference dataset to compare to the DLWP output. This early analysis showed that the DLWP output tended to have *more* PWAT and *less* SBCAPE than the ERA-5 raw variables– i.e., the opposite of the results that were shown when only 13 vertical ERA-5 levels were used in the PWAT and SBCAPE calculations. Baño-Medina et al. (2024) also showed in their work that PWAT values in the same DLWP models as this study tended to be consistently higher than the ERA-5 full-profile PWAT. However, they found that using the native PWAT from FourCastNetv2 (and version 1) did not lead to this effect. Thus, it is apparent that 1) the limited vertical levels in the DLWP models have serious implications on computing derived variables across layer depths and 2) there

is motivation for predicting these variables directly, as has already been suggested by previous work (Feldmann et al., 2024). This second point is further supported by preliminary examinations of other derived DLWP variables in the CIRA dataset (not discussed here) that rely on fine-scale information about the lower atmosphere, such as convective inhibition and mixed-layer CAPE. Early analyses of these DLWP model derived variables have shown that they exhibit some strange behaviors, such as mixed-layer CAPE values exceeding most-unstable CAPE values.

Despite these shortcomings, these results do offer some promise towards using derived DLWP data for severe certain weather prediction purposes. The case studies presented here show that the DLWP models can reasonably forecast the thermodynamics and dynamics of severe weather environments at lead times beyond a week (at least comparably to the present GFS). The guidance seems particularly useful for strongly-forced severe weather events, though it is helpful for weaker-forced events as well. More broadly, these results (and the results in Feldmann et al. (2024)) demonstrate that DLWP models are best used when the derived variable quantities are small (e.g., the cool season and overnight for instability and the warm season for shear). This knowledge is useful from a forecasting perspective, as a forecaster wanting to incorporate DLWP output in their work could adjust their trust in the model output for different situations. Further, given that aggregate evaluations of these models have shown that they capture large-scale atmospheric features with greater skill than existing NWP systems (Bi et al., 2023; Bonev et al., 2023; Lam et al., 2023) and that this work (as well as that presented by Feldmann et al. (2024)) shows that derived convective parameters from DLWP models are reasonable, it is fair to suggest that they can be integrated into operations in at least some capacity. However, it is imperative that forecasters

remain privy to their shortcomings (as they do with other model output) if they do choose to use them.

4.5 Summary and conclusions

This work examines convective environments and derived convective parameters over the CONUS in three DLWP models: GraphCast, Pangu-Weather, and FourCastNetv2. 22 months of DLWP model output and derived output from the Cooperative Institute for Research in the Atmosphere are compared to ERA-5 reanalysis and operational Global Forecasting System (GFS) forecasts. Three analyses of these forecasts are presented, each using an increasingly-detailed approach than the last. Those results are summarized here.

In the first analysis, derived values of precipitable water (PWAT), surface-base convective available potential energy (SBCAPE), and vertical wind shear between the surface to 500 hPa (SHR500) were compared across the 22 months of DLWP forecasts and ERA-5 reanalysis. The ERA-5 parameters were computed with the same 13 vertical levels as the DLWP forecasts for a fairer comparison. Broadly, the DLWP models depicted more PWAT and less SBCAPE than ERA-5. SBCAPE differences tend to be the largest in FourCastNetv2, while PWAT differences are roughly equivalent among the three systems. SHR500 differences between the models and reanalysis were mixed. The PWAT and SBCAPE differences between the DLWP models and ERA-5 were largest during the warm season, while SHR500 differences were largest during the cool season.

In the second analysis, profiles of dew point and temperature were analyzed with respect to ERA-5 and GFS forecasts during “convectively-favorable" forecast days across five geographically-diverse sites: Albany, NY, Bismarck, ND, Birmingham, AL, Lincoln, IL, and Norman, OK. Convectively-favorable forecast days were defined as days where a Storm Prediction Center (SPC)

day-1 convective outlook of the “enhanced" or greater intersected the National Weather Service County Warning Area of that site across the 22-month forecast archive. This approach yielded 90 cases. Examining the averaged dew point and temperature profiles across the many convectively-favorable environments showed that dew points in the DLWP models tended to be drier than ERA-5 throughout most of the atmospheric column, which helps explain the tendency of the models to have lower PWAT than ERA-5. These averaged differences became smaller at lead times within 5 days in the low- to mid- troposphere, but in GraphCast and Pangu-Weather, this pattern remained at levels above approximately 400 hPa as well as near the surface. Compared to the GFS, GraphCast and Pangu-Weather dew points were also lower (especially aloft), yet FourCastNetv2 dew points tended to be higher than the GFS in the upper levels. Still, there was substantial variability in dew point behavior across individual forecasts, making it challenging to identify clear biases. Temperature profiles in the DLWP models showed a much greater degree of predictability than the dew point profiles, but they still differed from ERA-5 and the GFS. Compared to ERA-5, the DLWP models tended to have higher temperatures above 200 hPa and lower temperatures below 200 hPa, especially between 800 hPa and 200 hPa. Similar results were found in comparisons to the GFS. It is hypothesized that because the DLWP models tend to have lower temperatures in the mid-levels, which correlate with steeper lapse rates, the models thus tend to have larger SBCAPE values (despite the lower amounts of moisture present).

In the third analysis, DLWP forecasts were examined for two severe weather outbreaks: one strongly-forced event and one weakly-forced event. As was shown in the second analysis, the DLWP models struggled with predicting moisture (as seen in the dew point profiles), but they were much better at predicting temperature. However, due to their limited vertical resolution, the

DLWP failed to model shallow stable layers that are able to be captured by the GFS. Examining hodographs of the DLWP forecasts showed that the models were very effective at modeling the shear direction and evolution relative to ERA-5 reanalysis, however they fell short of capturing the true magnitude of the shear compared to the observed sounding hodographs. Lastly, there tended to be greater predictability at longer lead times in the DLWP model forecasts–both in the evolution of SBCAPE forecasts and predicted sounding profiles–in the strongly-forced case compared to the weakly-forced case.

This work only scratches the surface of operationally-focused analyses of the DLWP forecasts for convective events, and there are a number of opportunities for future work. For example, it would be worthwhile to investigate the DLWP forecasts for specific types of convective events, such as nocturnal events or cold season events. Characteristics of forecasts associated with prolific tornado versus hail versus severe wind events would be another interesting research avenue. It would also be worthwhile to identify potential failure modes in DLWP systems, such as if there are situations when they predict large co-located values of CAPE and shear, but severe storms do not occur. And lastly, future work should aim to identify additional flaws in the DLWP model output that have relevance to convective forecasting (such as their inability to capture inversions). Knowledge of these DLWP model shortcomings and successes is key for supporting their potential future use in operational forecasting.

Chapter 5

Summary and Conclusions

5.1 Summary

In this dissertation, three studies that aim to better understand machine learning (ML)-based forecasts of severe convective weather hazards and environments at short to medium-range lead times are presented. Forecasts from architecturally-diverse ML systems are investigated through a number of lenses. Specifically, Chapter 2 elucidates how environmental information is used to make ML-based probabilistic severe weather forecasts. Chapter 3 probes forecast skill across a variety of severe weather-producing regimes. And finally, Chapter 4 investigates how derived convection-related parameters and environments in deep learning weather prediction (DLWP) model output compare to reanalysis and operational forecasts. Summaries of each of these studies are presented here.

In Chapter 2, Tree Interpreter (TI; Saabas, 2014), an explainable artificial intelligence method for analyzing random forest (RF) ML-based forecasts, is harnessed to analyze two years of probabilistic severe weather forecasts from the Colorado State University Machine Learning Probabilities (CSU-MLP) system. Two years of CSU-MLP forecasts of tornadoes, severe convective wind, and severe hail at two- and three-day lead times, as well as "any" severe convective hazard at day 4. TI is used to disaggregate the daily CSU-MLP probabilities into "feature contributions". This approach allows for an examination of how each of the environmental input variables (i.e., fields from the Global Ensemble Forecast System, or GEFS) contribute to the probabilistic forecasts temporally and spatially. Overarching results of this work show that characteristics of the feature

contributions that comprise the CSU-MLP probabilities resemble aspects of severe convective environments. For example, CSU-MLP probabilistic severe weather forecasts are strongly influenced by environmental input variables that are intimately tied to observed severe convection, including surface-based convective available potential energy (SBCAPE), wind shear, and LCL height. Additionally, there is seasonal variability in the magnitude to which these variables influence the forecasts, with thermodynamic input variables contributing to the probabilities more strongly than dynamic variables in the warm season (and vice-versa in the cool season). This variability follows seasonal availability of these ingredients in the real world: thermodynamic ingredients tend to be more plentiful in the summer, whereas kinematic ingredients tend to be more abundant in the winter. Examining relationships between the values of the feature contributions and the values of the environmental inputs from the GEFS further elucidates the relationships between them. For instance, larger values of SBCAPE and shear tend to correlate with larger SBCAPE and shear contributions (which would enhance CSU-MLP probabilities), whereas surface-based convective inhibition (SBCIN) often tends to negatively contribute to the probabilities. While much of this work focuses on understanding characteristics of the CSU-MLP feature contributions in aggregate over many forecasts, finer details can be gleaned by examining feature contributions in an individual case. Applying TI to the now-operational CSU-MLP forecasts for both aggregated and the individual forecasts can offer benefits to forecasting. Namely, the aggregate analysis illuminates that the CSU-MLP system makes its predictions in ways that are consistent with physical relationships in the atmosphere, which could enhance overall trust in the product. For individual forecasts, a forecaster can use TI to probe how specific environmental information contributes to

the CSU-MLP predictions–analogous to the "ingredients-based forecasting" framework that most are already well-acquainted with.

Chapter 3 also analyzes forecast output from the CSU-MLP system, but this project focuses on forecast skill across regimes. To identify regimes, self-organizing maps (SOMs) are leveraged to identify synoptic scale weather regimes in reanalysis data (ERA-5) over a two-year period. Two SOMs are constructed using 2100 UTC daily ERA-5 reanalysis standardized daily anomalies. One SOM is trained on SBCAPE and 10m-850hPa wind shear, and the other trained on SBCAPE and 10m-500hPa wind shear. Each SOM is trained to generate 9 regimes with meteorologically-distinct characteristics. Day-2 probabilistic forecasts of tornadoes, wind, and hail from the CSU-MLP system are sorted into their respective regimes that describe the ambient conditions during each of their forecast periods. Only day-2 forecasts that have a maximum probability at least in the Storm Prediction Center-defined "slight" risk category are considered (i.e., the maximum daily CSU-MLP forecast probability must exceed 5% for tornadoes or 15% for hail or wind to be considered a case). The CSU-MLP forecast skill is compared to forecast skill of the Storm Prediction Center (SPC) convective outlook skill across the various nodes and hazard predictions. Composites of environmental variables across each of the nodes show that there is overlap between the types of regimes identified by each of the SOMs, but each SOM also diagnoses a few of their own unique regimes. Frequency of CSU-MLP forecasts across the nodes varies spatially for each of the hazard forecasts, suggesting that some SOM regimes are closely related to CSU-MLP forecast probabilities in certain regions. Brier skill score is used to identify the best- and worst-performing CSU-MLP forecasts over the SOM nodes, and results show that the nodes characterized by the best- and worst-performing forecast skill varies by hazard type. Examining nodes of the SOM trained on

SBCAPE and 10m-850 hPa shear (“SOM0") and in greater detail shows that CSU-MLP best- and worst-performing CSU-MLP tornado forecasts both occur in strongly forced synoptic regimes, but the best-performing tornado forecasts appears to have a greater magnitude of forcing compared to the worst-performing regime based on environmental composite analysis. Best-performing wind forecasts also occur in a regime with stronger synoptic forcing compared to the regime where its worst-performing forecast tend to occur. Interestingly, the mean best- and mean worst-performing hail forecasts occur within the same node in the first SOM, suggesting that both the most-and least-successful forecasts can occur under similar synoptic conditions, and additional analysis may be needed to further diagnose catalysts for the forecast performance. Lastly, the most-skillful CSU-MLP and SPC forecasts tend to be associated with many reports, while the less-skillful forecasts tend to be correlated with fewer reports. This finding implies that poor skill likely stems from a combination of false alarms and isolated events.

