UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

# COMPARING FORECAST TECHNIQUES FOR OVERLAPPING TORNADO AND FLASH FLOOD EVENTS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

by CHRISTIAN MCGINTY

Norman, Oklahoma

2026

COMPARING FORECAST TECHNIQUES FOR OVERLAPPING TORNADO AND FLASH FLOOD EVENTS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Aaron Hill (Chair)

Dr. Erik Nielsen

Dr. Jason Furtado

Dr. Harold Brooks

# Acknowledgments

First and foremost, I'd like to thank my advisor Aaron Hill for his incredible wisdom and support throughout my time here at OU. I could not have completed this thesis without his guidance and it has been such an honor to work with him. I'd also like to thank my committee members, Dr. Erik Nielsen, Dr. Jason Furtado, and Dr. Harold Brooks for their added support, guidance, and feedback which significantly improved this study. Special thanks to my hilarious and amazing office mates in 5310 who made every day special and were an incredible support system as I was writing this thesis. I also could not have done this without the support of my family who have been at my side since day one cheering me on.

# Table of Contents

# List of Tables

# List of Figures

# Abstract

Tornadoes and flash floods are considered two of the most hazardous weather phenomena in the United States. On rare occasions, these hazards can overlap in space and time, an event known as a TORFF (TORnado-Flash-Flood). These overlaps are especially dangerous because the rarity of combined events makes them difficult to forecast, and the protective guidance of overlapping warnings often contradict each other. Tornado warnings will advise people to seek shelter in a basement or low-level structure, contrasting flash flood warnings which often advise moving to higher ground. This divergence in guidance has proved deadly in multiple instances, highlighting the importance of ample warning time and specific protective guidance for such events. Numerous studies have shed light on the meteorological characteristics favorable for TORFF events, but little research has demonstrated their forecasting potential or if current operational forecast systems are capable of highlighting TORFF risk with sufficient calibration or spatial discrimination. The purpose of this study, therefore, is to develop a probabilistic TORFF forecasting paradigm that accurately highlights areas of dual-hazard risk to inform forecasters, emergency managers, and the public. Forecasters could use such a product to shift focus on areas where both hazards are possible to improve warning coordination, and emergency managers could use it to prepare additional resources for high-risk areas that may experience both hazards.

In order to forecast TORFFs within a probabilistic framework, the statistical relationship between the individual events must be studied to discern how to best model the joint probability. Very little is known about the statistical relationship between tornadoes and flash floods, especially whether they are to be treated as independent or dependent events. To address this knowledge gap, a linear regression method is utilized to model the correlation between tornado and flash flood events across the United States (US). Results do not indicate a robust correlation across the US, but a notable region of positive correlation was present from northern Mississippi into central Iowa, motivating the use of both independent and conditional probability to derive joint probabilities. As a result, multiple forecast methods are evaluated including independent joint probabilities derived from operational Storm Prediction Center (SPC) tornado outlooks and Weather Prediction Center (WPC) excessive rainfall outlooks, conditional probability methods that account for statistical relationships between tornadoes and flash floods, and finally a random forest machine learning model trained on Global Ensemble Forecast System reforecast data.

No significant difference in forecast skill is evident between regions with and without a significant tornado-flash flood correlation. Forecast methods which utilize both SPC and WPC outlooks provide the best spatial discrimination for TORFF events, but probabilities are not well calibrated. Methods that utilize conditional probability in conjunction with the SPC tornado outlook are much better calibrated, yet have worse spatial discrimination. The machine learning model is also not well calibrated but demonstrates enhanced spatial discrimination in test cases over other forecast methods, suggesting future work with machine learning is a promising avenue for forecasting TORFF events.

# Chapter 1

# Introduction

## 1.1 Background

Tornadoes and flash floods are two of the deadliest natural hazards that people face in the United States, with 30yr average death rates of 72 and 89, respectively (NOAA 2024). These extremely dangerous hazards are amplified when both occur over the same area at the same time, an event known as a TORFF (TORnado-Flash-Flood). Although quite rare, TORFFs present an especially life threatening situation for those caught in their path, as warning guidance for tornadoes and flash floods are often contradictory. For example, a tornado warning will advise people to seek shelter inside on the lowest possible floor, whereas a flash flood warning may advise people to seek shelter on high ground to escape rising flood waters. This difference in warning guidance can create an exceptional challenge for forecasters, broadcasters, and emergency managers (EMs) when they occur at the same time as they must decide which hazard is prioritized in communication with the public.

National Weather Service (NWS) forecasters and broadcast meteorologists often amplify the tornado threat over the flash flood threat when both are present, due to a general belief that tornadoes are more dangerous than flash floods (Henderson et al. 2020; Ernst et al. 2024). Amplifying one hazard over the other may result in a life-threatening situation if both or only the attenuated hazard verifies, leaving the public vulnerable to an unexpected weather event (Henderson et al. 2020). It has also been demonstrated that those who live in manufactured homes are at an enhanced risk for TORFFs as they are more likely to encounter flooded roadways when evacuating from a tornado since they cannot safely shelter in their homes (Henderson et al. 2020; First et al. 2022; Chaney and Weaver 2010). Situations such as these highlight the need to better understand the risk of dual-hazard weather events, as well as when and where they may occur.

The majority of studies on TORFF events have occurred within the past decade, and relatively few exist overall, indicating that TORFF events are still an emerging topic of study with much still to be understood about their meteorological characteristics. TORFF studies up to this point have largely focused on their climatology, storm mode, and supportive environments, but due to the lack of a universally accepted definition for a TORFF event, results vary considerably depending on the datasets and methods used.

## 1.2 TORFF Characteristics and Climatology

Long before the term 'TORFF' was coined, Maddox and Dietrich (1982) were the first to examine dual-hazard weather events, specifically flash flooding that occurred simultaneously with severe thunderstorms that produced tornadoes. They analyzed the 'meteorological aspects' of 11 such cases and found they were largely reminiscent of frontal and mesohigh type flash floods, but had considerably higher conditional instability and lower-tropospheric moisture thanks to strong southerly advection of warm moist air from the low-level jet (LLJ). They also noted that these events typically occurred within a relatively weak synoptic-scale setup with weak winds and vertical wind shear, but close to a mid-tropospheric ridge. Rogash and Smith (2000) analyzed a single case in eastern Arkansas/western Tennessee that produced a violent tornado outbreak with severe flash flooding. Similar to Maddox and Dietrich (1982), they found this event occurred in a mesohigh heavy rain pattern and had significantly high levels of lower-tropospheric moisture and instability. They also noted strong storm relative helicity (SRH) and strong synoptic-scale forcing from an upper-tropospheric jet streak that became collocated with forcing from surface boundaries. Rogash and Racy (2002) examined 31 cases where significant tornadoes occurred in proximity to flash flooding. They found cases occurred mostly in the afternoons of spring and summer and, similar to Rogash and Smith (2000), had high instability, high lower-tropospheric moisture, moderate SRH, and strong synoptic-scale vertical motion from a well-defined upper-tropospheric trough and jet streak.

Nielsen et al. (2015) was a pivotal study that revisited TORFFs and ended a significant drought in research after Rogash and Racy (2002). They examined the climatology and environments of TORFF events using two datasets: one consisting of intersecting tornado and flash flood warning polygons issued within 30 minutes of each other, and one consisting of clustered, verified tornado and flash flood storm reports. They estimate that warning intersections occur approximately 400 times a year, with a maximum frequency over the lower Mississippi River valley and during the spring and early summer. Using clustered storm reports, TORFF environments were found to have stronger low-level synoptic forcing, higher low-level moisture, and stronger vertical wind shear when compared with tornado-only environments. Despite this finding, they mention that TORFF environments generally have very similar characteristics to typical tornadic environments, making them difficult to distinguish with much confidence. Ellis et al. (2023) similarly investigated TORFF climatology with intersecting warnings, but only focusing on six NWS county warning areas (CWAs) in the southeastern United States where TORFFs occur most frequently. They found that TORFFs in this region were most likely to occur during the day and in spring, with a median intersecting warning duration between 23.5 and 36 minutes. They also showed that intersecting warnings where both hazards verified only occurred 15% of the time, and that the Memphis, TN CWA had the highest proportion of tornado warnings that occurred with flash flood warnings (51.3%).

Most recently, Burow et al. (2025) analyzed TORFF environments with the same intersecting warning dataset and regions as Ellis et al. (2023), but using a random forest machine learning algorithm trained on ECMWF Re-analysis (ERA5) (Hersbach et al. 2020) proximity sounding data to characterize environments associated with the 4 most common TORFF storm modes for subsequent classification predictions. Results showed that tropical cyclone (TC) storm modes were associated with warmer temperatures and higher moisture in the upper atmosphere, quasi-linear convective system (QLCS) modes had higher near-surface winds but lower deep-layer shear, discrete cell modes had lower upper-level moisture and greater deep-layer shear, and cluster modes had very similar characteristics to discrete mode but with slightly higher moisture. These results demonstrate that the environmental characteristics of TORFFs can vary considerably with different storm modes, further complicating the task of distinguishing them from non-TORFF environments.

In addition to the previously discussed results, Nielsen et al. (2015), Ellis et al. (2023), and Burow et al. (2025) all analyzed the distribution of TORFF events across storm modes through manual classification of radar data. Nielsen et al. (2015) found that the distribution was nearly uniform among all modes (QLCS, discrete, cluster, and TC), whereas both Ellis et al. (2023) and Burow et al. (2025) found that cluster modes were the most common. The difference in distributions is likely due to the regions analyzed; Ellis et al. (2023) and Burow et al. (2025) analyzed only the southeast US while Nielsen et al. (2015) analyzed the entire continental United States (CONUS).

This is also supported by related research showing there to be regional differences in the distribution of convective storm modes across the CONUS (Sobash et al. 2023; Ashley et al. 2019; Gallus et al. 2008). Furthermore, Nielsen et al. (2015) were only able to classify a relatively small sample of TORFF events (68), suggesting more research is necessary to fully characterize the distribution of TORFF storm mode across the broader CONUS.

In summary, previous studies agree on various synoptic-scale conditions that favor TORFF environments including strong forcing for ascent, strong to moderate wind shear, and significant warm, moist air advection in the lower-troposphere. Despite these findings, there remains difficulty in distinguishing TORFF environments from typical tornadic environments that have these characteristics yet don't produce flash flooding (Nielsen et al. 2015; Burow et al. 2025). Climatologically, there is agreement that TORFFs occur most often in spring, in the southeast US, and during the afternoon, however this is also reminiscent of typical tornado climatology (Brooks et al. 2003).

## 1.3 Research Motivation

While previous studies have agreed on numerous synoptic-scale environmental characteristics indicative of TORFF events, there is still a fair amount of uncertainty due to differences in methodology and the extreme rarity of the events which makes it difficult to generate effective sample sizes. This uncertainty is especially notable at the storm-scale where there is much less agreement on the distribution of storm mode, and the environmental conditions specific to certain modes have not been sufficiently studied for TORFFs specifically. Furthermore, many parallels are observed between TORFF and typical tornadic environments in terms of climatology and large-scale meteorological conditions, making it difficult to characterize TORFFs as belonging to a unique class of convective environments. Moreover, the true death rate of TORFFs is unknown since tornado and flash flood deaths are categorized separately (P dos Santos 2016), making it extremely difficult to demonstrate the significance of TORFF events and their impact on the public. As a result of these sizable challenges, research on tornado environments and forecasting is heavily prioritized over such research for TORFFs, so a forecasting paradigm for TORFF events has yet to be established that helps communicate risk with sufficient lead time. Given the incredible risk to life and property that TORFF events pose and the need for specific guidance ahead of time to prepare the public for potentially conflicting warning guidance, a forecast that is specific to their occurrence is necessary to improve public safety. Such a forecast would help forecasters and EMs better understand the risk of TORFFs in their area to improve hazard communication and provide additional lead time and resources for necessary protective actions. Furthermore, a forecast for convective dual-hazard weather events has never been demonstrated in literature, so a TORFF forecast would be an excellent initial demonstration of the potential for dual-hazard forecasting in complex convective environments.

To address the gap in knowledge specific to TORFF forecasting, this study aims to develop a framework that demonstrates the predictability of TORFFs with various joint forecasting methods. The joint forecasts developed in this work are designed to be probabilistic and utilized at the outlook-scale to effectively highlight where synoptic-scale conditions are favorable for TORFFs. An outlook-scale forecast is currently the most effective method to highlight TORFF risk given that the synoptic-scale characteristics of TORFFs are generally more understood than the storm-scale, and multi-day lead time is necessary for forecasters and EMs to prepare additional resources for dual-hazard weather days. Communicating TORFF risk in a probabilistic manner is also necessary to convey the uncertainty in their environments which benefits users who rely on probabilistic thresholds to make decisions. Day 1 TORFF forecasts will be the focus of this work to develop the initial framework and to emphasize differences in forecast method performance across a single lead time.

## 1.4 TORFF Forecast Framework

To establish a probabilistic TORFF forecasting framework, the methods and performance of operational forecasts for tornadoes and flash floods are explored to consider how they can be leveraged in developing joint forecast products. Currently, operational probabilistic forecasts for tornadoes and flash floods are handled separately by the Storm Prediction Center (SPC) and Weather Prediction Center (WPC), respectively.

### 1.4.1 SPC Tornado Forecast

The SPC is a national forecast office branched from the National Weather Service (NWS) responsible for issuing forecasts and watches for severe thunderstorms, tornadoes, and fire weather across the CONUS (NOAA 2026b). SPC severe weather forecasts are communicated through their Convective Outlook – a collection of probabilistic forecast graphics that illustrate areas of severe weather risk out to Day 8. Day 1 and Day 2 Convective Outlooks can be separated into individual outlooks for tornadoes, hail, and wind, but only the tornado outlook (TO) is considered in this work. The SPC TO is a probabilistic forecast product that depicts the probability of a tornado occurring within 25 miles of a point. SPC forecasters utilize a blend of model data, surface observations, soundings, and satellite imagery to determine and categorize tornado risk. Tornado probabilities are discretized into bins of $\geq 2\%$, $\geq 5\%$, $\geq 10\%$, $\geq 15\%$, $\geq 30\%$, $\geq 45\%$, and $\geq 60\%$ (Figure 1.1) which have varying correspondence to categorical forecast areas. SPC forecasters may also overlay hatching that indicates the severity of forecasted hazards relative to climatology, but that information is not relevant to this study.

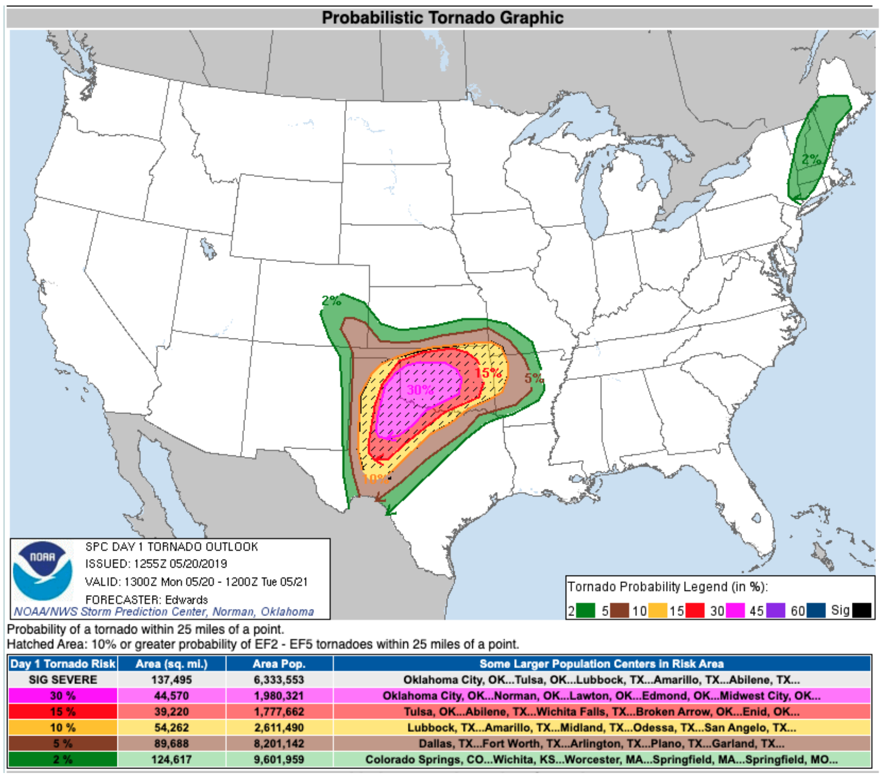


Figure 1.1: Day 1 SPC Tornado Outlook 1300 UTC 20 May, 2019.

Investigations into the performance of the SPC TO have demonstrated that it tends to under-forecast compared to forecasts for wind and hail (Herman et al. 2018), but has better calibration during significant tornado outbreaks and when synoptic-scale forcing is greater (Hitchens and Brooks 2017; Mercer and Bates 2019). These results indicate that the performance of the SPC TO is sensitive to the rarity of tornadoes, but is still a skillful forecast relative to climatology, especially at short lead times (Herman et al. 2018).

### 1.4.2 WPC Excessive Rainfall Forecast

The WPC is a national forecast office branched from the NWS responsible for short- to medium-range forecasts of synoptic-scale weather features, winter weather, and a particular emphasis on quantitative and probabilistic precipitation forecasting (NOAA 2026a). Under this umbrella, WPC is responsible for forecasting nationwide flash flood risk which they convey via the Excessive Rainfall Outlook (ERO). This outlook provides a probabilistic forecast of flash flood guidance (FFG) *exceedance* within 25 miles of a point, utilizing FFG from NWS River Forecast Centers (RFC), model data, environmental conditions, and local NWS office input to determine and categorize risk out to Day 5. As of February 2022[1], WPC discretizes exceedance probabilities into bins of $\geq$ 5%, $\geq$ 15%, $\geq$ 40%, and $\geq$ 70% for marginal, slight, moderate, and high risks, respectively (Figure 1.2).

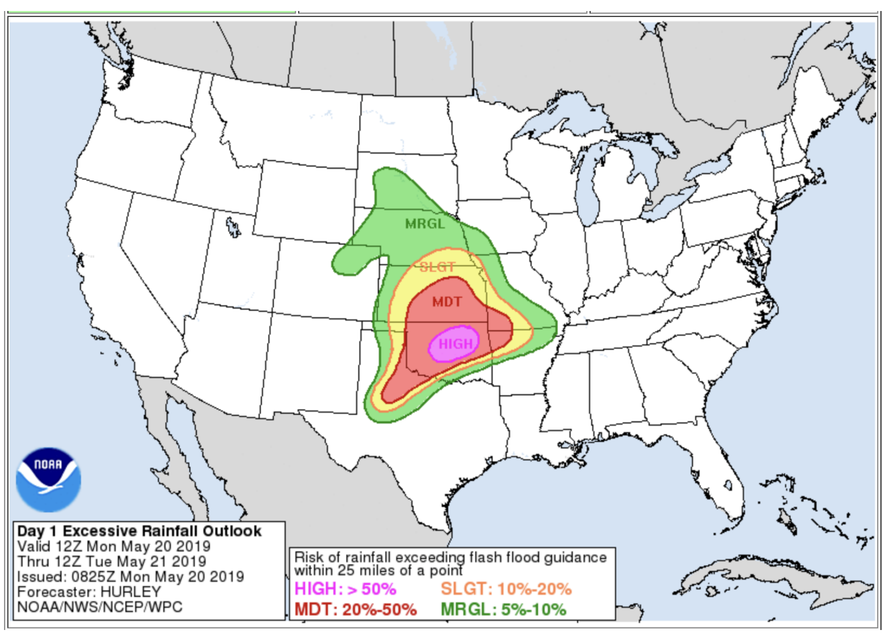


Figure 1.2: Day 1 WPC Excessive Rainfall Outlook 0830 UTC 20 May, 2019.

[1]Prior to February 2022, bins were defined as 5%-10%, 10%-20%, 20%-50%, and $\geq$ 50%. WPC retroactively applied the 2022 changes to *all* existing ERO files in the archive

Investigations into the performance of the WPC ERO have also shown that the forecast is skillful relative to climatology and well calibrated to FFG exceedance (Erickson et al. 2019, 2021). Similar to the SPC TO, these studies found that the ERO was best at shorter lead times and for storm systems that are more organized with stronger synoptic-scale forcing. Although the WPC ERO does not explicitly forecast for flash flood occurrences, research has shown that FFG has the best correspondence with flash flood reports and demonstrates considerable value as a threshold for the onset of flash flooding (James and Schumacher 2024). Furthermore, the ERO is said to be largely calibrated to the frequency and spatial density of flash flood occurrences (Burke et al. 2022), making it a suitable and skillful forecast for flash floods.

### 1.4.3 Joint Outlook Forecast

Given the positive skill of the SPC TO and WPC ERO relative to climatology and adequate calibration to tornadoes and flash floods respectively, there is considerable potential in a joint forecast for TORFFs that combines the probabilities of the two operational products together. It is also very beneficial that both forecasts use the same 25 mile event radius for calibration, enabling a joint forecast to be created without any changes to probabilistic calibration. Furthermore, little to no communication occurs between the SPC and WPC offices regarding these forecasts (Henderson et al. 2020), so a joint forecast could help bridge the gap in paradigms and improve dual-hazard communication. Therefore, one of the main goals of this study will be to assess the performance of combining the SPC and WPC outlooks to determine if these existing forecast products are sufficient in highlighting the probability and risk area of TORFFs. Notably, this forecast method assumes that tornadoes and flash floods are independent events such that their respective outlooks probabilities can be multiped together. As such, this forecast product is henceforth referred to as the independent joint outlook

(IJO) forecast.

### 1.4.4 Independence and Conditional Probability

A major hurdle in creating a joint forecast for TORFF events is that the statistical relationship between tornadoes and flash floods is not well understood. As previously mentioned, combining SPC and WPC outlooks outright would assume that the probabilities of the two hazards are independent of one another, but this has never been specifically addressed in previous literature. Given that TORFF environments are often hard to distinguish from those conducive to a single hazard (Nielsen et al. 2015; Maddox and Dietrich 1982), and both hazards are shown to occur within similar environments and storm modes (Burow et al. 2021; Nielsen et al. 2015; Nielsen and Schumacher 2020; Jiang et al. 2025), it's reasonable to hypothesize that this assumption of independence does not hold true in most cases, and that there is likely a positive correlation between tornado and flash flood occurrences. To test this hypothesis, the statistical relationship between tornadoes and flash floods is explored using a linear regression model to determine if a significant correlation exists between their occurrences. This analysis is divided into regions defined by the NWS CWAs in the CONUS to asses any spatial patterns in correlation and to demonstrate if correlation is robust CONUS-wide.

To account for the hypothesized case of non-independence, joint forecasts that utilize the *conditional* probability between tornadoes and flash floods are explored as alternatives to the IJO forecast. Conditional probability assumes dependence between variables such that the joint probability is the probability of one event occurring given that the other has occurred. Spatial distributions of conditional probability (i.e., the probability of a tornado given a flash flood at a point and vice versa) are generated for the CONUS using warnings and outlooks for each hazard to represent occurrences.

These spatial distributions are then used in conjunction with the respective SPC or WPC outlook to generate a conditional joint forecast for TORFFs. It is hypothesized that conditional joint forecast methods will perform better in any CWA regions found to have a significant tornado-flash flood correlation as given by the linear regression model, and that the IJO forecast will perform better elsewhere. Additionally, it's hypothesized that the spatial distributions of conditional probability will exhibit patterns that match the results of the linear regression analysis. Specifically, locally higher conditional probability of either hazard should be seen in CWA regions with a higher/positive tornado-flash flood correlation, and the opposite for CWA regions with a lower/negative correlation. Furthermore, its expected that there will be locally higher conditional probability in CWA regions with a higher frequency of TORFF events.

### 1.4.5 Random Forest Machine Learning

Machine learning methods have been shown to add considerable value to probabilistic forecasts of severe weather (Hill et al. 2020; McGovern et al. 2017). Given the uncertainty in atmospheric conditions favorable for TORFFs, there is strong potential in a statistical approach to forecasting as an alternative to the joint forecast methods. Previous studies have shown particular success with random forest machine learning (RFML) models for both severe weather (tornado, hail, wind) (Hill et al. 2020, 2023; White and Hill 2026) and excessive rainfall (Herman and Schumacher 2018a; Schumacher et al. 2021; Loken et al. 2019; Hill and Schumacher 2021), the majority of which have seen improvements in calibration and spatial discrimination over operational forecasts for the respective hazards. A machine learning model in this case would eliminate the need for conditional probability, would not be sensitive to the performance of operational forecasts, and be specifically calibrated to the frequency of TORFF events. An RFML model is designed and tested in this study to asses the capabilities of machine learning for rare/dual-hazard events and to determine if it is capable of highlighting TORFF risk with sufficient calibration and spatial discrimination. The hypothesis is that the RFML model will have improved spatial discrimination in test cases when compared to the joint forecasts, and will also have sufficient calibration to TORFF frequency.