Finally, in Chapter 4, focus is shifted towards analyzing a series of different ML-based forecasts. In this work, output from three DLWP systems–GraphCast Operational, Pangu-Weather, and FourCastNetv2-small–is examined in the context of severe convection. In this work, 22 months of Global Forecasting System (GFS)-initialized daily forecasts from these three models are used to generate a number of derived parameters that have relevance to severe convective forecasting. Environments and parameters from these model forecasts are analyzed in three ways. First, three of the derived parameters (precipitable water (PWAT), SBCAPE, and surface-500hPa shear (SHR500)) are compared to ERA-5 reanalysis spatially and seasonally over the 22 month period. Differences in PWAT and SBCAPE are greatest during the warm season, whereas shear differences are largest during the winter months. Average PWAT values tend to be lower than ERA-5 overall,

SBCAPE values tend to be generally larger than ERA-5 (with largest discrepancies occurring in the summer and over the Rockies), and results for SHR500 are mixed. Second, forecast point data from five geographically-diverse radiosonde sites are compared to ERA-5 reanalysis and the GFS output. In these data, it is shown that temperature profile data between the DLWP models, GFS, and ERA-5 show strong agreement, even at lead times of a week or more. Profiles of dew point have much larger differences between the DLWP models and both the GFS and reanalysis, with moisture differences between the datasets at times exceeding 10°C or more even at lead times as short as two days. Many of these characteristics are consistent across the different sites, suggesting that they may represent broader characteristics of the forecasts. Third, DLWP forecasts from two case studies of severe weather events–one strongly-forced case and one weekly-forced case–are examined. Soundings and hodographs show that the DLWP models struggle to capture moisture and may underestimate wind shear magnitude, though their predictions of temperature and directional shear are comparable to ERA-5 and observations. The DLWP also fail to predict low-level stable layers present in the observations in the GFS forecasts and the observations for these cases, which is likely caused by their limited vertical resolution. Understanding differences in the native variable forecasts in convective environments presents can help with hypotheses on the causes of discrepancies in the derived parameters in the models compared to reanalysis. For instance, perhaps the DLWP models' inability to capture shallow stable layers may be contributing to larger derived SBCAPE values. This work demonstrates that while derived convective parameters from some DLWP may offer insights on convective environments, the degree to which they are useful varies by model and variable, parameter, and even on a forecast-by-forecast basis.

5.2 Future work

Though the CSU-MLP probabilistic severe products are now operational at NOAA, there are a number of additional studies that could be conducted with the modeling system that have not yet been explored. For example, the studies presented in Chapters 2 and 3 inherently focus on days when the CSU-MLP system *does* issue probabilities for severe hazards. An investigation of days when the system *does not* issue probabilities (both for correct negatives and misses) would also be interesting, particularly if explainability approaches (such as TI) are used. Further, qualitative survey results containing Storm Prediction Center forecaster evaluations of the CSU-MLP severe probabilities have not been thoroughly explored. These results likely offer invaluable suggestions for how the probabilities could be better presented for operational forecasting in the future. With this said, there are a number of parameters in the CSU-MLP system that could be tuned in future experiments to support efforts to optimize its performance. Potential experiments include, but are not limited to: adjusting environmental input variables (e.g., computing shear parameters over fixed rather than pressure-based layer depths), changing predictor assembly techniques (e.g., increasing/decreasing the spatial radius of inputted grid points), or using environmental data from other models for initialization (e.g. from the ECMWF Integrated Forecasting System, or IFS).

Development of DLWP systems that emulate NWP has increased dramatically in recent years, and with new systems becoming available (and open-source) frequently, nearly endless opportunities exist to study their output using explainability methods and statistics, as well as leverage their output for specific tasks related to convective environments. As a starting place, there are a number of convection-pertinent parameters derived from the CIRA archive of deep learning forecasts that were not examined here, such as lifted index, storm relative helicity, and convective inhibition.

Exploring the quality of these parameters would be a logical next step from the work presented in Chapter 4, though it is hypothesized that they may be poor given the limited vertical resolution of the DLWP models. Additionally, CIRA has begun to run additional DLWP forecasts than those used in this dissertation, including the ECMWF's Artificial Intelligence Forecasting System (AIFS; Lang et al., 2024) and IFS-initialized forecasts for Pangu-Weather, GraphCast, and FourCastNetv2. Comparing convective environments and derived convective parameters from these forecasts with those used in Chapter 4 present another avenue for future work. Exploring these parameters from the GFS-initialized forecasts in other locations globally (building on work by Feldmann et al. (2024)) or from the perspective of other convective hazards (e.g., flooding) may also be worthwhile.

Beyond analyzing output of the DLWP models, these output data could also be harnessed to generate post-processed output for predicting severe weather. A few studies have already begun to use them for this purpose (e.g., Flora and Potvin, 2024; Hill and Radford, 2024), but there are many existing data-driven forecasting systems that could experiment with incorporating DLWP output as inputs to the models (as opposed to relying on NWP output). If successful, these endeavors could allow such forecast products to be generated and viewed by forecasters in a more timely manner.

5.3 Final thoughts

ML offers a promising new frontier for weather forecasting. With multiple new ML-based weather prediction systems being released practically each month, this field will continue to evolve very quickly in the coming years, and it is impossible to forecast how it may look a decade from now. As developers continue to advance and fine-tune these models, it is essential that scientists continue to dedicate time to studying ML forecasts from physics-informed perspectives. Because

a large percentage of this development is being pioneered in the private sector at present (largely because of their computing advantages), perhaps academia will feature a smaller role in developing ML-based forecasting models and a larger role in employing their subject-area expertise to understand ML predictions (Bauer, 2024). These scientific efforts are essential to bringing these datasets to operational weather forecasters: as ML forecast output becomes more widely disseminated, forecasters will continue to have increased interest in understanding their characteristics and biases, so they can better use them in their day-to-day predictions.

This dissertation research supports these operationally-informed objectives by analyzing output from just a couple of ML-based weather forecasting models to understand and communicate their strengths, shortcomings, tendencies, and "quirks". It is hoped that the results presented herein underscore the importance of scientifically investigating these new types of forecast output before they are used in decision-making. While this work supports broader efforts towards trust and transparency of ML prediction systems, there are still a number of ongoing discipline-specific barriers (e.g., McGovern et al., 2022a), such as forecaster training, that will need to be overcome before such systems can be fully assimilated into the weather enterprise.

Bibliography

Allen, J. T., and M. K. Tippett, 2015: The Characteristics of United States Hail Reports: 1955-2014. *EJSSM*, **10 (3)**, 1–31, https://doi.org/10.55599/ejssm.v10i3.60.

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2017: Self-Organizing Maps for the Investigation of Tornadic Near-Storm Environments. *Weather and Forecasting*, **32 (4)**, 1467–1475, https://doi.org/10.1175/WAF-D-17-0034.1.

Arcodia, M., and Coauthors, 2022: Applied Machine Learning Tutorial for Earth Scientists. URL https://github.com/eabarnes1010/ml_tutorial_csu.

Bauer, P., 2024: What if? Numerical weather prediction at the crossroads. arXiv, URL http://arxiv.org/abs/2407.03787, arXiv:2407.03787 [physics].

Baño-Medina, J., A. Sengupta, A. Michaelis, L. D. Monache, and D. Watson-Parris, 2024: Harnessing AI data-driven global weather models for climate attribution: An analysis of the 2017 Oroville Dam extreme atmospheric river. arXiv.

Ben Bouallègue, Z., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, https://doi.org/10.1175/BAMS-D-23-0162.1.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619 (7970)**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bodnar, C., and Coauthors, 2024: Aurora: A Foundation Model of the Atmosphere. arXiv.

Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, 2023: Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere. arXiv, URL http://arxiv.org/abs/2306.03838, arXiv:2306.03838 [physics].

Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, 5–32.

Brier, G. W., 1950: Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78 (1)**, 1–3.

Brooks, H. E., 2007: Ingredients-Based Forecasting. *Atmospheric Convection: Research and Operational Forecasting Aspects*, D. B. Giaiotti, R. Steinacker, and F. Stel, Eds., Vol. 475, Springer Vienna, Vienna, 133–140, https://doi.org/10.1007/978-3-211-69291-2_12, URL http://link.springer.com/10.1007/978-3-211-69291-2_12, series Title: CISM International Centre for Mechanical Sciences.

Brooks, H. E., and J. P. Craven, 2002: A database of proximity soundings for significant severe thunderstorms, 1957–1993. *Preprints*, American Meteorological Society, San Antonio, Texas, 639–642.

Brooks, H. E., J. W. Lee, and J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, **67-68**, 73–94, https://doi.org/10.1016/S0169-8095(03)00045-0.

Bunkers, M. J., S. R. Fleegel, T. Grafenauer, C. J. Schultz, and P. N. Schumacher, 2020: Observations of Hail–Wind Ratios from Convective Storm Reports across the Continental United States. *Weather and Forecasting*, **35 (2)**, 635–656, https://doi.org/10.1175/WAF-D-19-0136.1.

Burke, A., N. Snook, D. J. Gagne Ii, S. McCorkle, and A. McGovern, 2020: Calibration of Machine Learning–Based Probabilistic Hail Predictions for Operational Forecasting. *Weather and Forecasting*, **35 (1)**, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1.

Cains, M. G., C. D. Wirz, J. L. Demuth, A. Bostrom, D. J. Gagne, A. McGovern, R. A. Sobash, and D. Madlambayan, 2024: Exploring NWS Forecasters' Assessment of AI Guidance Trustworthiness. *Weather and Forecasting*, **39 (8)**, 1219–1241, https://doi.org/10.1175/WAF-D-23-0180.1.

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. *Weather and Forecasting*, **37 (8)**, 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. *Weather and Forecasting*, **38 (8)**, 1271–1293, https://doi.org/10.1175/WAF-D-22-0187.1.

Chen, L., X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li, 2023: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim Atmos Sci*, **6 (1)**, 190, https://doi.org/10.1038/s41612-023-00512-1.

Clark, A., and Coauthors, 2021: Spring Forecasting Experiment 2021: Preliminary Findings and Results. Tech. rep., 86 pp. URL https://hwt.nssl.noaa.gov/sfe/2021/docs/HWT_SFE_2021_Prelim_Findings_FINAL.pdf.

Clark, A., and Coauthors, 2022: Spring Forecasting Experiment 2022: Preliminary Findings and Results. Tech. rep., 95 pp. URL https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE_2022_Prelim_Findings_FINAL.pdf.

Clark, A. J., and E. D. Loken, 2022: Machine Learning–Derived Severe Weather Probabilities from a Warn-on-Forecast System. *Weather and Forecasting*, **37 (10)**, 1721–1740, https://doi.org/10.1175/WAF-D-22-0056.1.

Clark, A. J., and Coauthors, 2023: Spring Forecasting Experiment 2023: Preliminary Findings and Results. Tech. rep., 75 pp. URL https://hwt.nssl.noaa.gov/sfe/2023/docs/HWT_SFE_2023_Prelim_Findings_v1.pdf.

DeMaria, M., J. L. Franklin, G. Chirokova, J. Radford, R. DeMaria, K. D. Musgrave, and I. Ebert-Uphoff, 2024: Evaluation of Tropical Cyclone Track and Intensity Forecasts from Artificial Intelligence Weather Prediction (AIWP) Models. arXiv, URL https://arxiv.org/abs/2409.06735#:~:text=The%20AIWP%20models%20almost%20always,at%20the%20longer%20time%20periods., https://doi.org/https://doi.org/10.48550/arXiv.2409.06735.

Doswell, C. A., 1980: Synoptic-Scale Environments Associated with High Plains Severe Thunderstorms. *Bull. Amer. Meteor. Soc.*, **61 (11)**, 1388–1400, https://doi.org/10.1175/1520-0477(1980)061<1388:SSEAWH>2.0.CO;2.

Doswell, C. A., 1987: The Distinction between Large-Scale and Mesoscale Contribution to Severe Convection: A Case Study Example. *Wea. Forecasting*, **2 (1)**, 3–16, https://doi.org/10.1175/1520-0434(1987)002<0003:TDBLSA>2.0.CO;2.

Doswell, C. A., H. E. Brooks, and M. P. Kay, 2005: Climatological Estimates of Daily Local Nontornadic Severe Thunderstorm Probability for the United States. *Weather and Forecasting*, **20 (4)**, 577–595, https://doi.org/10.1175/WAF866.1.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash Flood Forecasting: An Ingredients-Based Methodology. *Wea. Forecasting*, **11 (4)**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Doswell, C. A., and D. M. Schultz, 2006: On the Use of Indices and Parameters in Forecasting Severe Storms. *EJSSM*, **1 (3)**, 1–22, https://doi.org/10.55599/ejssm.v1i3.4.

Ebert-Uphoff, I., and K. Hilburn, 2023: The outlook for AI weather prediction. *Nature*, **619 (7970)**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

ECMWF, 2023: Massive Online Open Course Machine Learning in Weather and Climate. URL https://www.spc.noaa.gov/misc/about.html.

Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and Climatological Impacts of Convective Wind Estimations. *Journal of Applied Meteorology and Climatology*, **57 (8)**, 1825–1845, https://doi.org/10.1175/JAMC-D-17-0306.1.

Escobedo, J. A., 2022: CSU-MLP GEFS Day-1 "First Guess" Excessive Rainfall Forecasts: Aggregate Evaluation and Synoptic Regimes of Best- and Worst-Performing Forecasts. Ph.D. thesis, Colorado State University.

Escobedo, J. A., and R. S. Schumacher, 2024: Patterns of Performance from a Machine Learning Prediction System for Excessive Rainfall. *In Review with Weather and Forecasting*.

Feldmann, M., T. Beucler, M. Gomez, and O. Martius, 2024: Lightning-Fast Thunderstorm Warnings: Predicting Severe Convective Environments with Global Neural Weather Models. arXiv, URL http://arxiv.org/abs/2406.09474, arXiv:2406.09474 [physics].

Flora, M., C. Potvin, A. McGovern, and S. Handler, 2022: Comparing Explanation Methods for Traditional Machine Learning Models Part 1: An Overview of Current Methods and Quantifying Their Disagreement. arXiv, URL http://arxiv.org/abs/2211.08943, arXiv:2211.08943 [physics, stat].

Flora, M. L., and C. Potvin, 2024: WoFSCast: A machine learning model for predicting thunderstorms at watch-to-warning scales. URL https://essopenarchive.org/users/829074/articles/1223249-wofscast-a-machine-learning-model-for-predicting-thunderstorms-at-watch-to-warning-scales?commit=28a29eb79f581eb711296cc6e87b572a2f091cc4, https://doi.org/10.22541/essoar.172574503.30734251/v1.

Flora, M. L., C. K. Potvin, A. McGovern, and S. Handler, 2024: A Machine Learning Explainability Tutorial for Atmospheric Sciences. *Artificial Intelligence for the Earth Systems*, **3 (1)**, e230 018, https://doi.org/10.1175/AIES-D-23-0018.1.

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. *Monthly Weather Review*, **149 (5)**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, **32 (5)**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting Tornadoes Using Convection-Permitting Ensembles. *Weather and Forecasting*, **31 (1)**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended Probabilistic Tornado Forecasts: Combining Climatological Frequencies with NSSL–WRF Ensemble Forecasts. *Weather and Forecasting*, **33 (2)**, 443–460, https://doi.org/10.1175/WAF-D-17-0132.1.

Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the U.S. using ERA5 proximity soundings. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-21-0056.1.

Guibas, J., M. Mardani, Z. Li, A. Tao, A. Aanandkumar, and B. Catanzaro, 2022: Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers. 15.

Guyer, J. L., and A. R. Dean, 2010: Tornadoes within weak CAPE environments across the continental United States. American Meteorological Society, Denver, CO, Vol. 1.5, URL https://ams.confex.com/ams/25SLS/webprogram/Paper175725.html.

Hamill, T. M., and Coauthors, 2022: The Reanalysis for the Global Ensemble Forecast System, Version 12. *Monthly Weather Review*, **150 (1)**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Heinselman, P. L., and Coauthors, 2024: Warn-on-Forecast System: From Vision to Reality. *Weather and Forecasting*, **39 (1)**, 75–95, https://doi.org/10.1175/WAF-D-23-0147.1.

Herman, G. R., and R. S. Schumacher, 2018a: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Monthly Weather Review*, **146 (5)**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Herman, G. R., and R. S. Schumacher, 2018b: "Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation. *Mon. Wea. Rev.*, **146 (6)**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

Hersbach, H., 2023: ERA5 reanalysis now available from 1940. URL https://www.ecmwf.int/en/newsletter/175/news/era5-reanalysis-now-available-1940.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Q.J.R. Meteorol. Soc.*, **146 (730)**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hewitson, B., and R. Crane, 2002: Self-organizing maps: applications to synoptic climatology. *Clim. Res.*, **22**, 13–26, https://doi.org/10.3354/cr022013.

Hilburn, K. A., I. Ebert-Uphoff, and S. D. Miller, 2021: Development and Interpretation of a Neural-Network-Based Synthetic Radar Reflectivity Estimator Using GOES-R Satellite Observations. *Journal of Applied Meteorology and Climatology*, **60 (1)**, 3–21, https://doi.org/10.1175/JAMC-D-20-0084.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. *Monthly Weather Review*, **148 (5)**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hill, A. J., and J. T. Radford, 2024: Postprocessing Data-Driven AI Forecasting Models for Hazardous Weather Prediction. Virginia Beach, VA.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-21-0026.1.

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. *Weather and Forecasting*, **38 (2)**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Hua, Z., and A. K. Anderson-Frey, 2022: Self-Organizing Maps for the Classification of Spatial and Temporal Variability of Tornado-Favorable Parameters. *Monthly Weather Review*, **150 (2)**, 393–407, https://doi.org/10.1175/MWR-D-21-0168.1.

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2019: Classifying Convective Storms Using Machine Learning. *Weather and Forecasting*, **35 (2)**, 537–559, https://doi.org/10.1175/WAF-D-19-0170.1.

Johns, R. H., 1984: A Synoptic Climatology of Northwest-Flow Severe Weather Outbreaks. Part II: Meteorological Parameters and Synoptic Patterns. *Mon. Wea. Rev.*, **112 (3)**, 449–464, https://doi.org/10.1175/1520-0493(1984)112<0449:ASCONF>2.0.CO;2.

Johns, R. H., and C. A. Doswell, 1992: Severe Local Storms Forecasting. *Weather and Forecasting*, **7 (4)**, 588–612, https://doi.org/https://doi.org/10.1175/1520-0434(1992)007<0588:SLSF>2.0.CO;2.

Kis, A. K., and J. M. Straka, 2010: Nocturnal Tornado Climatology*. *Weather and Forecasting*, **25 (2)**, 545–561, https://doi.org/10.1175/2009WAF2222294.1.

Kiviluoto, K., 1996: Topology Preservation in Self-organizing Maps. *Proceedings of International Conference on Neural Networks (ICNN'96)*, Vol. 1, 294–299 vol.1.

Kohonen, T., 1982: Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, **43 (1)**, 59–69, https://doi.org/10.1007/BF00337288.

Kohonen, T., 1990: The self-organizing map. *Proc. IEEE*, **78 (9)**, 1464–1480, https://doi.org/10.1109/5.58325.

Kohonen, T., 2013: Essentials of the self-organizing map. *Neural Networks*, **37**, 52–65, https://doi.org/10.1016/j.neunet.2012.09.018.

Kolczynski, W. C., and J. P. Hacker, 2014: The Potential for Self-Organizing Maps to Identify Model Error Structures. *Monthly Weather Review*, **142 (4)**, 1688–1696, https://doi.org/10.1175/MWR-D-13-00189.1.

Krocak, M. J., and H. E. Brooks, 2018: Climatological Estimates of Hourly Tornado Probability for the United States. *Weather and Forecasting*, **33 (1)**, 59–69, https://doi.org/10.1175/WAF-D-17-0123.1.

Krocak, M. J., and H. E. Brooks, 2020: An Analysis of Subdaily Severe Thunderstorm Probabilities for the United States. *Weather and Forecasting*, **35 (1)**, 107–112, https://doi.org/10.1175/WAF-D-19-0145.1.

Lagerquist, R., A. McGovern, C. R. Homeyer, D. J. Gagne Ii, and T. Smith, 2020: Deep Learning on Three-Dimensional Multiscale Data for Next-Hour Tornado Prediction. *Monthly Weather Review*, **148 (7)**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind. *Weather and Forecasting*, **32 (6)**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which Polarimetric Variables Are Important for Weather/No-Weather Discrimination? *Journal of Atmospheric and Oceanic Technology*, **32 (6)**, 1209–1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382 (6677)**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lang, S., and Coauthors, 2024: AIFS - ECMWF's data-driven forecasting system. arXiv, URL http://arxiv.org/abs/2406.01465, arXiv:2406.01465 [physics].

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests. *Weather and Forecasting*, **35 (4)**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Loken, E. D., A. J. Clark, and A. McGovern, 2022: Comparing and Interpreting Differently Designed Random Forests for Next-Day Severe Weather Hazard Prediction. *Weather and Forecasting*, **37 (6)**, 871–899, https://doi.org/10.1175/WAF-D-21-0138.1.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests. *Weather and Forecasting*, **34 (6)**, 2017–2044, https://doi.org/10.1175/WAF-D-19-0109.1.

Lundberg, S. M., and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*, **2 (1)**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Mamalakis, A., I. Ebert-Uphoff, and E. Barnes, 2022: Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science. *xxAI - Beyond Explainable AI*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., Vol. 13200, Springer International Publishing, Cham, 315–339, https://doi.org/10.1007/978-3-031-04083-2_16, URL https://link.springer.com/10.1007/978-3-031-04083-2_16, series Title: Lecture Notes in Computer Science.

May, R. M., and Coauthors, 2022: Metpy: A meteorological python library for data analysis and visualization. *Bulletin of the American Meteorological Society*, **103 (10)**, E2273 – E2284, https://doi.org/10.1175/BAMS-D-21-0125.1.

Mazurek, A. C., A. J. Hill, R. S. Schumacher, and H. J. McDaniel, 2025: Can ingredients-based forecasting be learned? disentangling a random forest's severe weather predictions. *Weather and Forecasting*, **40 (2)**, 237 – 258, https://doi.org/10.1175/WAF-D-23-0193.1.

McGovern, A., A. Bostrom, M. McGraw, R. J. Chase, D. J. Gagne, I. Ebert-Uphoff, K. D. Musgrave, and A. Schumacher, 2024: Identifying and Categorizing Bias in AI/ML for Earth

Sciences. *Bulletin of the American Meteorological Society*, **105 (3)**, E567–E583, https://doi.org/10.1175/BAMS-D-23-0196.1.

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A Review of Machine Learning for Convective Weather. *Artificial Intelligence for the Earth Systems*, **2 (3)**, e220 077, https://doi.org/10.1175/AIES-D-22-0077.1.

McGovern, A., I. Ebert-Uphoff, D. J. Gagne II, and A. Bostrom, 2022a: The Need for Ethical, Responsible, and Trustworthy Artificial Intelligence for Environmental Sciences. *Environ. Data Science*, **1**, e6, https://doi.org/10.1017/eds.2022.5.

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

McGovern, A., and Coauthors, 2022b: NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). *Bulletin of the American Meteorological Society*, **103 (7)**, E1658–E1668, https://doi.org/10.1175/BAMS-D-21-0020.1.

McNulty, R. P., 1995: Severe and Convective Weather: A Central Region Forecasting Challenge. *Wea. Forecasting*, **10 (2)**, 187–202, https://doi.org/10.1175/1520-0434(1995)010<0187:SACWAC>2.0.CO;2.

Mecikalski, J. R., T. N. Sandmæl, E. M. Murillo, C. R. Homeyer, K. M. Bedka, J. M. Apke, and C. P. Jewett, 2021: Random Forest Model to Assess Predictor Importance and Nowcast

Severe Storms using High-Resolution Radar–GOES Satellite–Lightning Observations. *Monthly Weather Review*, https://doi.org/10.1175/MWR-D-19-0274.1.

Molina, M. J., and Coauthors, 2023: A Review of Recent and Emerging Machine Learning Applications for Climate Variability and Weather Phenomena. *Artificial Intelligence for the Earth Systems*, 1–46, https://doi.org/10.1175/AIES-D-22-0086.1.

Molnar, C., 2022: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.)*. URL https://christophm.github.io/interpretable-ml-book/.

Moosavi, V., S. Packmann, and I. Vallés, 2014: SOMPY: A Python Library for Self Organizing Map (SOM).

Nixon, C. J., and J. T. Allen, 2022: Distinguishing between Hodographs of Severe Hail and Tornadoes. *Weather and Forecasting*, **37 (10)**, 1761–1782, https://doi.org/10.1175/WAF-D-21-0136.1.

Nixon, C. J., J. T. Allen, and M. Taszarek, 2023: Hodographs and Skew-Ts of Hail-Producing Storms. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-23-0031.1.

NOAA National Centers for Environmental Information, 2023: NCEI storm events database. URL https://www.ncdc.noaa.gov/stormevents/.

NOAA National Centers for Environmental Information, 2024: Billion-Dollar Weather and Climate Disasters. URL https://www.ncei.noaa.gov/access/billions/.

NOAA Storm Prediction Center, 2023a: Severe Weather Maps, Graphics, and Data Page: 30-year Severe Weather Climatology (1986-2015). URL https://www.spc.noaa.gov/wcm/.

NOAA Storm Prediction Center, 2023b: SPC Products. URL https://www.spc.noaa.gov/misc/about.html.

Nowotarski, C. J., and A. A. Jensen, 2013: Classifying Proximity Soundings with Self-Organizing Maps toward Improving Supercell and Tornado Forecasting. *Weather and Forecasting*, **28 (3)**, 783–801, https://doi.org/10.1175/WAF-D-12-00125.1.

Nowotarski, C. J., and E. A. Jones, 2018: Multivariate Self-Organizing Map Approach to Classifying Supercell Tornado Environments Using Near-Storm, Low-Level Wind and Thermodynamic Profiles. *Weather and Forecasting*, **33 (3)**, 661–670, https://doi.org/10.1175/WAF-D-17-0189.1.

Olivetti, L., and G. Messori, 2024: Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather and GraphCast. URL https://egusphere.copernicus.org/preprints/2024/egusphere-2024-1042/, https://doi.org/10.5194/egusphere-2024-1042.

Pathak, J., and Coauthors, 2022: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. arXiv, URL http://arxiv.org/abs/2202.11214, arXiv:2202.11214 [physics].

Pathak, J., and Coauthors, 2024: Kilometer-Scale Convection Allowing Model Emulation using Generative Diffusion Modeling.