## 1.5 Research Questions (RQs)

Based on the goals of the TORFF forecast framework and respective hypotheses, five main research questions will be addressed in this study to address the gaps in research and establish a TORFF forecasting framework:

1. Can tornadoes and flash floods be treated as independent events? In which CWA regions is there a significant correlation between these hazards?

2. What is the conditional probability of tornado and flash flood occurrences as estimated by warnings and outlooks? Do patterns in the spatial distributions align with the results of the linear regression and/or TORFF frequency?

3. Is there a significant difference in the performance of joint forecast methods in CWA regions with and without a significant tornado-flash flood correlation?

4. How well do different joint forecast methods perform when compared with each other?

5. Is an RFML model capable of highlighting TORFF risk with sufficient calibration? Does this model demonstrate improved spatial discrimination compared to the joint forecast methods across test cases?

# Chapter 2

# Data and Methods

## 2.1 Data

### 2.1.1 Tornado and Flash Flood Warnings

To answer the first and second research questions regarding the statistical relationship and conditional probability between tornadoes and flash floods, NWS warning polygons for each hazard are utilized as one way to represent their occurrences. Warnings were chosen over local storm reports (LSR) due to the relative consistency and lack of population bias that warnings offer, both of which are major issues for spotter-based datasets (P dos Santos 2016). Warnings are also sufficient to indicate where conditions are favorable for the respective hazard even if they do not verify, providing considerably more data points to determine the broader statistical relationships. Archived tornado and flash flood warning polygons from January 2008 to December 2024 were obtained from the Iowa State Environmental Mesonet NWS Watch, Warning, and Advisory archive page for the CONUS (Herzman 2024). The 2008 start date was chosen as it is the first full year that NWS utilized storm-based warnings (NOAA 2007). For the linear regression model and calculations of conditional probability, the coordinates of each polygon's centroid are used to represent the location of a tornado or flash flood event to simplify computation.

### 2.1.2 SPC and WPC Outlooks

SPC and WPC outlooks are used in this study to develop joint forecast methods utilizing both independent and conditional probabilities. In addition, outlooks are used as an alternative to warnings to estimate conditional probabilities.

Day 1 WPC EROs are issued at 0100 UTC, 0830 UTC, and 1600 UTC, however, only the 0830 UTC product is valid for a full 24 hour period (12 UTC to 12 UTC). Therefore, the 0830 UTC outlook is chosen to represent the Day 1 flash flood forecast in this study. The WPC archive of EROs is limited to August 2016 and beyond, so Day 1 outlooks are retrieved from January 2017 to December 2024 for a complete 8 year period. EROs are archived as netCDF files using a 276x721 latitude-longitude forecast grid with a grid spacing of 0.09-degrees. For the purposes of this study, the resolution of ERO forecast grids is unnecessarily high and greatly increases computation time, so outlooks are re-projected to a grid with 0.5 degree grid spacing which is approximately equal to 55 km horizontal spacing at the equator.

Day 1 SPC TOs are issued daily at 0600 UTC, 1300 UTC, 1630 UTC, 2000 UTC, and 0100 UTC, but only the 1300 UTC outlook is considered for this study which is valid from 1300 UTC to 1200 UTC. Although this outlook is not valid for a full 24 hour period, it is superior to the 0600 UTC outlook as it utilizes critical atmospheric data provided by the nationwide 1200 UTC upper-air soundings and forecast guidance from 1200 UTC numerical weather model runs (Sherburn et al. 2019). To ensure forecast continuity for subsequent verification, the 1300 UTC outlook is assumed valid for the previous hour in this study (valid 1200 UTC to 1200 UTC), allowing the forecast to represent a full 24 hour period. To align with the WPC ERO, Day 1 SPC TOs are also obtained from January 2017 to December 2024. SPC TOs are archived in ESRI Shapefile format which preserves forecast areas as geometric polygons. Polygons for each SPC TO probability contour are translated to the 55km grid by assigning probabilities to grid points that fall within the polygon's bounding coordinates.

### 2.1.3 Intersecting Warnings

Intersecting tornado and flash flood warnings are used to represent observed TORFF events for forecast evaluation, targets for training the RFML model, and to generate basic climatological statistics. Similar to the reasoning provided for the use of tornado and flash flood warnings, intersecting warnings provide significantly more data points than LSRs and indicate where conditions are favorable for TORFFs even if the warnings do not verify.

For the purposes of this study, a tornado and flash flood warning are said to be intersecting if they intersect spatially and temporally such that there is a non-zero time overlap between the expiration date of the first-issued warning and the issuance date of second-issued warning. This method matches what is done by Ellis et al. (2023) and Burow et al. (2025) but differs from Nielsen et al. (2015) who only considered warnings that were issued with 30 minutes of each other. The method used by Ellis et al. (2023) and Burow et al. (2025) was chosen to ensure that overlapping warnings occurring at the later end of the first-issued warning are not overlooked, especially considering flash flood warnings may last for several hours (Doswell et al. 1996).

Intersecting warning polygons are generated from January 2008 to December 2024 using the GeoSeries intersect function with the tornado and flash flood warning polygon datasets. Intersection polygons with an area less than 1 km$^2$ are omitted to remove unusually small intersections and geometric artifacts. For each valid intersection polygon, the coordinates of the polygon's centroid and the issuance time of the later-issued warning are used to represent the location and time of a TORFF event. Figure 2.1 demonstrates an example of an intersecting warning scenario.

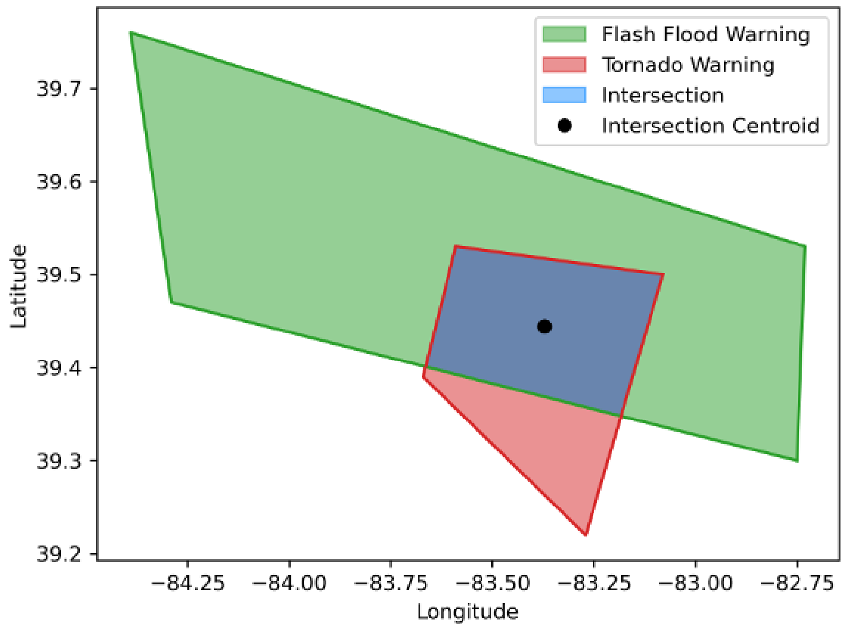


Figure 2.1: Example tornado warning (red polygon), flash flood warning (green polygon), and intersection (blue polygon) with centroid (black dot).

### 2.1.4 Global Ensemble Forecast System

Data used to train the RFML model comes from NOAA's Second Generation Global Ensemble Forecast System Re-forecast (GEFS/R). The GEFS/R is a global, convection-paramaterized, 11-member ensemble with 55 km grid spacing and is initialized once a day at 0000 UTC (Hamill et al. 2013). Data is retrieved from the AWS archive from January 2008 to December 2019 for a 12 year training period. This is the maximum possible length of the training period since the intersecting warning dataset used for model targets starts in 2008, and the GEFS/R dataset end in 2019. The length of the training period was maximized to combat the relative small sample size of targets which can lead to overfitting (Charilaou and Battat 2022). Forecast files are archived as grib2 files which are then combined and converted to netCDF for ease of use. The GEFS/R archive maintains 5 members of the original ensemble, and the ensemble median at each grid point is considered as input for training. An initial subset of 14 model predictors from GEFS/R data are chosen to include a blend of thermodynamic and kinematic variables at surface and upper levels with known connections to TORFF environments. In addition to atmospheric variables, 3 static variables are included: grid point latitude, grid point longitude, and meteorological season of the forecast day (encoded as 1, 2, 3, and 4 for DJF, MAM, JJA, and SON, respectively). The full list of predictors used is listed in table 2.3. For the purposes of this study, only Day 1

| **Symbol** | **Description** | **Class** |
|---|---|---|
| APCP | Precipitation accumulation in past (3) 6 hr | None |
| CAPE | Surface-based convective available potential energy | Thermodynamic |
| CIN | Surface-based convective inhibition | Thermodynamic |
| PWAT | Total precipitable water | Thermodynamic |
| Q2M | Specific humidity two meters above ground | Thermodynamic |
| T2M | Air temperature two meters above ground | Thermodynamic |
| U10 | Zonal component of 10-m wind | Kinematic |
| V10 | Meridional component of 10-m wind | Kinematic |
| U500 | Zonal component of 500 hPa wind | Kinematic |
| V500 | Meridional component of 500 hPa wind | Kinematic |
| Q500 | Specific humidity at 500 hPa | Thermodynamic |
| T500 | Air temperature at 500 hPa | Thermodynamic |
| Z500 | Geopotential height at 500 hPa | Kinematic |
| SRH03 | Storm relative helicity from surface to 3 km | Kinematic |
| Latitude | Grid point latitude | Static |
| Longitude | Grid point longitude | Static |
| Season | Meteorological forecast day season | Static |

Table 2.1: Predictor variables used in the RF model.

GEFS/R forecasts are considered which are archived in 3 hour intervals. For a 1200 UTC to 1200 UTC forecast, this equates to 9 forecast hours per day. The ensemble median of each predictor for each forecast hour is aggregated which generates 153 model features per grid point each day. A predictor radius was not used in this study as it can significantly increase computation time, and non-zero predictor radii typically do not improve forecast performance by a meaningful amount (Herman and Schumacher

2018a).

## 2.2 Methods

### 2.2.1 Linear Regression Model

As previously introduced, one of the most significant gaps in TORFF research lies in the understanding of the statistical relationship between tornadoes and flash floods, specifically if they are to be treated as independent or dependent events. If tornadoes and flash floods are truly independent, then the occurrence of one would have no impact on the likelihood of the other occurring. Determining whether or not a significant relationship exists between these two hazards will improve the overall understanding of TORFF events and allow for the development of a proper forecasting paradigm. One way to demonstrate if two variables are independent is through a linear regression model to predict the value of one variable given a value of the other. If a significant correlation is indicated by the model, then the two variables are said to be dependent on each other.

The linear regression analysis is performed on daily counts of tornado and flash warnings from January 2008 to December 2024 to illustrate the linear relationship between their occurrences. The analysis is segmented into the NWS CWAs (NOAA 2025) as they act as the natural boundaries for warning polygons and allow for the correlation to be visualized in different regions. The number of tornado warnings and number of flash flood warnings issued on each day within a CWA are aggregated to first understand how hazard warning issuance may vary with one another. Each CWA is analyzed over the 17-year period and warning issuance pairs are generated for each day. Naturally, the distributions for tornado and flash flood warning counts on a single day are extremely exponential due to the overwhelming majority of days that have 1 or less warning. To account for these skewed distributions, only days with at least 1 of each warning are considered to prevent a false negative correlation. Additionally, values greater than 4 standard deviations from the mean for each CWA are omitted since very large outliers can significantly skew exponential distribution correlations (Yuan and Bentler 2001).

To build the linear regression model, tornado warning counts are used as the dependent/$x$ variable and flash flood warning counts are used as the independent/$y$ variable. A two-sided hypothesis test is then used to determine the significance ($p$) of the correlation. Due to the underlying exponential distributions, a Monte Carlo resampling method is used for hypothesis testing with random exponential distributions resampled 5000 times for each correlation calculation. The linear regression analysis uses the least squares method to determine the line of best fit, and the Pearson correlation coefficient ($r$) is calculated with the following formula:

$$r = \frac{\sum(x-\bar{x})\sum(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \tag{2.1}$$

The Monte Carlo hypothesis test uses the following exponential distribution function to test the significance of $r$:

$$f(x) = \frac{1}{\beta}\exp(-\frac{x}{\beta}) \tag{2.2}$$

where $\beta$ is the scale parameter. $\beta$ is set to 3 to match the shape of the exponential distributions observed in the warning counts. The chosen significance or alpha ($\alpha$) value for this study is 0.05 which corresponds to a 95% confidence interval. Correlation coefficients and significance are calculated for every CWA with at least 30 data points to ensure an accurate linear regression analysis.

### 2.2.2 Independent Joint Outlook

In the case that tornadoes and flash floods *can* be treated as independent events, the most straightforward method to produce a TORFF forecast involves multiplying the SPC TO and WPC ERO probabilities together to produce the joint probability of both events occurring. This method was defined previously as the independent joint outlook (IJO), and follows the basic rule of joint probability between two independent variables:

$$P(A \cap B) = P(A) * P(B) \tag{2.3}$$

In terms of tornado and flash flood probability, (2.3) can be rewritten as:

$$P(\text{TORFF}) = P(\text{TOR}) * P(\text{FF}) \tag{2.4}$$

where $P(\text{TORFF})$ is the probability of a TORFF event, $P(\text{TOR})$ is the probability of a tornado, and $P(\text{FF})$ is the probability of a flash flood. Since $P(\text{TOR})$ and $P(\text{FF})$ can be estimated with the SPC and WPC outlooks, respectively, (2.4) becomes:

$$P(\text{TORFF})_{\text{IJO}} = P(\text{TOR})_{\text{SPC}} * P(\text{FF})_{\text{WPC}} \tag{2.5}$$

where $P(\text{TOR})_{\text{SPC}}$ is the probability of a tornado with 25 miles of point as given by the SPC TO and $P(\text{FF})_{\text{WPC}}$ is the probability of a flash flood (FFG exceedance) within 25 miles of a point as given by the WPC ERO. $P(\text{TORFF})_{\text{IJO}}$ therefore gives the probability of a tornado and FFG exceedance occurring at the same time and area within 25 miles of a point, effectively estimating the probability of a TORFF event occurring. For each day with a valid SPC TO and WPC ERO from Jaunary 2017 to December 2024, forecasts for $P(\text{TORFF})_{\text{IJO}}$ are generated which are valid from 1200

UTC to 1200 UTC. Furthermore, a Gaussian smoothing filter is applied to make the forecast probabilities more continuous, which are shown to verify better than those that are discretized (Herman et al. 2018). The Gaussian filter uses the following N-dimensional Gaussian equation:

$$G(x) = \frac{1}{(\sqrt{2\pi}\sigma)^N} \exp(-\frac{||x||^2}{2\sigma^2}) \tag{2.6}$$

where $N$ is the number of dimensions of the input array and $\sigma$ is the standard deviation of the Gaussian which is set to 0.5. Figure 2.2 illustrates a high-end probability case for the IJO forecast.

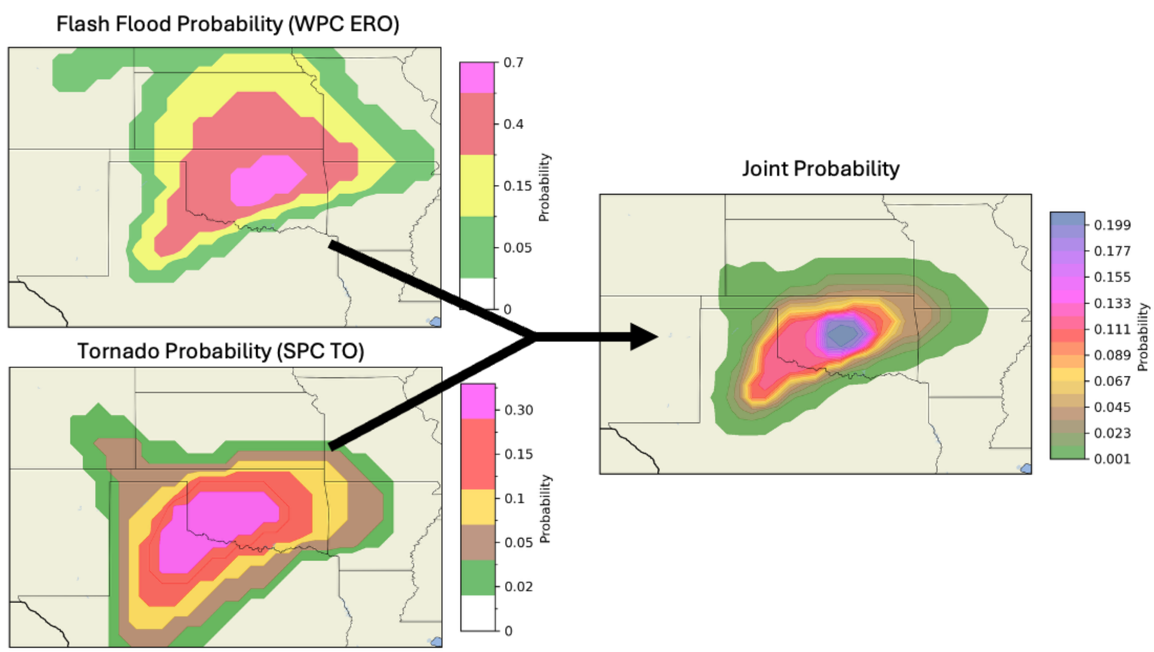


Figure 2.2: Gridded 0830 UTC WPC ERO (top left), 1300 UTC SPC TO (bottom left), and IJO forecast (right) for 20 May, 2019.

### 2.2.3 Conditional Probability

Given the hypothesis that tornadoes and flash floods are not independent of one another, conditional probability is necessary to show the likelihood of one event occurring given that the other has occurred. This will further inform how tornadoes and flash floods are related to improve the understanding of TORFF environments. Additionally, spatial distributions of conditional probability will provide a way to forecast TORFF events without having to combine both outlooks (which assumes independence).

In the TORFF framework, $P(\text{FF}|\text{TOR})$ is the conditional probability that a flash flood occurs given a tornado has occurred, and $P(\text{TOR}|\text{FF})$ is the inverse. Two distinct methods are used to approximate $P(\text{TOR}|\text{FF})$ and $P(\text{FF}|\text{TOR})$: one based on tornado and flash flood warnings, and another using the SPC and WPC outlooks. These methods are classified as warning-scale and outlook-scale, respectively. For both scales, conditional probability is determined with the following generalized formulae:

$$P(\text{TOR}|\text{FF}) = \frac{P(\text{TOR} \cap \text{FF})}{P(\text{FF})} \tag{2.7}$$

$$P(\text{FF}|\text{TOR}) = \frac{P(\text{TOR} \cap \text{FF})}{P(\text{TOR})} \tag{2.8}$$

#### 2.2.3.1 Warning Scale

By using daily tornado and flash flood warnings to quantify conditional probability, (2.7) and (2.8) equate to the ratio between the number of days when both warnings are issued and the number of days each individual warning type is issued. (2.7) and (2.8) can therefore be rewritten as:

$$P(\text{TOR}|\text{FF})_{\text{Warn}} = P(\text{TOR-Warn}|\text{FF-Warn}) = \frac{P(\text{TOR-Warn} \cap \text{FF-Warn})}{P(\text{FF-Warn})} \tag{2.9}$$

$$P(\text{FF}|\text{TOR})_{\text{Warn}} = P(\text{FF-Warn}|\text{TOR-Warn}) = \frac{P(\text{TOR-Warn} \cap \text{FF-Warn})}{P(\text{TOR-Warn})} \tag{2.10}$$

Ideally, (2.9) and (2.10) are calculated for each point in the forecast grid, but the spatial extent of warnings does not generate sufficient data for individual points. Instead, these ratios are calculated for each NWS CWA with at least 10 days where both warnings occur between January 2008 and December 2016. This date range is chosen to separate the conditional probability data from the forecast period (2017-2024) to prevent inflated skill in conditional forecasts. Therefore, $P$(TOR-Warn) is the daily probability of a tornado warning within the CWA, $P$(FF-Warn) is the daily probability of a flash flood warning within the CWA, and $P$(TOR-Warn $\cap$ FF-Warn) is the daily probability of both warnings occurring within the CWA (not necessarily intersecting).

#### 2.2.3.2 Outlook Scale

The use of warnings to quantify conditional probability presents numerous caveats such as limited spatial representation and differences in warning procedure between Weather Forecast Offices (WFO) (Harrison and Karstens 2017). In response to this, Day 1 SPC TOs and WPC EROs are considered as an alternative to quantify conditional probability. Although outlooks do not specify where events occurred, they indicate grid points where atmospheric conditions are favorable for the respective hazard. This method therefore highlights the general synoptic-scale conditional probability of tornado and flash flood environments. Here, (2.7) and (2.8) equate to the ratio between the number of days when both outlooks (irrespective of probability) are issued and the number of days each individual outlook (irrespective of probability) is issued. (2.7) and (2.8) can therefore be rewritten as:

$$P(\text{TOR}|\text{FF})_{\text{Otlk}} = P(\text{TOR-Otlk}|\text{FF-Otlk}) = \frac{P(\text{TOR-Otlk} \cap \text{FF-Otlk})}{P(\text{FF-Otlk})} \tag{2.11}$$

$$P(\text{FF}|\text{TOR})_{\text{Otlk}} = P(\text{FF-Otlk}|\text{TOR-Otlk}) = \frac{P(\text{TOR-Otlk} \cap \text{FF-Otlk})}{P(\text{TOR-Otlk})} \tag{2.12}$$

where $P$(FF-Otlk) is the probability of a WPC ERO of any probability being issued at the grid point, $P$(TOR-Otlk) is the probability of an SPC TO of any probability being issued at the grid point, and $P(\text{TOR-Otlk} \cap \text{FF-Otlk})$ is the probability of both outlooks being issued at the grid point. (2.11) and (2.12) are calculated for each grid point with at least 10 overlapping outlook days between January 2017 and December 2024, representing the full range of available outlooks.

### 2.2.4 Conditional Joint Forecasts

In the case that tornadoes and flash floods are not independent, conditional probability must be used instead which estimates the probability of a TORFF based only on the knowledge of one event's likelihood and the conditional probability of the other hazard. Therefore, $P$(TORFF) is given by one of the following formulae:

$$P(\text{TORFF}) = P(\text{TOR})_{\text{SPC}} * P(\text{FF}|\text{TOR}) \tag{2.13}$$

$$P(\text{TORFF}) = P(\text{FF})_{\text{WPC}} * P(\text{TOR}|\text{FF}) \tag{2.14}$$

In an ideal probability space, (2.13) and (2.14) give the same value, but sampling variability between tornadoes, flash floods, and their respective forecasts makes this an unrealistic assumption in this case. Therefore, it is necessary to use both $P$(TOR|FF) and $P$(FF|TOR) and create separate joint forecasts with each conditional probability distribution. Since $P$(TOR|FF) and $P$(FF|TOR) were estimated using warnings and outlooks, there are 4 conditional joint forecasts methods to be considered:

1. Flash Flood Warning Conditional (FWC):
   $P(\text{TORFF})_{\text{FWC}} = P(\text{TOR})_{\text{SPC}} * P(\text{FF}|\text{TOR})_{\text{Warn}}$

2. Tornado Warning Conditional (TWC):

   $P(\text{TORFF})_{\text{TWC}} = P(\text{FF})_{\text{WPC}} * P(\text{TOR}|\text{FF})_{\text{Warn}}$

3. Flash Flood Outlook Conditional (FOC):

   $P(\text{TORFF})_{\text{FOC}} = P(\text{TOR})_{\text{SPC}} * P(\text{FF}|\text{TOR})_{\text{Otlk}}$

4. Tornado Outlook Conditional (TOC):

   $P(\text{TORFF})_{\text{TOC}} = P(\text{FF})_{\text{WPC}} * P(\text{TOR}|\text{FF})_{\text{Otlk}}$

Flash flood conditional forecasts (FWC and FOC) multiply $P(\text{TOR})_{\text{SPC}}$ by the conditional probability of a flash flood as estimated by the warning and outlook scale conditional probability methods, respectively. Conversely, tornado conditional forecasts (TWC and TOC) multiply $P(\text{FF})_{\text{WPC}}$ by the conditional probability of a tornado as estimated by the warning and outlook scale conditional probability methods, respectively. For each of these forecasts, grid points that have insufficient conditional probability data are replaced with the IJO forecast to ensure forecast continuity. Similar to the IJO forecast, conditional joint forecasts are generated for each day with a valid SPC TO and WPC ERO outlook from Jaunary 2017 to December 2024, and are valid from 1200 UTC to 1200 UTC. The Gaussian smoothing filter is also applied (after multiplication) but with $\sigma = 1$ to enhance smoothing with conditional probability grids.