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

Radford, J. T., I. Ebert-Uphoff, J. Q. Stewart, K. D. Musgrave, R. DeMaria, N. Tourville, and K. Hilburn, 2025: Accelerating community-wide evaluation of ai models for global weather prediction by facilitating access to model output. *Bulletin of the American Meteorological Society*, **106 (1)**, E68 – E76, https://doi.org/10.1175/BAMS-D-24-0057.1.

Radford, J. T., and G. M. Lackmann, 2023a: Assessing Variations in the Predictive Skill of Ensemble Snowband Forecasts with Object-Oriented Verification and Self-Organizing Maps. *Weather and Forecasting*, **38 (9)**, 1673–1693, https://doi.org/10.1175/WAF-D-23-0004.1.

Radford, J. T., and G. M. Lackmann, 2023b: Improving High-Resolution Ensemble Forecast (HREF) System Mesoscale Snowband Forecasts with Random Forests. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-23-0005.1.

Rasp, S., and Coauthors, 2024: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *J Adv Model Earth Syst*, **16 (6)**, e2023MS004 019, https://doi.org/10.1029/2023MS004019.

Roebber, P. J., and S. Smith, 2023: Prospects for Machine Learning Activity within the United States National Weather Service. *Bulletin of the American Meteorological Society*, **104 (7)**, E1333–E1344, https://doi.org/10.1175/BAMS-D-22-0181.1.

Saabas, A., 2014: Interpreting random forests. URL https://blog.datadive.net/interpreting-random-forests/.

Schmude, J., and Coauthors, 2024: Prithvi WxC: Foundation Model for Weather and Climate. arXiv, URL http://arxiv.org/abs/2409.13598, arXiv:2409.13598 [physics].

Schumacher, R. S., A. J. Hill, M. Klein, J. A. Nelson, M. J. Erickson, S. M. Trojniak, and G. R. Herman, 2021: From Random Forests to Flood Forecasts: A Research to Operations Success Story. *Bulletin of the American Meteorological Society*, **102 (9)**, E1742–E1755, https://doi.org/10.1175/BAMS-D-20-0186.1.

Schwartz, C. S., G. S. Romine, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015: A Real-Time Convection-Allowing Ensemble Prediction System Initialized by Mesoscale Ensemble Kalman Filter Analyses. *Weather and Forecasting*, **30 (5)**, 1158–1181, https://doi.org/10.1175/WAF-D-15-0013.1.

Scott, D. W., 1979: On optimal and data-based histograms. *Biometrika*, **66 (3)**, 605–610, https://doi.org/10.1093/biomet/66.3.605.

Shapley, L., 1953: A value for n-person games. *Contributions to the Theory of Games*, 307–317.

Sherburn, K. D., and M. D. Parker, 2014: Climatology and Ingredients of Significant Severe Convection in High-Shear, Low-CAPE Environments. *Weather and Forecasting*, **29 (4)**, 854–877, https://doi.org/10.1175/WAF-D-13-00041.1.

Sherburn, K. D., M. D. Parker, J. R. King, and G. M. Lackmann, 2016: Composite Environments of Severe and Nonsevere High-Shear, Low-CAPE Convective Events. *Weather and Forecasting*, **31 (6)**, 1899–1927, https://doi.org/10.1175/WAF-D-16-0086.1.

Smith, B. T., T. E. Castellanos, A. C. Winters, C. M. Mead, A. R. Dean, and R. L. Thompson, 2013: Measured Severe Convective Wind Climatology and Associated Convective Modes of Thunderstorms in the Contiguous United States, 2003–09. *Weather and Forecasting*, **28 (1)**, 229–236, https://doi.org/10.1175/WAF-D-12-00096.1.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) Severe Weather and Aviation Products: Initial Operating Capabilities. *Bulletin of the American Meteorological Society*, **97 (9)**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Weather and Forecasting*, **26 (5)**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

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

Song, F., Z. Feng, L. R. Leung, R. A. Houze Jr., J. Wang, J. Hardin, and C. R. Homeyer, 2019: Contrasting Spring and Summer Large-Scale Environments Associated with Mesoscale Convective Systems over the U.S. Great Plains. *Journal of Climate*, **32 (20)**, 6749–6767, https://doi.org/10.1175/JCLI-D-18-0839.1.

Taszarek, M., J. T. Allen, T. Púčik, K. A. Hoogewind, and H. E. Brooks, 2020: Severe Convective Storms across Europe and the United States. Part II: ERA5 Environments Associated with Lightning, Large Hail, Severe Wind, and Tornadoes. *Journal of Climate*, **33 (23)**, 10 263–10 286, https://doi.org/10.1175/JCLI-D-20-0346.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close Proximity Soundings within Supercell Environments Obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18 (6)**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:

CPSWSE>2.0.CO;2.

Thompson, R. L., B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective Modes for Significant Severe Thunderstorms in the Contiguous United States. Part II: Supercell and QLCS Tornado Environments. *Weather and Forecasting*, **27 (5)**, 1136–1154, https://doi.org/10.1175/WAF-D-11-00116.1.

Tirone, E., S. Pal, W. A. Gallus, S. Dutta, R. Maitra, J. Newman, E. Weber, and I. Jirak, 2024: A Machine Learning Approach to Improve the Usability of Severe Thunderstorm Wind Reports. *Bulletin of the American Meteorological Society*, https://doi.org/10.1175/BAMS-D-22-0268.1.

Trojniak, S., and J. Correia, 2022: 2022 Flash Flood and Intense Rainfall (FFaIR) Final Report: Results and Findings. Tech. rep., Hydrometeorology Testbed, NOAA Weather Prediction Center.

Vaughan, A., and Coauthors, 2024: Aardvark Weather: end-to-end data-driven weather forecasting. arXiv, URL http://arxiv.org/abs/2404.00411, arXiv:2404.00411 [physics].

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, 1999: Self-organizing map in Matlab: the SOM Toolbox, Espoo, Finland. 35–40.

Warren, R. A., H. Richter, and R. L. Thompson, 2021: Spectrum of Near-Storm Environments for Significant Severe Right-Moving Supercells in the Contiguous United States. *Monthly Weather Review*, **149 (10)**, 3299–3323, https://doi.org/10.1175/MWR-D-21-0006.1.

Wendt, N. A., and I. L. Jirak, 2021: An Hourly Climatology of Operational MRMS MESH-Diagnosed Severe and Significant Hail with Comparisons to Storm Data Hail Reports. *Weather and Forecasting*, **36 (2)**, 645–659, https://doi.org/10.1175/WAF-D-20-0158.1.

Wetzel, S. W., and J. E. Martin, 2001: An Operational Ingredients-Based Methodology for Forecasting Midlatitude Winter Season Precipitation. *Wea. Forecasting*, **16 (1)**, 156–167, https://doi.org/10.1175/1520-0434(2001)016<0156:AOIBMF>2.0.CO;2.

Xu, J., W. Zhou, Z. Fu, H. Zhou, and L. Li, 2021: A Survey on Green Deep Learning. arXiv, URL http://arxiv.org/abs/2111.05193, arXiv:2111.05193 [cs].

Yao, H., X. Li, H. Pang, L. Sheng, and W. Wang, 2020: Application of random forest algorithm in hail forecasting over Shandong Peninsula. *Atmospheric Research*, **244**, 105 093, https://doi.org/10.1016/j.atmosres.2020.105093.

Zhong, X., L. Chen, X. Fan, W. Qian, J. Liu, and H. Li, 2024: FuXi-2.0: Advancing machine learning weather forecasting model for practical applications. arXiv, URL http://arxiv.org/abs/2409.07188, arXiv:2409.07188 [physics].

Zhou, X., and Coauthors, 2022: The Development of the NCEP Global Ensemble Forecast System Version 12. *Weather and Forecasting*, **37 (6)**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Appendix A

Supplementary Material for Chapter 3

A.1 Example SOM parameter tuning experiments

A few SOM parameter tuning experiments are provided here as justification for the inputs used in the final SOM configurations. Details on each experiment are provided in the figure captions. All tuning experiments shown here were conducted with the SOM0 environmental inputs (i.e., SBCAPE and 10m-850hPa shear) with a 3x3 rectangular lattice structure.

These experiments show that relationships between the SOM parameters and SOM skill are non-linear, and it is difficult to parse some of the patterns. However, the results show that changes to the parameters generally do not result in remarkable skill differences with respect to topographic and quantization errors. A few overarching findings from these experiments can be summarized as follows:

- Increasing rough and fine-tuning training lengths beyond a few epochs does not seem to result in large skill differences. Fine-tuning training lengths seem to have a particularly small influence on skill (c.f., Figs. A.1; A.2).

- Despite little correlation between training length and skill, differences in skill across various random seeds seems to converge across when rough training length is increased (Fig. A.1b).

- There is some variability in SOM skill across the random seeds (e.g., Figs. A.4; A.5). However, these skill differences seem relatively small.

- For some random seeds, the SOM can get “stuck" and try to sort all cases into 3 nodes rather than 9. This behavior results in quantization errors that are larger than other random seeds

and topographic errors of zero (e.g., Figs. A.1; A.5). Such random seeds are not used for the SOM training.

- The impacts of rough and fine-tuning training radius on skill are not entirely clear, however it seems that keeping fine-tuning radius relatively small may result in smaller errors (Fig. A.4).

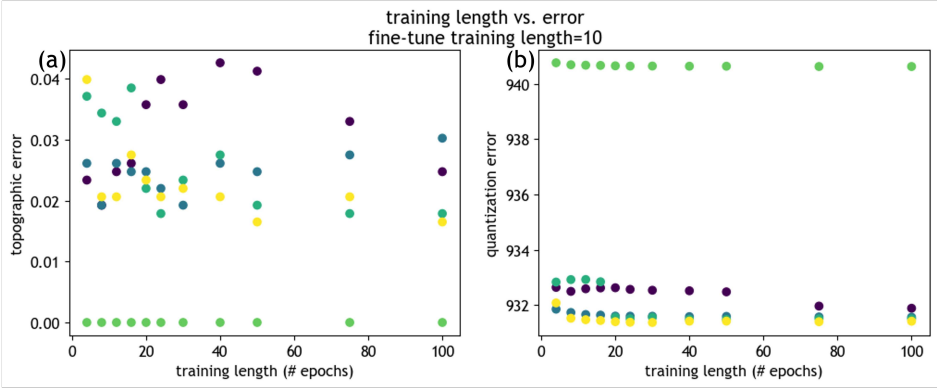

Figure A.1: (a) topographic error and (b) quantization error for assorted rough training lengths. Fine tuning training length is held constant at 10 epochs, and rough and fine-tuning training radii are 3 and 1 respectively. Colors correspond to assorted random seeds.

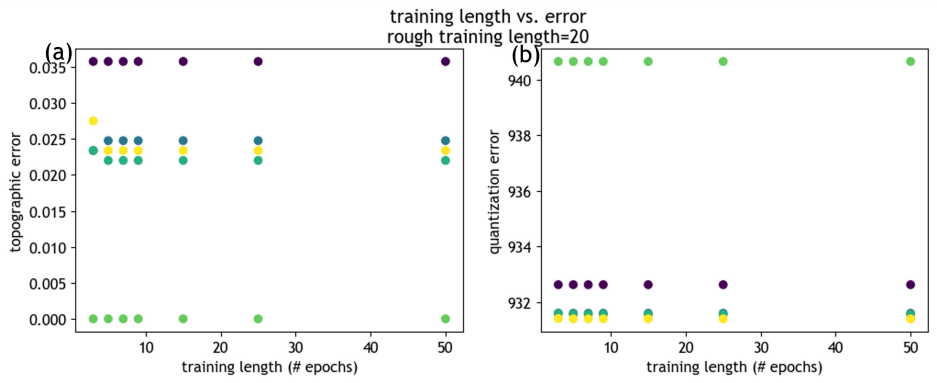

Figure A.2: As in Fig. A.1, but fine-tuning training length is varied, and rough training length is held constant at 20 epochs.

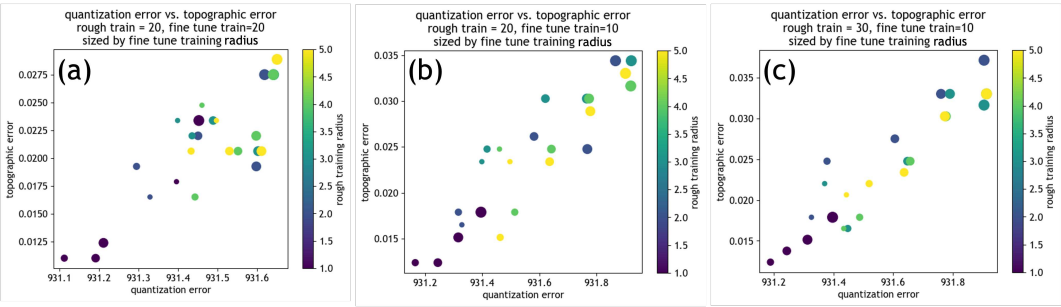

Figure A.3: Quantization error vs. topographic error for various rough and fine tuning training lengths and radii. Data points are colored by rough training radius and sized according to fine tuning training radius.