#### 2.2.4.1 Conditional Averaging

As a result of using conditional probability, the joint forecast methods outlined in this section can only utilize one of the two operational outlooks. Given that both outlooks provide valuable forecast information for TORFFs, it can be argued that disregarding one would result in a less skillful forecast. Furthermore, a user of conditional joint forecasts would have to choose between one or the other which can further complicate the forecast process. To resolve this, a third forecast method is proposed for both scales that averages the two conditional forecasts:

$$P(\text{TORFF}) = \frac{P(\text{TOR}) * P(\text{FF}|\text{TOR}) + P(\text{FF}) * P(\text{TOR}|\text{FF})}{2} \tag{2.15}$$

This method therefore makes use of both operational outlooks and conditional probability so that valuable forecast information is not disregarded. Averaging is done prior to smoothing by the Gaussian filter on the individual conditional joint forecasts, then the filter with $\sigma = 1$ is applied. The names and equations for these forecasts are as follows:

1. Warning Conditional Average (WCA):
   $P(\text{TORFF})_{\text{WCA}} = \frac{P(\text{TOR})_{\text{SPC}} * P(\text{FF}|\text{TOR})_{\text{Warn}} + P(\text{FF})_{\text{WPC}} * P(\text{TOR}|\text{FF})_{\text{Warn}}}{2}$
2. Outlook Conditional Average (OCA):
   $P(\text{TORFF})_{\text{OCA}} = \frac{P(\text{TOR})_{\text{SPC}} * P(\text{FF}|\text{TOR})_{\text{Otlk}} + P(\text{FF}) * P(\text{TOR}|\text{FF})_{\text{Otlk}}}{2}$

With the inclusion of conditional average forecasts (WCA and OCA), there are 7 joint forecast methods in total that estimate $P(\text{TORFF})$. For reference, the names, acronyms, and equations for each forecast are summarized in Table 2.2.

### 2.2.5 Forecast Evaluation

Forecasts are evaluated in a grid-based manner such that each grid point is treated as an independent forecast from neighboring grid points. Intersecting warnings are projected to the 55 km grid and a 40 km neighborhood radius is used to encode any grid point within 40 km of a event as a 1, otherwise 0 for nonevents. The neighborhood radius was chosen to approximately match the 25 mile event radius that SPC and WPC outlooks forecast for. This neighborhood radius also ensures that grid points with a

| **Forecast name** | **Acronym** | **Equation** |
|---|---|---|
| Independent Joint Outlook | IJO | $P(\mathrm{TOR})_{\mathrm{SPC}} * P(\mathrm{FF})_{\mathrm{WPC}}$ |
| Flash Flood Warning Conditional | FWC | $P(\mathrm{TOR})_{\mathrm{SPC}} * P(\mathrm{FF}\vert\mathrm{TOR})_{\mathrm{Warn}}$ |
| Tornado Warning Conditional | TWC | $P(\mathrm{FF})_{\mathrm{WPC}} * P(\mathrm{TOR}\vert\mathrm{FF})_{\mathrm{Warn}}$ |
| Flash Flood Outlook Conditional | FOC | $P(\mathrm{TOR})_{\mathrm{SPC}} * P(\mathrm{FF}\vert\mathrm{TOR})_{\mathrm{Otlk}}$ |
| Tornado Outlook Conditional | TOC | $P(\mathrm{FF})_{\mathrm{WPC}} * P(\mathrm{TOR}\vert\mathrm{FF})_{\mathrm{Otlk}}$ |
| Warning Conditional Average | WCA | $\frac{P(\mathrm{TOR})_{\mathrm{SPC}} * P(\mathrm{FF}\vert\mathrm{TOR})_{\mathrm{Warn}} + P(\mathrm{FF})_{\mathrm{WPC}} * P(\mathrm{TOR}\vert\mathrm{FF})_{\mathrm{Warn}}}{2}$ |
| Outlook Conditional Average | OCA | $\frac{P(\mathrm{TOR})_{\mathrm{SPC}} * P(\mathrm{FF}\vert\mathrm{TOR})_{\mathrm{Otlk}} + P(\mathrm{FF}) * P(\mathrm{TOR}\vert\mathrm{FF})_{\mathrm{Otlk}}}{2}$ |

Table 2.2: Forecast names, acronyms, and equations for $P(\mathrm{TORFF})$.

high density of event coordinates are normalized. This normalization is important as the intersecting warning dataset may contain multiple coordinates for the same TORFF event such as when multiple warnings are issued for the same storm due to CWA boundaries or if a warning is re-issued after expiration.

#### 2.2.5.1 Reliability and Sharpness

Reliability and sharpness diagrams are used in this study to visualize the performance of probabilistic forecasts. A reliability diagram indicates how well the probabilities of a forecast match the observed frequency of the events, illustrating overall calibration (Brier 1950; Murphy 1973). Probabilities for each forecast method are separated into bins and plotted against the event frequencies occurring within each bin to create a single line on the diagram. A perfectly reliable forecast, therefore, is one that creates a straight diagonal line from (0,0) to (1,1). Reliability is assessed by visually examining the departure of the forecast line from the diagonal. Values left of the line indicate under-forecasting (observed frequency > forecast) and values to the right indicate over-forecasting (observed frequency $<$ forecast). The sharpness diagram is a companion plot to reliability which illustrates the relative frequency that forecast probabilities in each bin are issued. Following the methods of Bröcker and Smith (2007), a bootstrapping approach is employed to resample forecast grids of each probability bin within the given time period 5000 times with replacement to generate a 95% confidence interval for each observed frequency in the reliability diagram.

#### 2.2.5.2 Brier Skill Score

The Brier Skill Score (BSS) is forecast metric that quantifies a probabilistic forecast's performance relative to a climatological reference forecast (Wilks 2011). A BSS above 0 indicates that the forecast is an improvement over the reference, and a score below 0 indicates the forecast is worse than consistently forecasting the reference. The reference climatological forecast is computed by taking the mean value of all event grids (which gives the climatological frequency) and interpolating that value to all grid points for all forecast days. The BSS is then computed with the following formula:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{forecast}}}{\mathrm{BS}_{\mathrm{ref}}} \tag{2.16}$$

where $\mathrm{BS}_{\mathrm{ref}}$ is the Brier score of the reference climatological forecast and $\mathrm{BS}_{\mathrm{forecast}}$ is the Brier score of the forecast. The Brier score is computed as:

$$\mathrm{BS} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{2.17}$$

where $N$ is the total number of grid points, $f$ is the forecast, and $o$ is the binary classification of an event. The Brier score on its own is a strictly proper scoring rule that calculates the mean squared error (MSE) between forecast probabilities and the observed event classification (Brier 1950). Therefore, a perfect Brier score is 0, and the worst is 1. Similar to reliability, a bootstrapping approach is employed to resample forecast grids within the given time period 5000 times to generate a 95% confidence interval for each BSS calculation.

#### 2.2.5.3 Contingency Tables

Contingency tables are a valuable tool used to derive a large variety of statistics that help characterize the skill of binary forecasts. Although probabilistic forecasts are not binary in nature, the probabilities can be thresholded to create a binary yes/no forecast area. For each thresholded area, the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are calculated based on the binary event classification of each grid point. In this study, each forecast is thresholded based on the distribution of it's probabilities, and contingency table statistics for each threshold are used to to determine the probability of detection (POD), success ratio (SR), and critical success index (CSI):

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.18}$$

$$\mathrm{SR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.19}$$

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.20}$$

Values of POD and SR for each threshold are plotted against each other to generate performance diagrams which illustrate the tradeoff between missed events and false alarms (Roebber 2009). A forecast with a perfect tradeoff (no missed events or false alarms) would have the maximum CSI of 1 which corresponds to the very top-right of a performance diagram where POD an SR are maximized. Comparing performance diagrams and CSI for each forecast will therefore illustrate which have the best overall spatial discrimination across their forecast probabilities.

### 2.2.6 Random Forest Machine Learning Model

An RFML model was designed as an alternative to using the SPC TO or WPC ERO forecasts to asses its capabilities with extremely rare events and to compare with the joint forecasts methods across test cases to evaluate spatial discrimination. A random forest is a machine learning technique that combines a large number of decision trees to improve forecast accuracy and reduce overfitting. Each tree is trained using a random subset of the data and predictors, allowing the model to capture nonlinear relationships and complex interactions among variables. The final prediction is produced by aggregating the outputs from all trees, typically through majority voting for classification or averaging for regression (Breiman 2001). Random forests are widely used in meteorological forecasting because they perform well with large datasets and can effectively model complex atmospheric processes.

#### 2.2.6.1 Model Domain

Figure 2.3 displays the domain used to train the RFML model. This domain was chosen to represent the area east of the Rockies where the majority of TORFF events occur (95%) and includes the gulf and east coasts to account for TORFFs that occur from landfalling tropical cyclones (Nielsen et al. 2015; Burow et al. 2021).

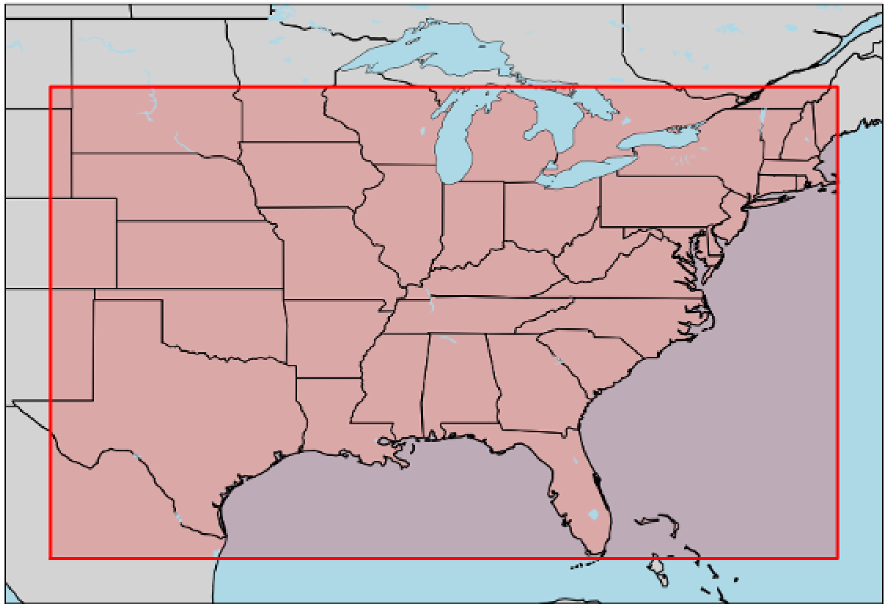

Figure 2.3: Training domain (red shaded rectangle) for RFML model

#### 2.2.6.2 K-fold Cross Validation

A common pitfall of machine learning for severe weather, especially with low sample size, is that the model may overfit to a particular training period which can inflate skill (Charilaou and Battat 2022). To address this, a K-Fold cross validation technique is implemented that splits the 12 year training period into 4 unique training/validation periods (4-fold). Each fold is a unique selection of 9 years used to train the model and the remaining 3 years are used for validation/testing. To ensure that each fold maintains roughly the same distribution of events and non-events, the folds are stratified by grouping the data by year and selecting folds that maintain class balance across all folds (Figure 2.4). Grouping the data by year ensures that no folds contain an uneven distribution of TORFF events from a particular season to prevent overfitting.

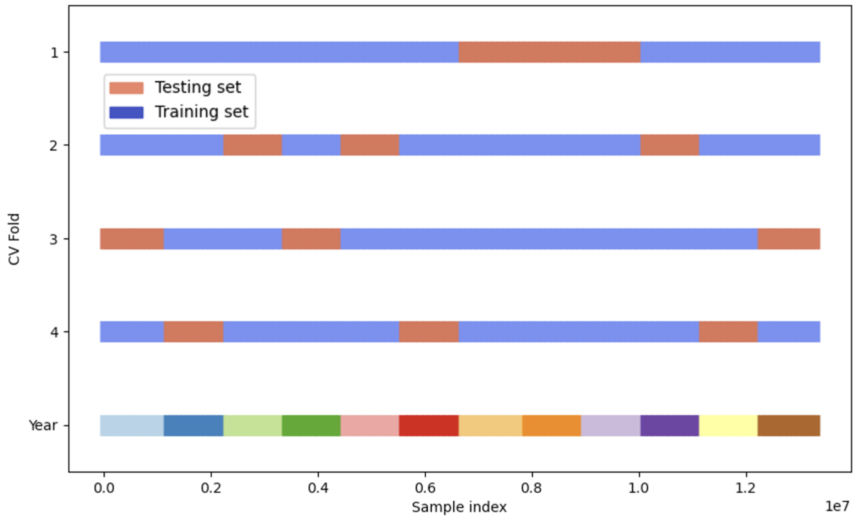


Figure 2.4: Testing (orange) and training (blue) segments for each cross validation fold. Groups (years) are visualized at the bottom with a different color representing each year within the sample index.

#### 2.2.6.3 Model Parameters

The RFML model is created using the Scikit-learn RandomForestClassifier which fits decision tree classifiers on various subsets of the data and uses averaging to prevent over-fitting and improve predictive accuracy (SciKit-Learn Developers 2026). The classifier can be tuned with a wide array of hyperparameters, the most important for this study being the number of trees, minimum sample split, and class weight. A manually implemented grid search technique was utilized to determine which values of these three hyperparameters maximizes the aggregate BSS across all 4 cross-validation folds; tested and final values are presented in Table 2.3. Split quality criterion and max tree depth are also shown in Table 2.3 as they were set to non-default values, but a grid search was not performed due to computational limitations. Instead, the values for these parameters reflect what is used by Herman and Schumacher (2018b) who demonstrated success with similar RFML configurations. All other model parameters are set to the defaults as defined by the classifier function.

| Parameter | Tested Values | Final Value |
|---|---|---|
| Number of trees | 100, 250, 500, 1000, 1500 | 1000 |
| Minimum sample split | 30, 60, 120, 180 | 60 |
| Split quality criterion | Entropy | Entropy |
| Max tree depth | 1 | 1 |
| Class weight | 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2 | 1 |

Table 2.3: Non-default RandomForestClassifier hyperparameter values and tested grid search values.

#### 2.2.6.4 Test Cases

The final model used for testing is the one validated on a segment that produced the highest BSS after hyperparameter tuning. Segment 3 produced the highest BSS of 0.04, so it's classifier will be used to generate forecasts. Three test cases are analyzed to compare the spatial discrimination of the forecasts generated by the RFML model to the joint forecasts. For each test case, operational GEFS forecast data are obtained from the AWS archive to generate input to the classifier which make a single forecast for the day. Similar to the training data, operational GEFS forecast data for each predictor variable is obtained from forecast hours 12 to 36 in 3-hr intervals. Initial tests with the model showed moderate noisiness, so the Gaussian smoothing filter used with previous forecast methods is applied to the forecast grid with $\sigma = 1$ to reduce noise.

# Chapter 3

# Results and Discussion

## 3.1 Warning and Outlook Frequency

Before diving into the analyses of tornado and flash flood relationships, counts of tornado, flash flood, and intersecting warnings within each CWA are analyzed from 2008 to 2024 to illustrate their regional frequency and discern any patterns between them. The spatial frequency of overlapping SPC TOs and WPC EROs from 2017 to 2024 is also analyzed to asses if regions with relatively high overlap frequency generally correspond with the frequency of intersecting warnings.

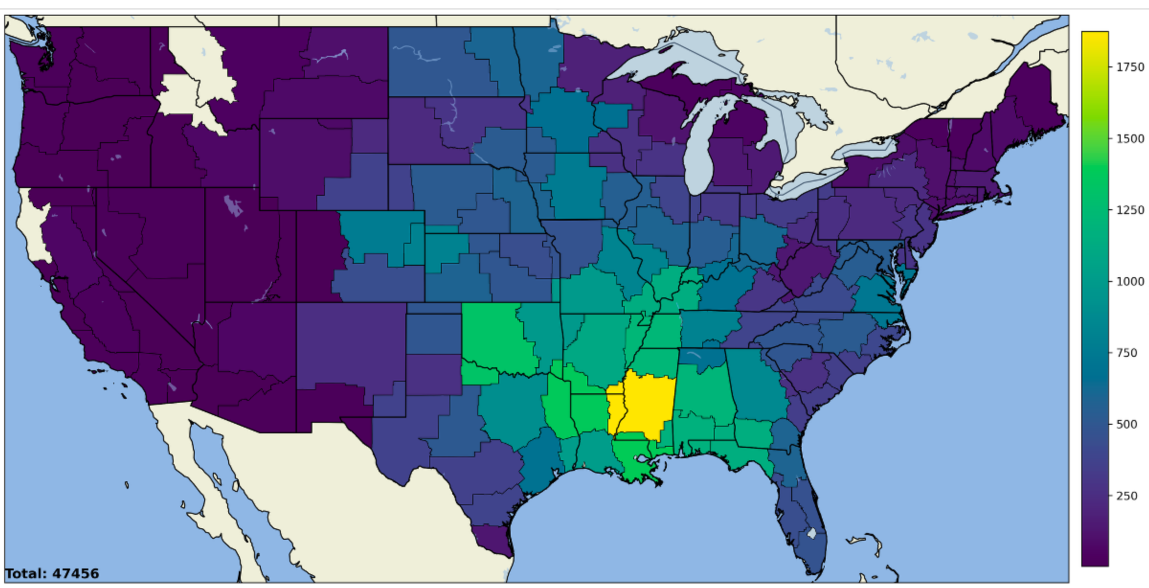


Figure 3.1: Total tornado warnings issued within each CWA from 2008 to 2024.

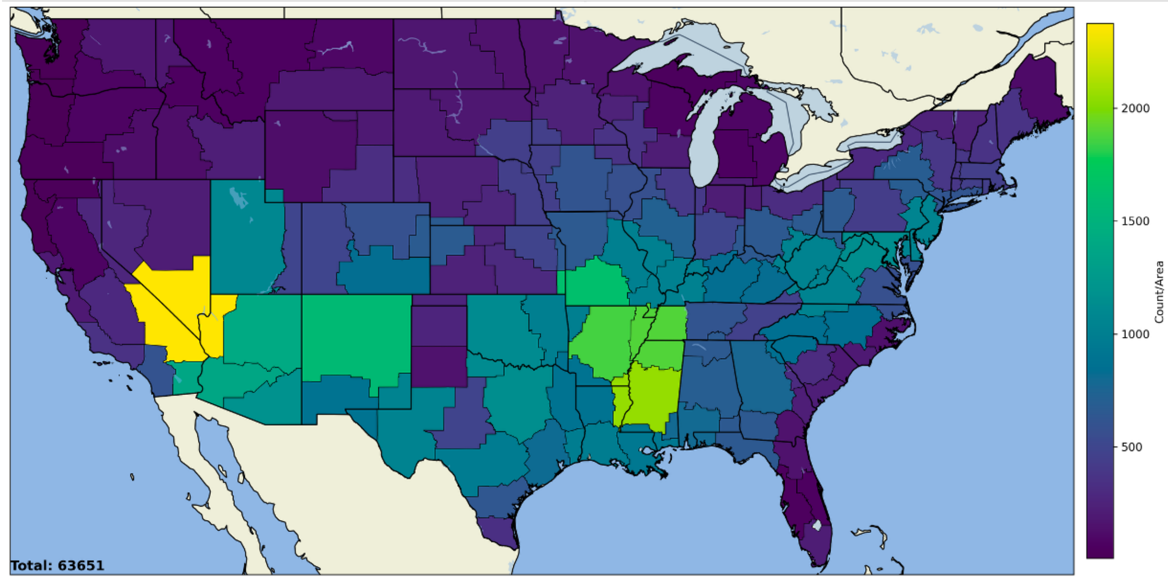


Figure 3.2: As in Figure 3.1 but for flash flood warnings

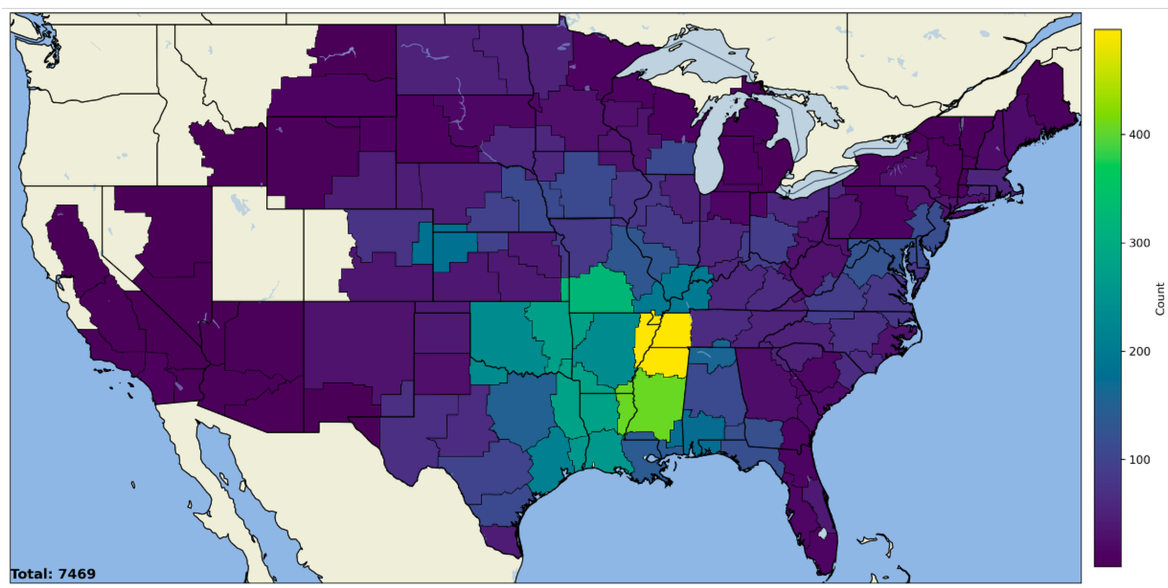


Figure 3.3: As in Figure 3.1 but for intersecting warnings

From 2008 to 2024, 47456 tornado warnings and 63651 flash flood warnings were issued across the CONUS CWAs. Tornado warnings occur most frequently in the southeast US, particularly in Oklahoma, eastern Texas, and the southern portion of the Mississipppi River valley (Figure 3.1). The Jackson, MS CWA (central Mississippi) has the most overall tornado warnings issued (1874) which is nearly 500 greater than the CWA with the next highest count. As expected, CWAs west of the Rockies have relatively few tornado warnings issued as well as CWAs in the Great Lakes region, New England, and Appalachia. Flash flood warnings occur most frequently in the desert southwest and a portion of the Mississippi River valley including Arkansas, west Tennessee, and Mississippi (Figure 3.2). The Las Vegas, NV CWA (southern Nevada/-southeastern California) has the most overall flash flood warnings issued (2375), and Jackson, MS has the second highest (2057). Flash flood warning counts are generally lower along the southern east coast, Great Lakes region, northern and central plains, Pacific Northwest, northern California, and the Northern Rockies.

For the same period, 7469 intersecting warnings occurred across CONUS CWAs which averages to approximately 440 a year, roughly matching what was found by Nielsen et al. (2015) although slightly higher. Intersecting warnings occur most frequently along the southern Mississippi River valley, Oklahoma, and west Texas, with a maximum (497) in the Memphis, TN CWA (western Tennessee/northern Mississippi) (Figure 3.3). This frequency generally matches what was found by Ellis et al. (2023) and Nielsen et al. (2015), especially the maximum in Memphis. Compared to tornado and flash flood warnings, intersecting warnings are confined to a smaller geographic area which aligns with where there is an intersection of relatively high tornado and flash flood warning counts. This confirms that the spatial distribution and frequency of intersecting warnings is largely dictated by where tornado and flash flood warnings are both more likely to occur.