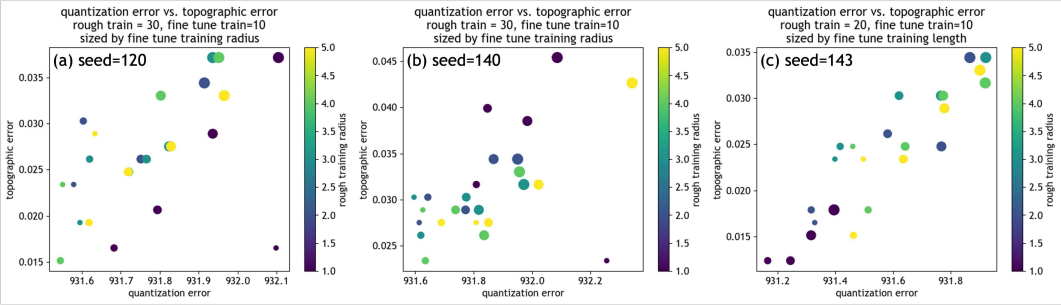

Figure A.4: Quantization vs. topographic error for various rough and fine-tuning radii across different random seeds. Data points are colored by rough training radius and sized according to fine tuning training radius.

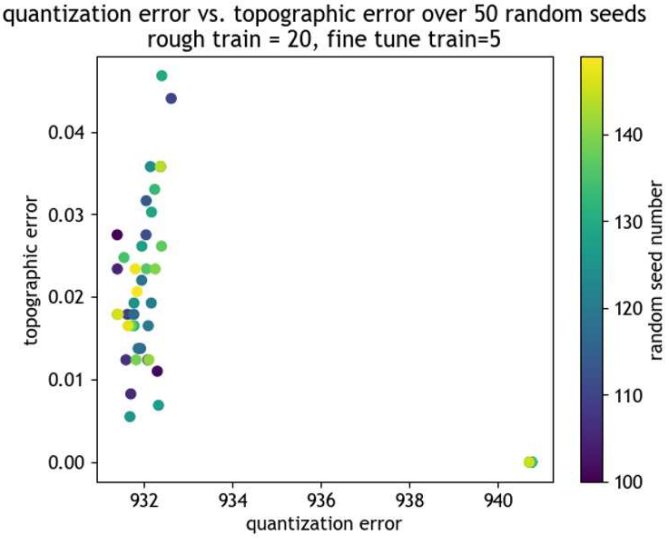

Figure A.5: Quantization error vs. topographic error for SOMs initialized over 50 random seeds. Rough and fine-tuning training radii are both set to 1, and training lengths are held constant at 20 ad 5 epochs respectively.

A.2 SOM regime characteristics

Additional characteristics of the trained SOM regimes presented in Chapter 3 are provided here. This includes environmental composites not included in the main text, as well as forecast frequency composites and skill-related assessments across the various nodes. Details are included in the captions.

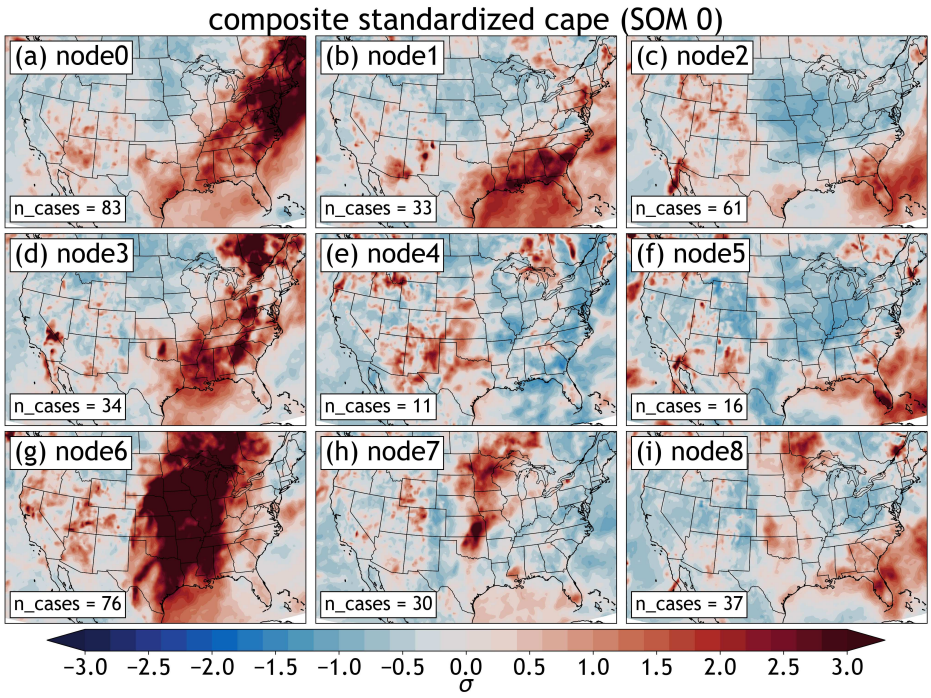

Figure A.6: Mean standardized daily surface-based CAPE anomalies, sorted by each node in SOM0. Node numbers and number of non-null forecast cases in each node are annotated in each panel.

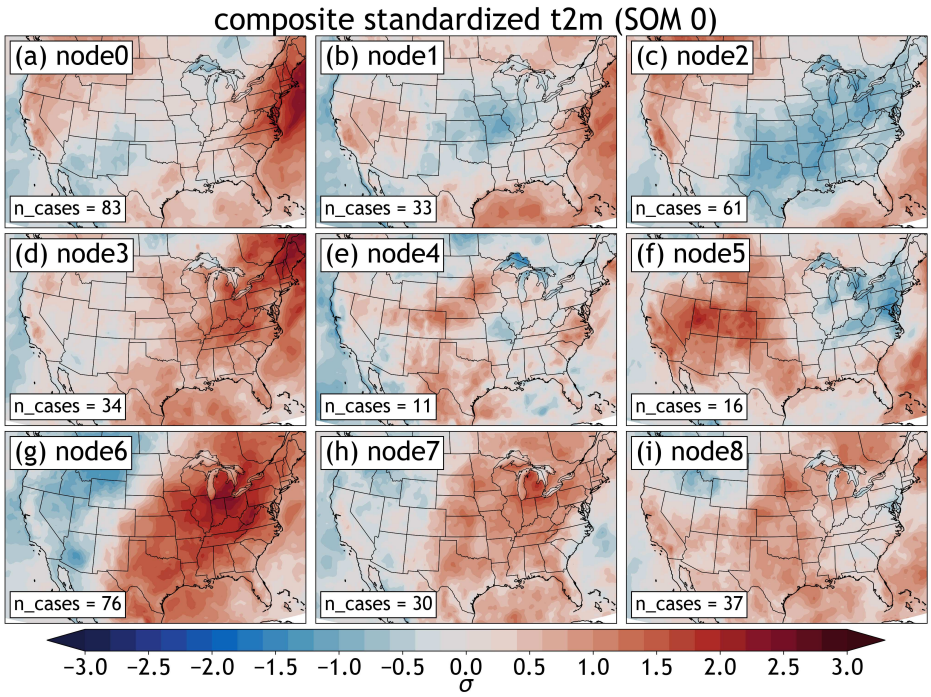

Figure A.7: As in Fig. A.6, but for 2-m temperature.

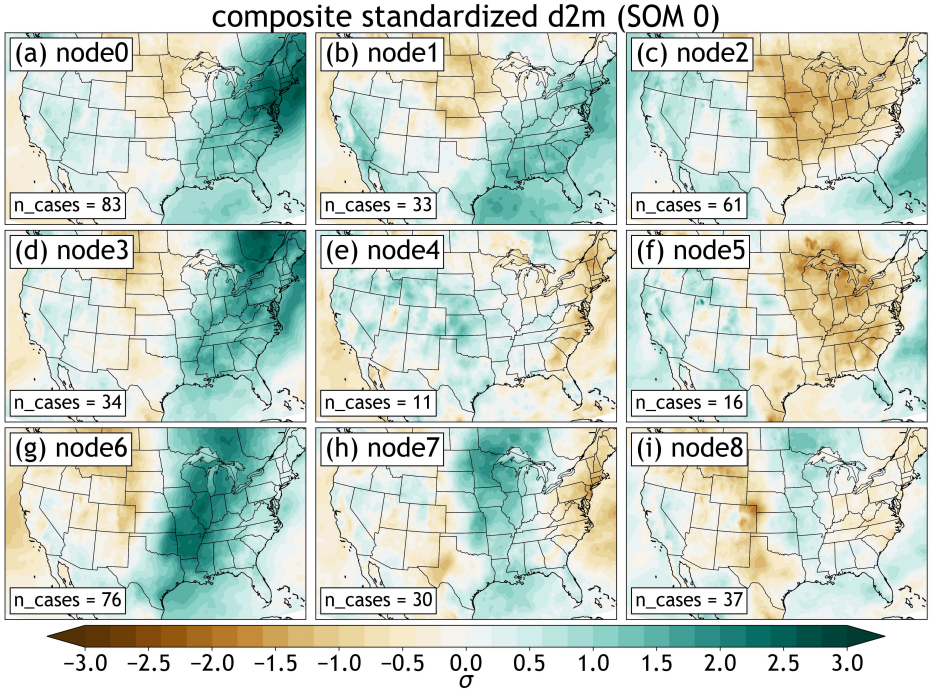

Figure A.8: As in Fig. A.6, but for 2-m dew point.

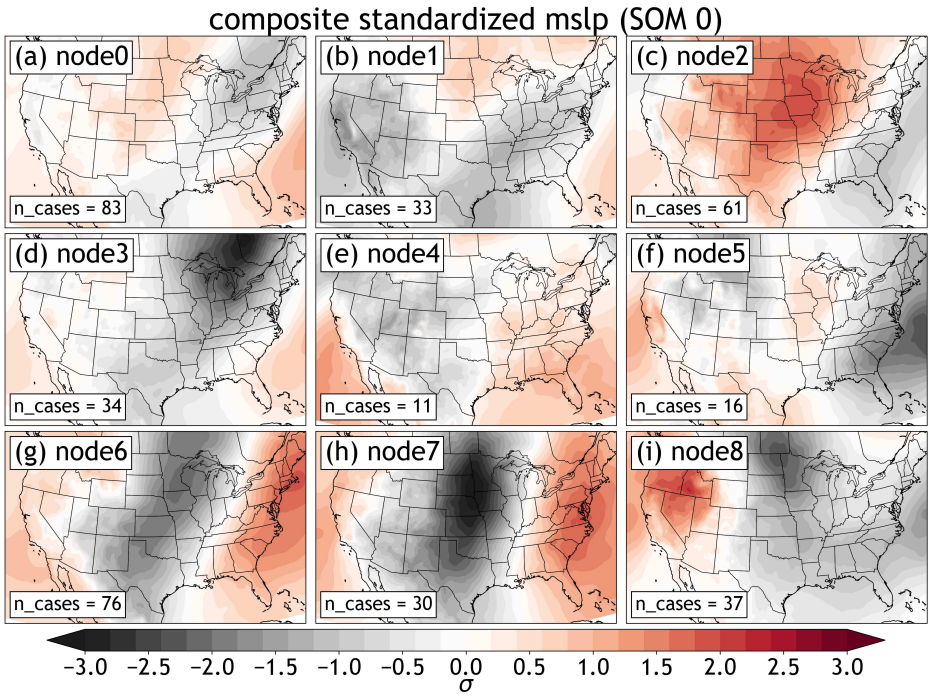

Figure A.9: As in Fig. A.6, but for mean sea level pressure.

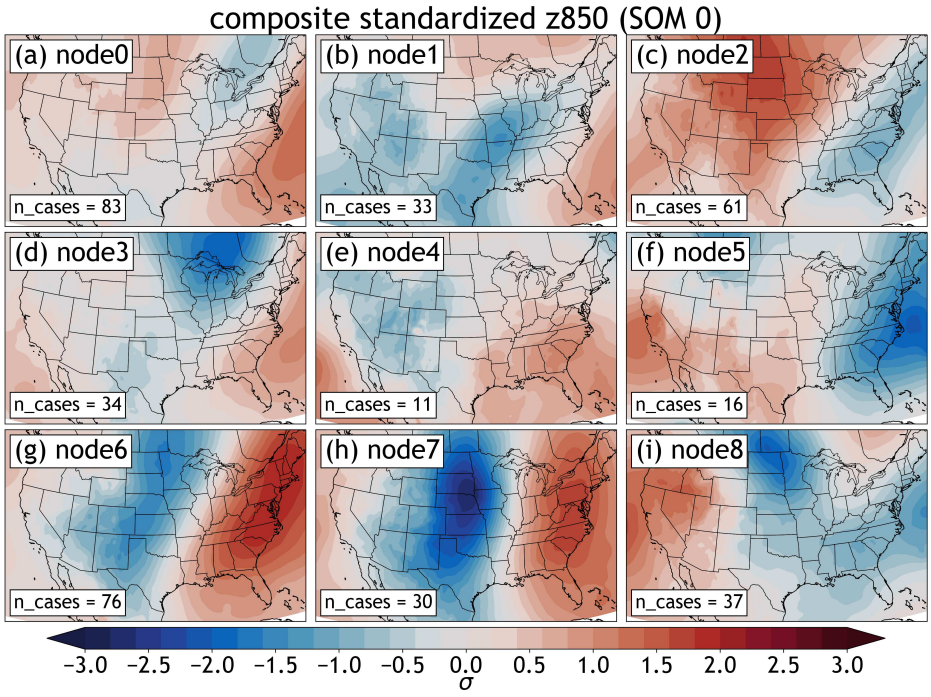

Figure A.10: As in Fig. A.6, but for 850 hPa geopotential heights.