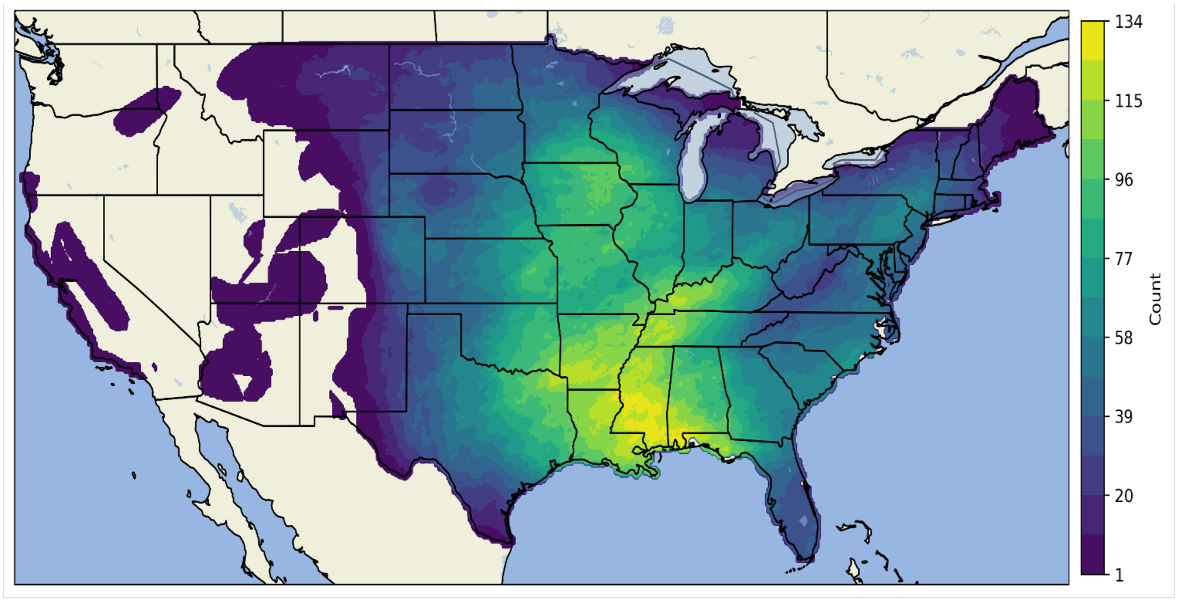


Figure 3.4: Number of overlapping SPC TOs and WPC EROs (of any probability) for each grid point from 2017 to 2024

Figure 3.4 shows that the SPC TO and WPC ERO overlap most frequently in the southern Mississippi River valley and the coasts of Mississippi and Alabama. A diagonal corridor of relatively high frequency is also observed from the southwest corner of Arkansas into north-central Kentucky. A less robust area of locally higher frequency is observed over Iowa, southeast Minnesota, southwest Wisconsin, and northwest Illinois. Similar to tornado warnings, there is relatively lower frequency in Appalachia, New England, and west of the Rockies. Comparing outlook overlap frequency (Figure 3.4) with intersecting warning frequency (Figure 3.3), there is some consistency in that both have generally the highest frequency over the lower Mississippi River valley; however, outlook overlap frequency is higher in southernmost part of this region near the Gulf. Furthermore, the region of locally higher outlook overlap frequency in the Iowa region is not observed with intersecting warnings. This suggests that the overlap of SPC TO and WPC ERO forecasts may be an overestimate of TORFF frequency in these regions. There is, however, a caveat in this analysis which is that the count of intersecting warnings is sensitive to the size of the CWA. As a result, Figure 3.3 may not accurately represent the true spatial distribution of TORFF events.

## 3.2 Tornado-Flash Flood Independence (RQ1)

A linear regression model was used to determine the statistical relationship between daily tornado and flash flood warning counts in each NWS CWA. The results of this analysis will illustrate which regions of the CONUS have a significant correlation between the occurrences of the two hazards, informing the appropriate approach to forecast TORFF events (whether or not to assume independence).

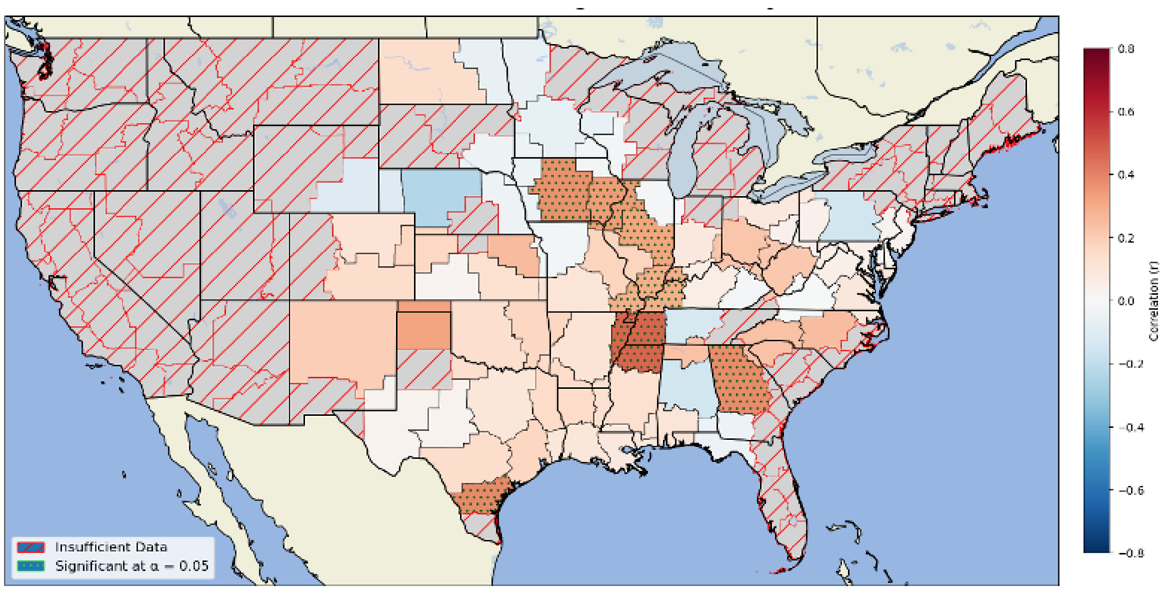


Figure 3.5: Correlation coefficient ($r$, shaded) and $\alpha = 0.05$ significance (green stippling) for each CWA. Grey shading with red hashing indicates CWAs with insufficient data (less than 30 data points).

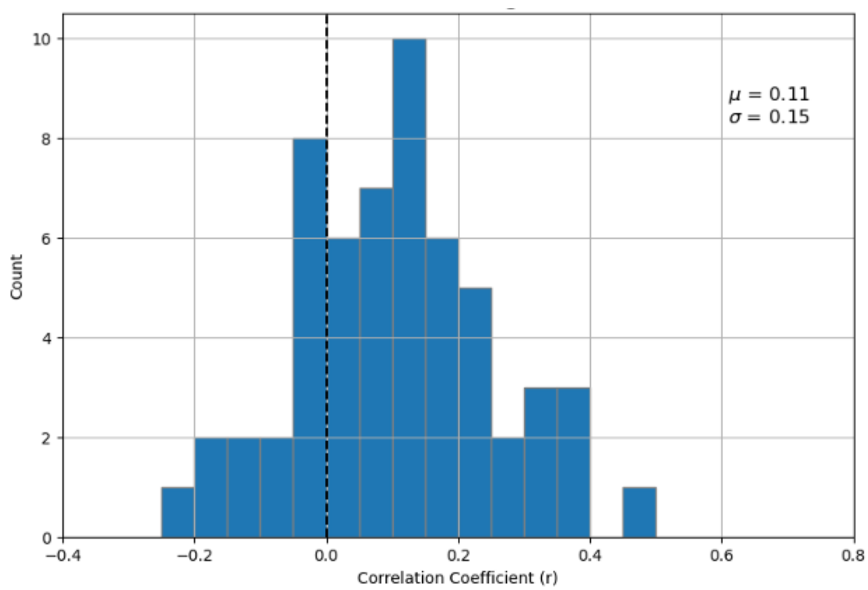


Figure 3.6: Histogram of Correlation Coefficients for each CWA with a bin interval of 0.05. Correlation mean ($\mu$) and standard deviation ($\sigma$) are shown in the top right. Black dashed line indicates 0.

Of the CONUS CWAs with sufficient data, 7 out of the 59 have a significant positive correlation between tornado and flash flood warning counts (Figure 3.5). Five of these CWAs are in the Midwest/Central US following a curved line extending from central Mississippi into central Iowa. These include the Memphis, Paducah, Lincoln, Quad Cities, and Des Moines CWAs. The CWA with the strongest correlation is Memphis ($r = 0.47$), and the lowest is North Platte (northwestern Nebraska) ($r = -0.24$). Two lone significant CWAs are also present: Corpus Christi, TX (southern Texas) and Atlanta, GA (north-central Georgia). In general, most CWAs have a weak positive correlation between tornado and flash flood warnings. However, there are a few CWAs, especially in the northern plains, that have a weak negative correlation. The distribution of correlation coefficients (Figure 3.6) is roughly uni-modal and has a slight positive (right) skew, with most values concentrated near weak positive correlations. The mean (0.11) and standard deviation (0.15) suggest that none of the correlations coefficients are significant outliers ($\pm 3\sigma$).

### 3.2.1 Discussion

Due to the lack of significance in the majority of CWAs, it is difficult to make a general conclusion on the independence of tornadoes and flash floods in the CONUS. The trend towards positive correlations somewhat supports the hypothesis that tornado and flash flood environments have similar characteristics and are therefore hard to distinguish from one another, but this conclusion cannot be sustained due to the lack of significance. Furthermore, there are no robust patterns in correlation that align with the intersecting warning (Figure 3.3) or overlapping outlook (Figure 3.4) frequencies, suggesting that tornado-flash flood warning correlation is not directly related to TORFF climatology. However, its notable that the Memphis CWA had the strongest correlation and highest intersecting warning frequency, both by a large margin. This initially suggests that regions with a significantly high tornado-flash flood correlation tend to see a higher frequency of TORFFs, but given that this is the only CWA with a correlation between the two statistics, there is not enough evidence to make this conclusion on a broader scale.

The region with 5 adjacent significant CWAs is still of particular interest as it is very unlikely that this is due to chance. Based simply off location, this pattern may be related to the synoptic-scale setup of mid-latitude cyclones which often form squall lines with heavy rain and strong convection in this region (Ashley et al. 2019; Gallus et al. 2008). Previous work has also shown that synoptic-scale squall lines tend to have a higher risk of overlapping tornado and flash flood warnings (Nielsen et al. 2015). In contrast, it is possible that the warning procedures implemented in these 5 CWAs' WFOs is similar due to their co-location, resulting in similar correlations. Henderson et al. (2020) demonstrated that the unique setup of WFOs can impact the pipeline of tornado and flash flood warning issuance; however, it is unfeasible to diagnose the setup and procedures of each individual WFO as this information is not publicly available. As a result, the explanation for this region of correlation remains unclear and is an avenue for future research.

A significant caveat with the linear regression analysis is the truncation of the warning data. As mentioned before, data points/days with zero tornado and/or flash flood warnings were omitted in order to prevent false negative correlations. Intuitively, days with only one of either warning should be considered as they represent a significant portion of the data, but this would only be possible if days with zero total warnings are also included in order to maintain the true distribution. When this was attempted, days with zero total warnings significantly skewed the data towards a mean of 0, making the results of any linear regression inconclusive. Even though significant outliers were removed, its possible that significant correlations are the result of large-scale storm systems that produced a significant amount of tornado and flash flood warnings which have a much larger effect on correlation than days with relatively few warnings. As a result, regions that are prone to frequent, larger-scale storm systems such as tropical and mid-latitude cyclones more likely to see a stronger correlation. Regions with less frequent or smaller-scale storms are more likely to have a higher proportion of truncated data that results in a weaker corelation that is not necessarily representative of the true relationship.

### 3.2.2 Forecast Implications

Despite the uncertainty in these results, the region of 5 adjacent correlated CWAs is significant enough to warrant further research into it's impact on TORFF forecasting. These results also validate the exploration of conditional probability as an alternative joint forecasting method since the IJO forecast is hypothesized to perform worse in regions with a significant correlation. To explore these ideas, two regions are delineated to separate the analysis of TORFF forecasts and compare their performance in correlated and uncorrelated regions. The first region is the combined geographic region of the 5 adjacent correlated CWA's, and the second region is the combined geographic region of all CWA's in the CONUS states east of the Rocky Mountains *excluding* the correlated CWA's (Figure 3.7). Despite it's size, the Correlated CWAs region contains a considerable proportion of events, mainly thanks to the Memphis CWA (Table 3.1).