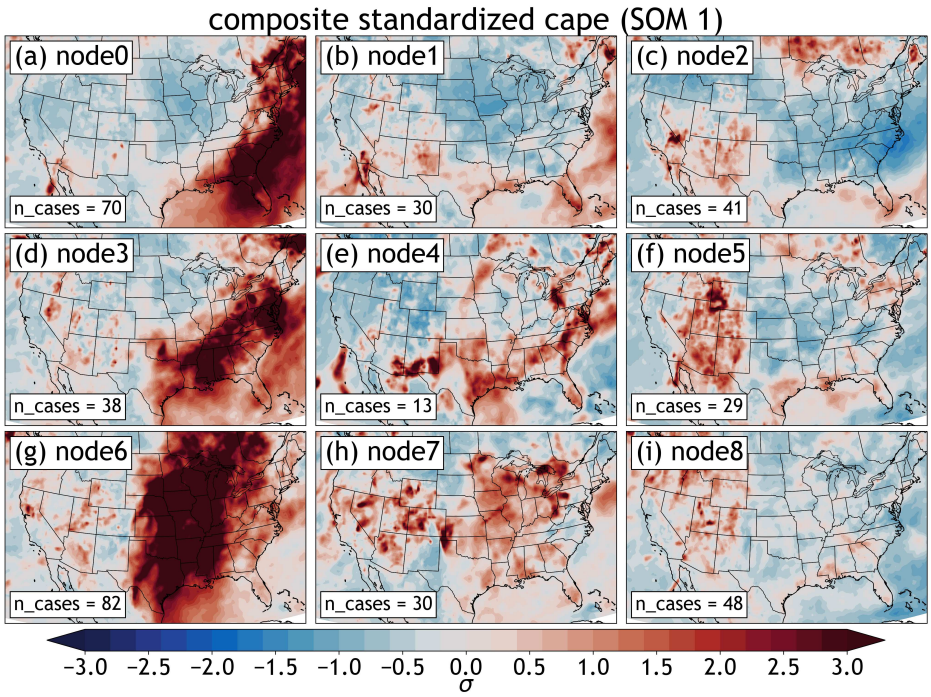

Figure A.11: Mean standardized daily surface-based CAPE anomalies, sorted by each node in SOM1. Node numbers and number of non-null forecast cases in each node are annotated in each panel.

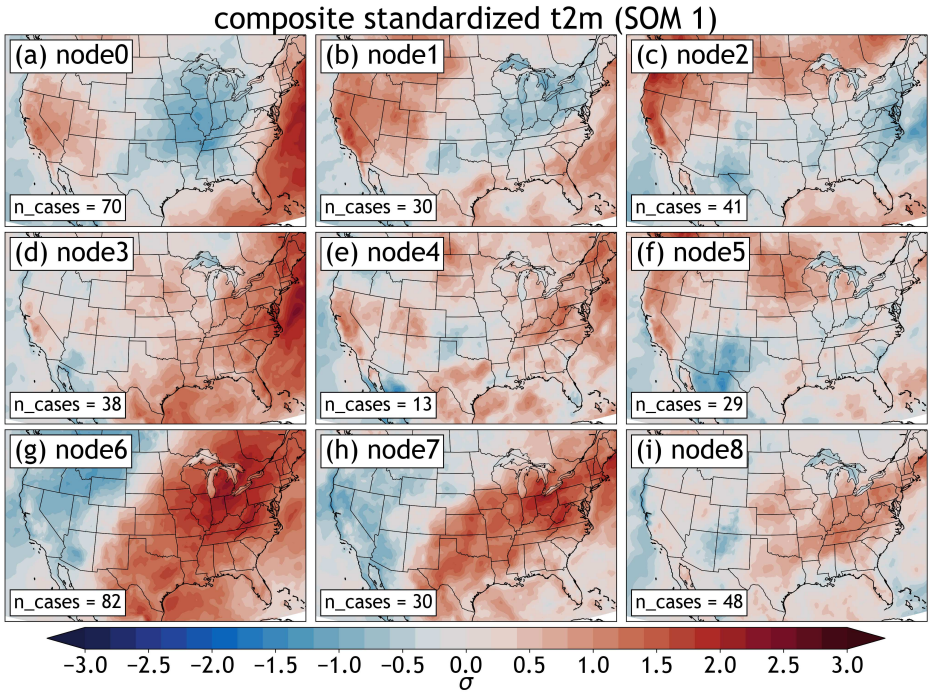

Figure A.12: As in Fig. A.11, but for 2-m temperature.

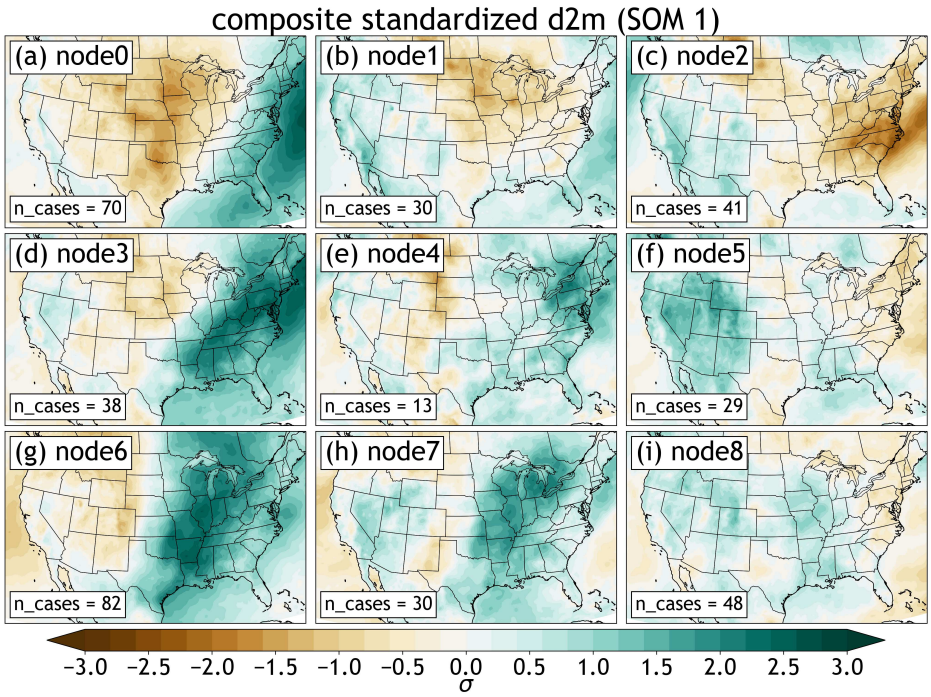

Figure A.13: As in Fig. A.11, but for 2-m dew point.

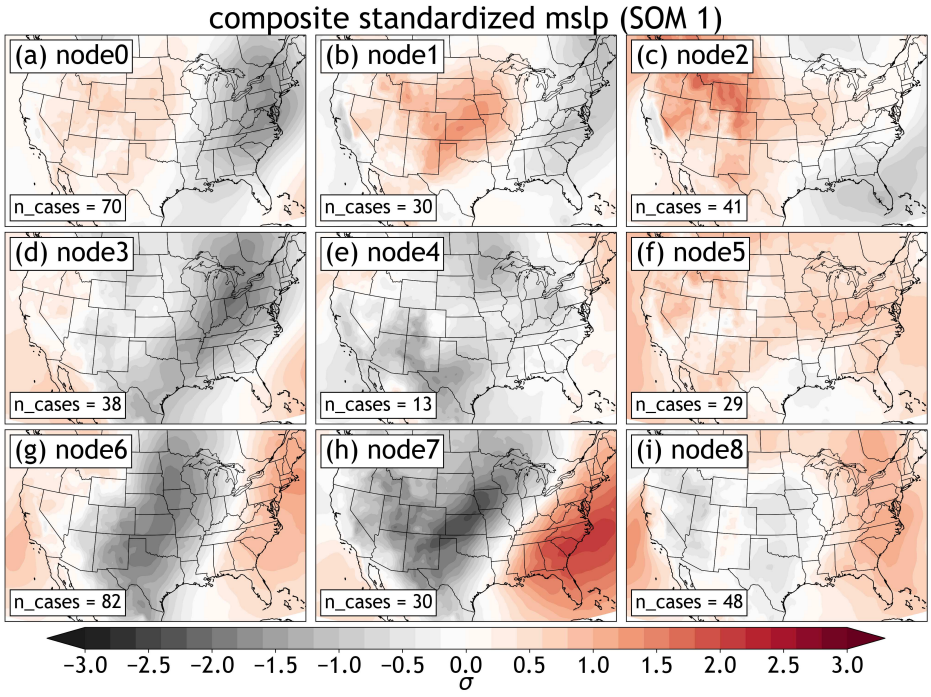

Figure A.14: As in Fig. A.11, but for mean sea level pressure.

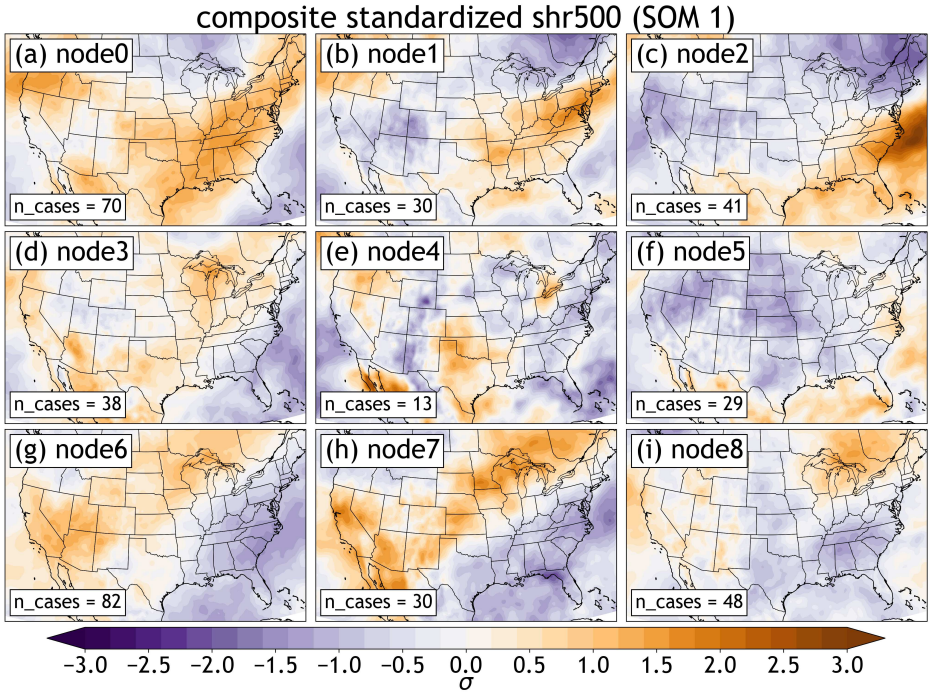

Figure A.15: As in Fig. A.11, but for 10-m to 500 hPa vertical wind shear.

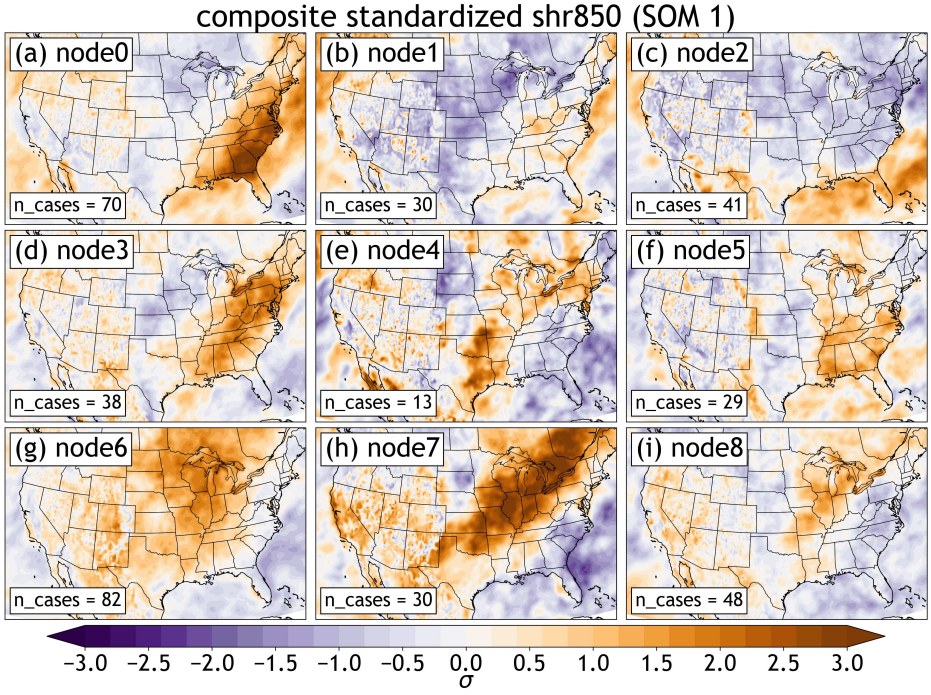

Figure A.16: As in Fig. A.11, but for 10-m to 850 hPa vertical wind shear.

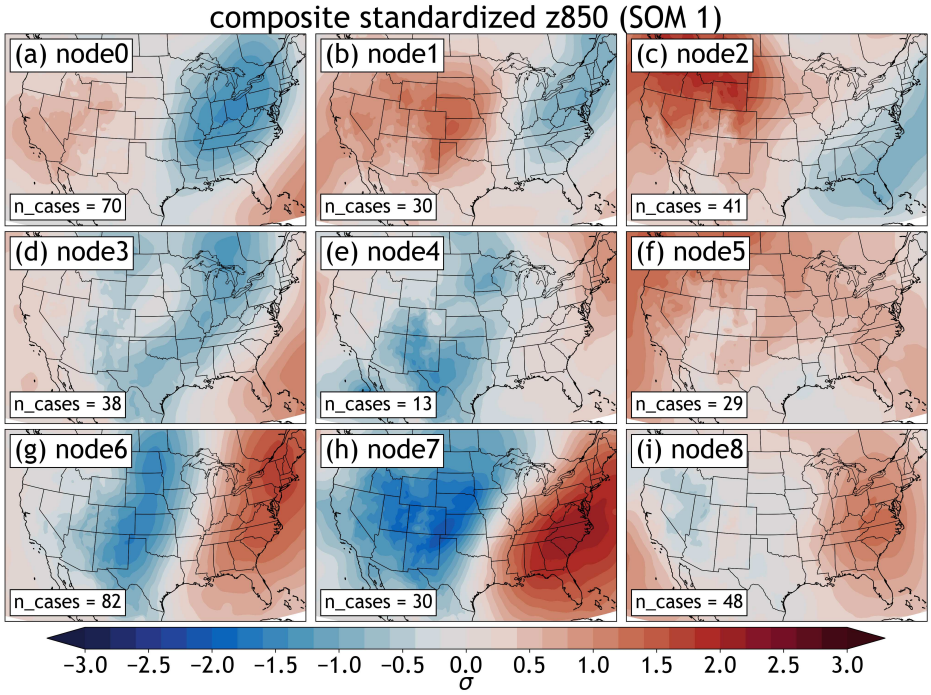

Figure A.17: As in Fig. A.11, but for 850 hPa geopotential heights.

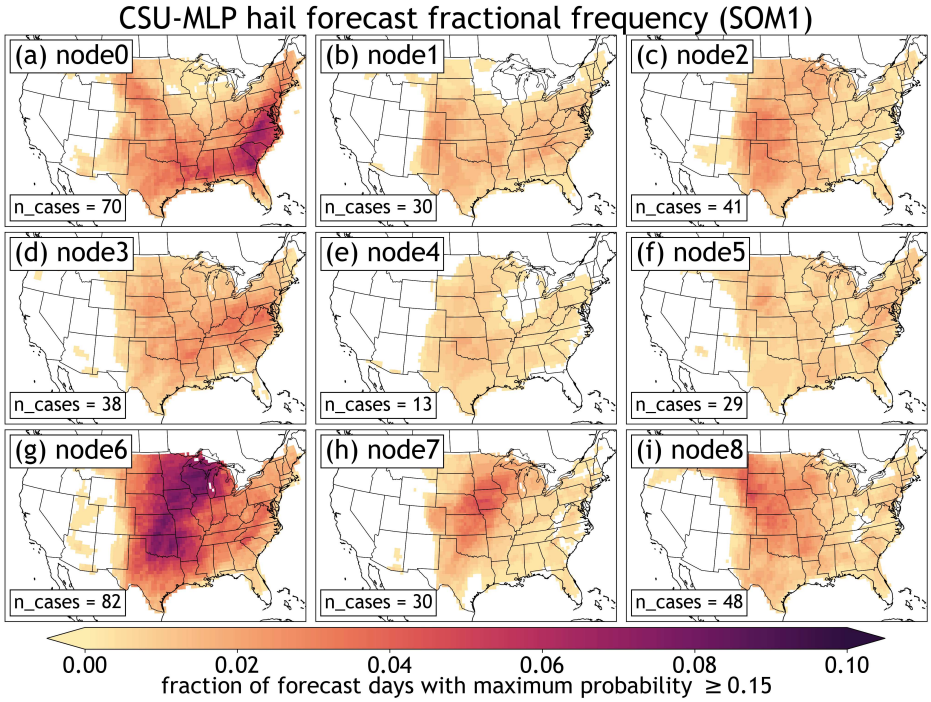

Figure A.18: As in Fig. 3.9 but for SOM1: fraction of non-null day-2 CSU-MLP hail forecasts out of total forecast days (381) at each node diagnosed at SOM1. A non-null forecast day is considered a forecast with a *maximum* hail probability of at least 15%; thus note that lower probabilities in the non-null cases are still considered here.

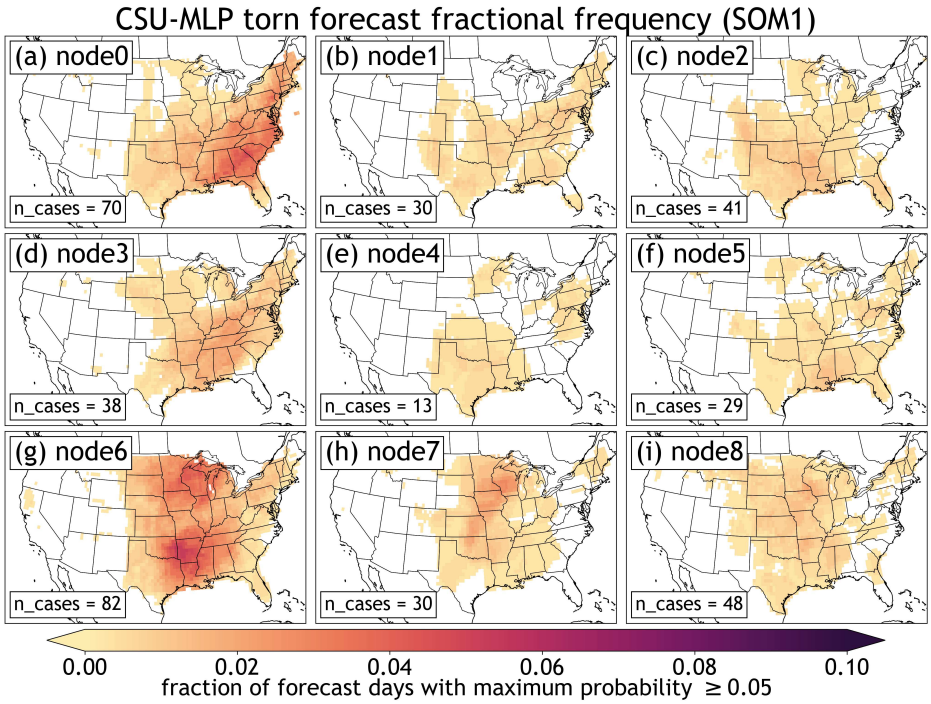

Figure A.19: As in Fig. A.18 but for day-2 CSU-MLP tornado forecasts. Note that the maximum daily probability must only exceed 5% to be considered a null case here.

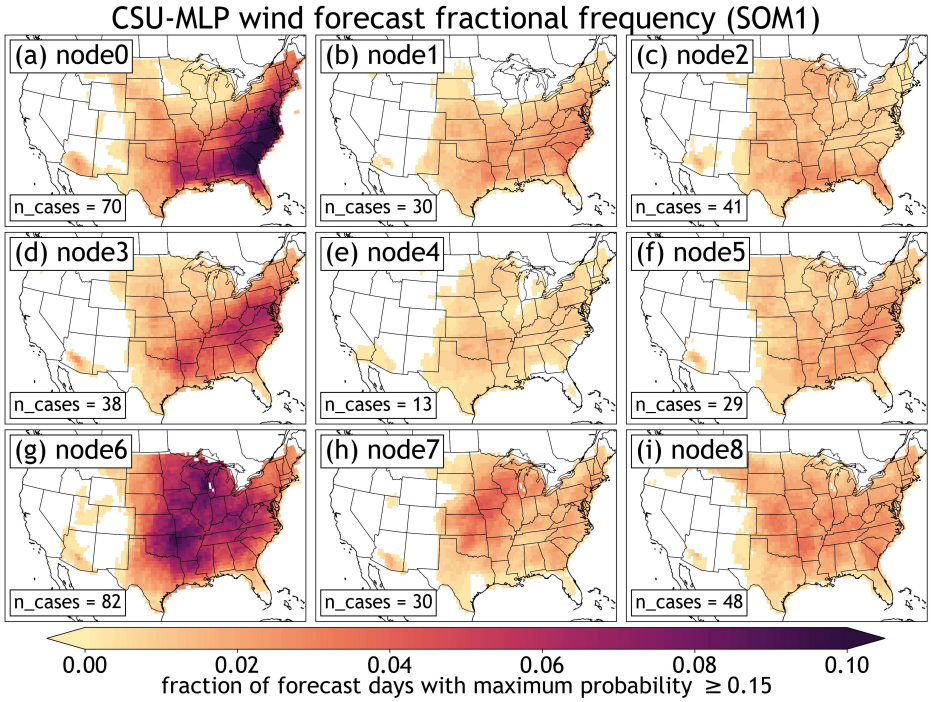

Figure A.20: As in Fig. A.18, but for wind forecasts.

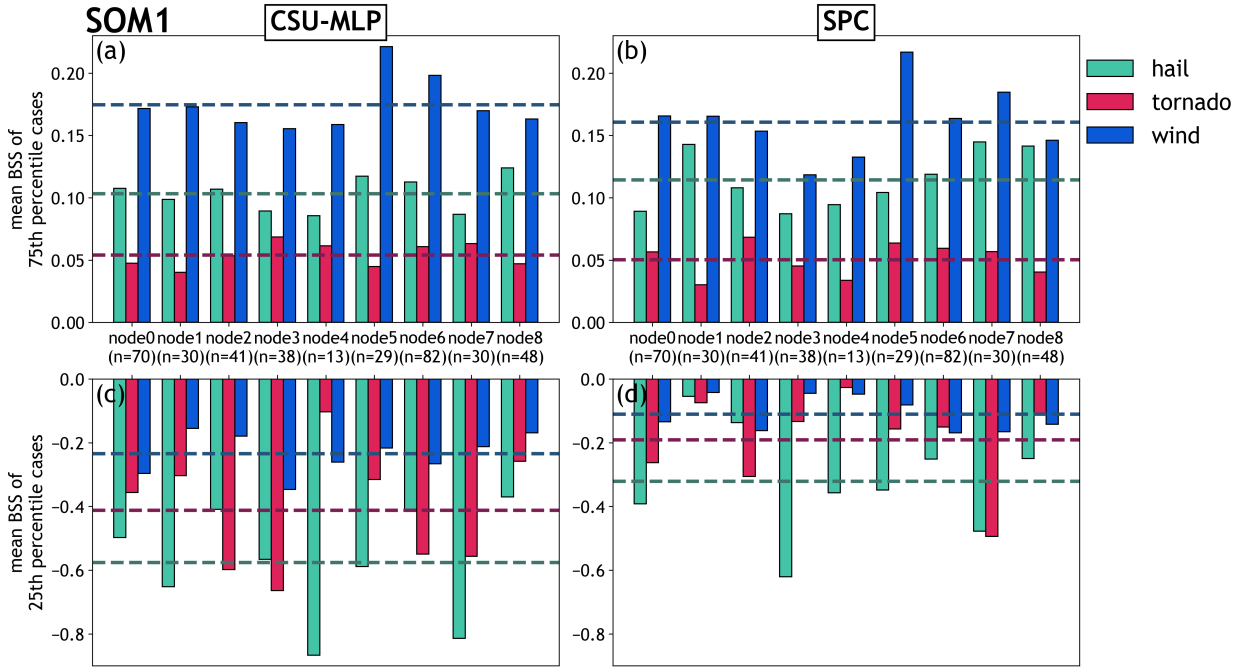

Figure A.21: As in Fig. 3.14, but for the SOM1 node configuration.

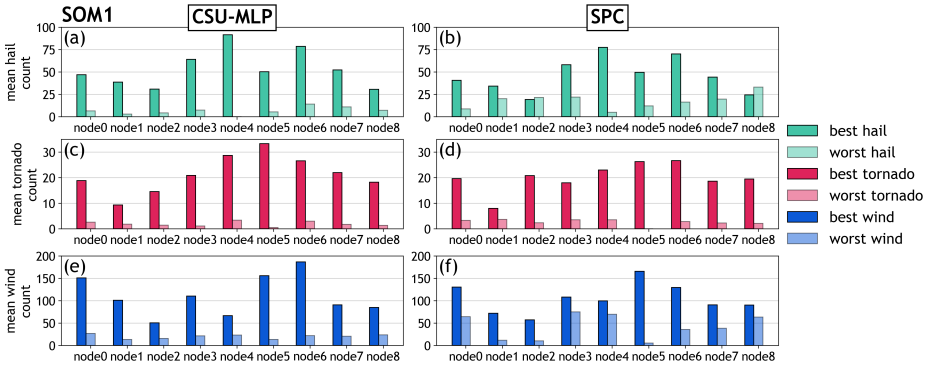

Figure A.22: Mean number of CSU-MLP grid points with at least one (a),(b) hail, (c), (d) tornado, or (e), (f) wind report among the best and worst CSU-MLP and SPC forecasts, separated by SOM1 nodes. Mean counts are shown for CSU-MLP forecasts in the left column and SPC forecasts in the right column.

Table A.1: Aggregate BSS for CSU-MLP and SPC forecasts in the 75th percentile ("best cases"). Aggregate scores are computed across each node for SOM0 and SOM1. The largest and second largest (i.e., best and second best) node-aggregated BSS are denoted by a double and single asterisk (respectively) for SOM0 and SOM1.

SOM	node	CSU 75%ile aggregate BSS hail	SPC 75%ile aggregate BSS hail	CSU 75%ile aggregate BSS torn	SPC 75%ile aggregate BSS torn	CSU 75%ile aggregate BSS wind	SPC 75%ile aggregate BSS wind
SOM0	node0	2.41**	1.86*	0.98*	1.00*	3.99*	2.30*
SOM0	node1	0.80	0.76	0.49	0.52	1.06	1.38
SOM0	node2	1.45*	1.59	0.52	0.41	1.52	1.31
SOM0	node3	0.87	0.80	0.55	0.63	1.62	1.87
SOM0	node4	0.54	0.32	0.02	0.05	0.00	0.14
SOM0	node5	0.67	0.86	0.03	0.10	0.83	1.06
SOM0	node6	1.20	1.96**	1.78**	1.59**	4.49**	3.77**
SOM0	node7	1.37	1.61	0.46	0.34	1.47	1.52
SOM0	node8	1.09	1.44	0.45	0.47	2.01	2.24
SOM1	node0	2.15**	1.43	0.81*	0.96*	4.12*	3.48*
SOM1	node1	0.99	1.00	0.28	0.21	0.69	0.50
SOM1	node2	1.18	1.84	0.43	0.34	1.92	2.00
SOM1	node3	0.45	0.44	0.62	0.36	1.40	0.59
SOM1	node4	0.34	0.38	0.18	0.13	0.48	0.40
SOM1	node5	0.82	0.31	0.27	0.45	1.55	1.30
SOM1	node6	2.14*	2.26*	1.77**	1.73**	4.16**	4.09**
SOM1	node7	0.35	0.72	0.51	0.57	1.36	1.48
SOM1	node8	1.98	2.83**	0.42	0.36	1.31	1.75

Table A.2: As in Table A.1, but for cases in the 25th percentile ("worst cases"). The smallest and second smallest (i.e., worst and second worst) node-aggregated BSS are denoted by a double and single asterisk (respectively) for SOM0 and SOM1.

SOM	node	CSU 25%ile aggregate BSS hail	SPC 25%ile aggregate BSS hail	CSU 25%ile aggregate BSS torn	SPC 25%ile aggregate BSS torn	CSU 25%ile aggregate BSS wind	SPC 25%ile aggregate BSS wind
SOM0	node0	-11.21*	-10.45**	-14.16**	-5.14*	-3.35	-2.09
SOM0	node1	-6.28	-3.41	-2.05	-0.80	-3.00	-0.50
SOM0	node2	-3.56	-0.97	-4.52	-2.17	-3.92*	-3.32*
SOM0	node3	-4.77	-3.07	-1.78	-0.95	-3.22	-1.53
SOM0	node4	-0.23	0.00	-0.57	-0.05	-0.20	-0.30
SOM0	node5	0.00	-0.17	-0.08	-0.35	-0.24	-0.15
SOM0	node6	-15.26**	-8.63*	-9.83*	-6.07**	-5.85**	-3.69**
SOM0	node7	-4.86	-1.36	-5.60	-1.03	-2.59	-0.73
SOM0	node8	-2.72	-3.79	-4.64	-3.07	-0.67	-0.40
SOM1	node0	-9.45*	-6.66	-8.20*	-6.04**	-4.74*	-2.82*
SOM1	node1	-1.96	-0.22	-1.52	-0.74	-0.62	-0.13
SOM1	node2	-2.46	-1.37	-4.79	-2.76	-2.16	-1.79
SOM1	node3	-8.50	-7.45*	-5.97	-1.07	-2.77	-0.45
SOM1	node4	-0.87	-0.72	-0.31	-0.05	-0.78	-0.19
SOM1	node5	-3.53	-2.09	-1.89	-0.78	-0.65	-0.16
SOM1	node6	-14.11**	-7.55**	-12.09**	-2.87	-6.39**	-3.38**
SOM1	node7	-6.52	-4.30	-6.13	-3.95*	-2.55	-1.65
SOM1	node8	-1.48	-1.50	-2.33	-1.34	-2.37	-2.14

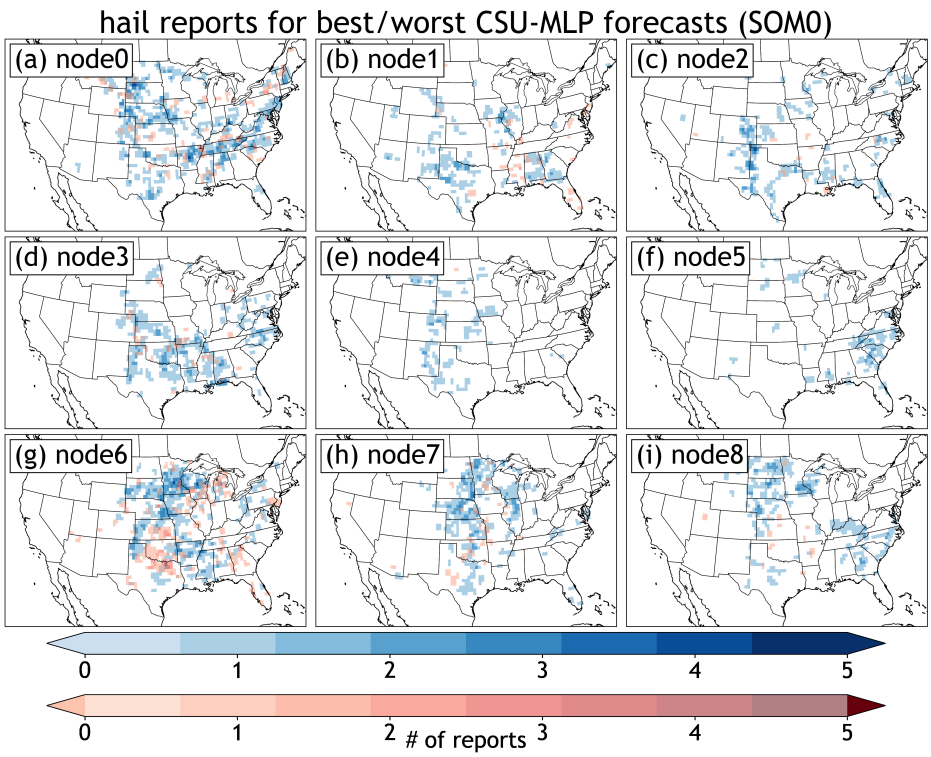

Figure A.23: Frequency plot of CSU-MLP grid points with at least one hail report over the 381 forecasts used in the study. Only reports that are associated with the 25% most-skilled CSU-MLP forecasts (blue) and 25% least-skilled forecasts (red) are shown. “Best" and “worst" thresholds are computed over all forecasts, but results are stratified by each SOM0 node.

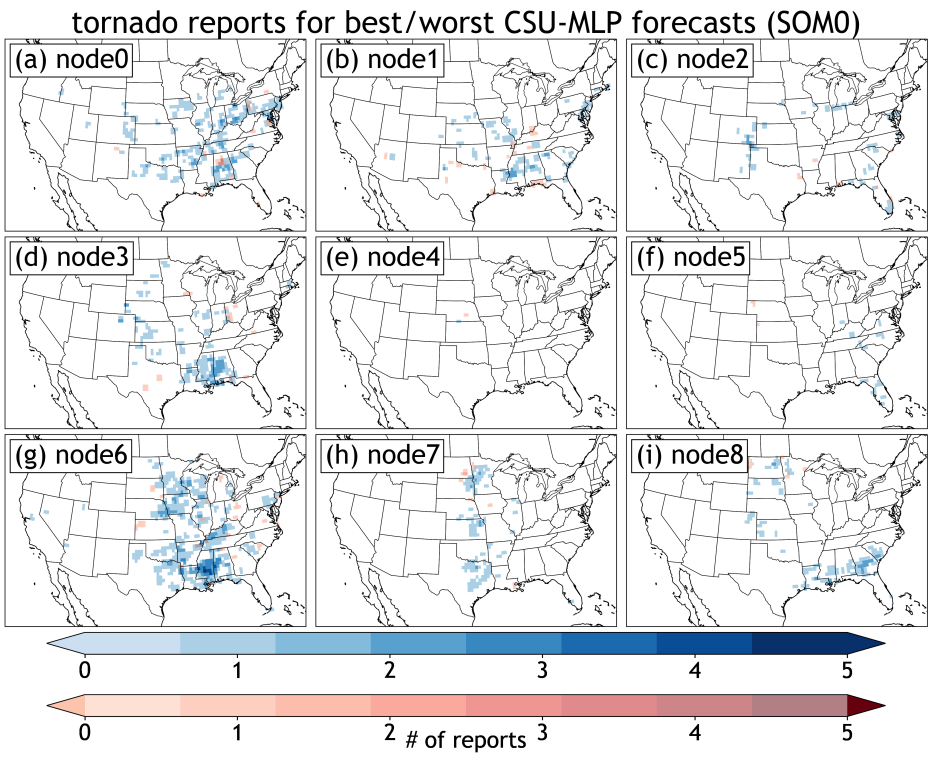

Figure A.24: As in Fig. A.23, but for tornado reports.

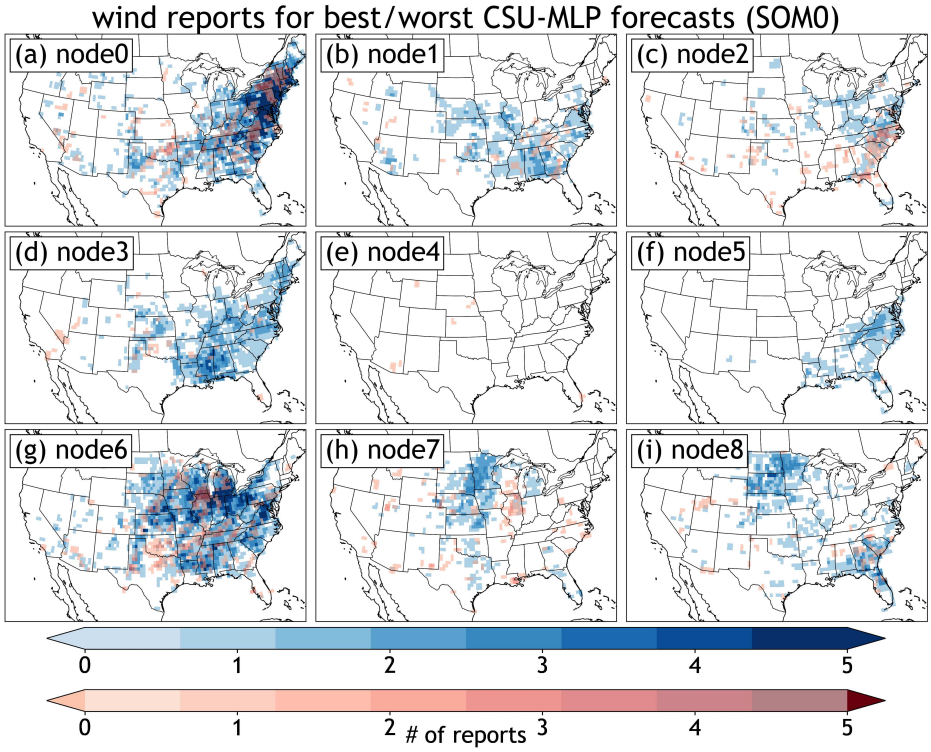

Figure A.25: As in Fig. A.23, but for wind reports.