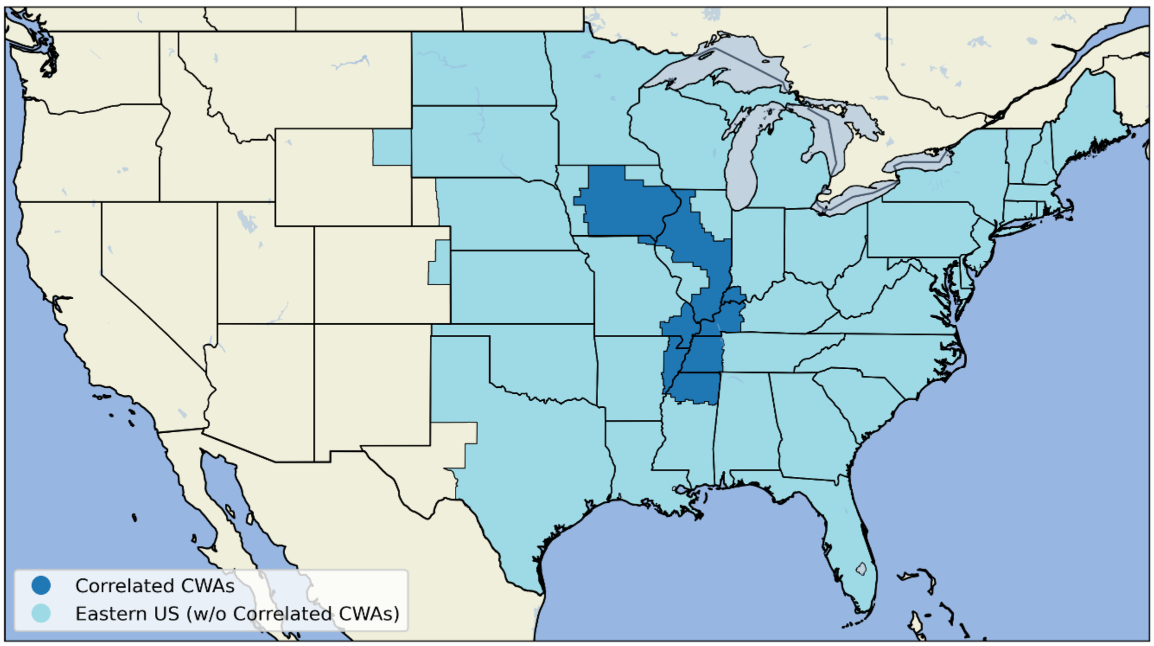


Figure 3.7: Forecast analysis regions: Correlated CWA's (dark blue) and Eastern US (w/o Correlated CWAs) (light blue)

| **Region** | **Event Proportion** |
|---|---|
| Eastern US | 81.44% |
| Correlated CWAs | 14.87% |

Table 3.1: Percentage of TORFF Events (intersecting warnings) occurring within each region

## 3.3 Conditional Probability (RQ2)

Given the results of the linear regression analysis that showed there to be several regions in the CONUS with significant tornado-flash flood correlation, the conditional probability of the two hazards is explored using warnings and outlooks as methods to estimate the relationship. The spatial distributions of conditional probabilities are compared to the results of the linear regression model (Figure 3.6) and the intersecting warning frequencies (Figure 3.3) to asses any patterns that may explain the tornado-flash flood relationship. Furthermore, the conditional probabilities found in this analysis will be used to create the conditional joint forecasts for TORFF events.

### 3.3.1 Warning Scale

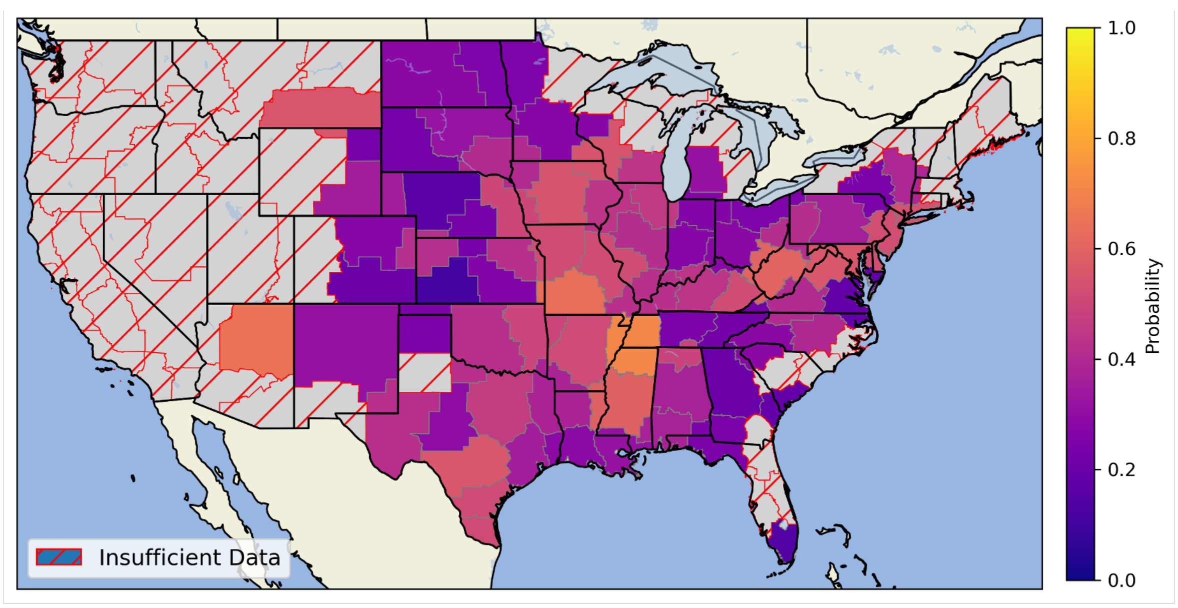


Figure 3.8: Conditional probability of a flash flood warning given a tornado warning ($P(\mathrm{FF|TOR})_{\mathrm{Warn}}$)(shaded) from 2008 to 2016 for each CWA with at least 10 dual-warning days. Red hashing indicates CWAs with insufficient data.

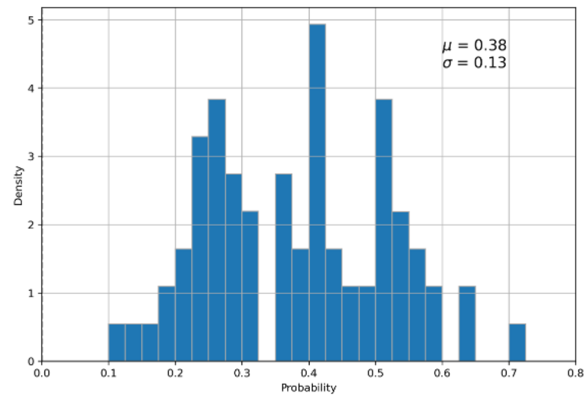


Figure 3.9: Distribution of conditional probabilities seen in Figure 3.8. The mean ($\mu$) and standard deviation ($\sigma$) of probabilities are shown in the top right.

$P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ is maximized in the central Mississippi River valley, with Memphis having the highest overall probability (0.71) (Figure 3.8). There are also relatively high probabilities in a corridor extending from West Virginia into New Jersey. Lowest probabilities are generally seen in the northern Great Plains, particularly in Kansas and Nebraska, with the Dodge City (southwestern Kansas) CWA having the lowest overall (0.11). There are also relatively low probabilities in the southeast US CWAs east of Alabama. The distribution of probabilities (Figure 3.9) is somewhat multi-modal with peaks at 0.25, 0.04, and 0.5, and it does not have a robust standard distribution shape. Given the mean (0.38) and standard deviation (0.13) of the data, none of the probabilities are considered outliers ($\pm 3\sigma$). However, the probability for Memphis does create a slight right tail.

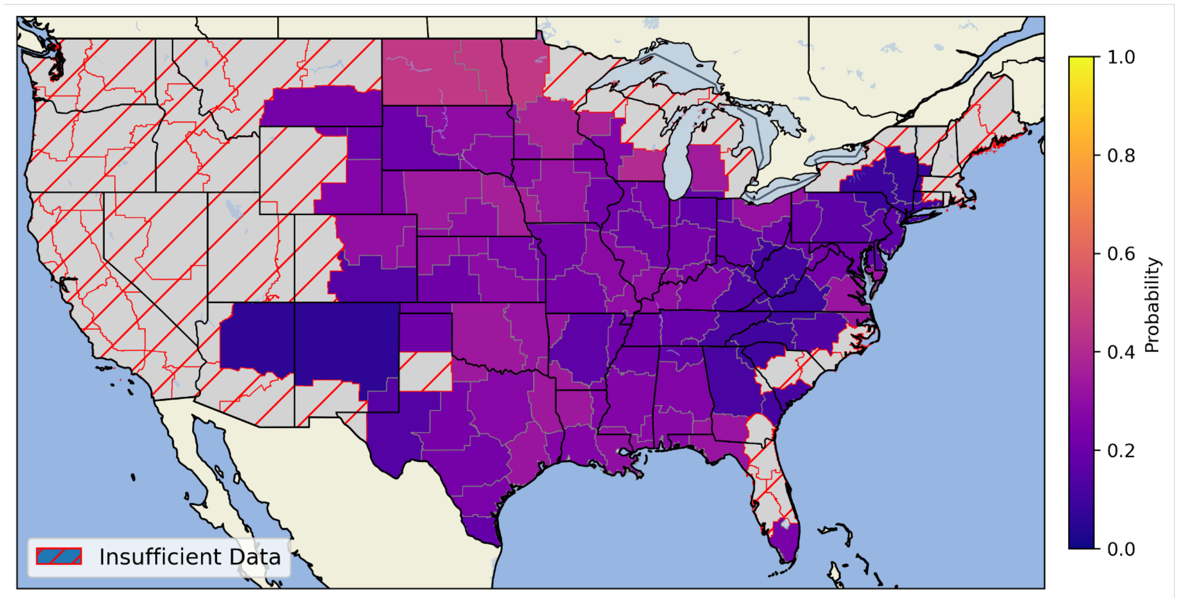


Figure 3.10: As in Figure 3.8 but for the conditional probability of a tornado warning given a flash flood warning ($P(\text{TOR}|\text{FF})_{\text{Warn}}$).

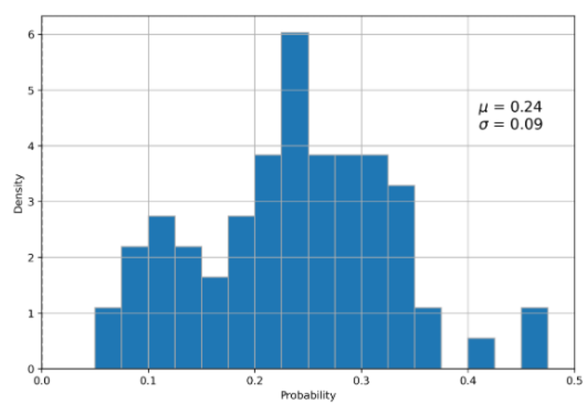


Figure 3.11: As in Figure 3.9 but for the probabilities seen in Figure 3.10

$P(\text{TOR}|\text{FF})_{\text{Warn}}$ is generally higher in the Northern Plains, maximizing in the North Dakota/Minnesota CWAs at ∼0.45 (Figure 3.10). Regions with lower probabilities are seen in the Appalachian Valley, New York, Northern New Mexico, and Northern Arizona. In the majority of the central US, probabilities are more uniform and there is little discernible spatial pattern. Figure 3.11 shows that the distribution of tornado warning conditional probabilities is bi-modal, with peaks at ∼0.1 and ∼0.225. Given the mean (0.24) and standard deviation (0.09), no probabilities are considered outliers ($\pm 3\sigma$). Despite this, there are two distinguishable probabilities that create a right tail at $\sim$0.4 and $\sim$0.45 which are the two CWAs in the North Dakota/Minnesota region.

Overall, the average value for $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ is $\sim$45% higher than $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Warn}}$. This finding is expected given that tornado warnings are issued less frequently than flash flood warnings in the majority of the United States (Figures 3.1, 3.2). The standard deviation and range of each distribution (Figures 3.9, 3.11) also indicate that $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ has more variance across CWA's than $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Warn}}$, suggesting that tornado warnings condition flash flood warnings more clearly than the inverse.

### 3.3.2 Outlook Scale

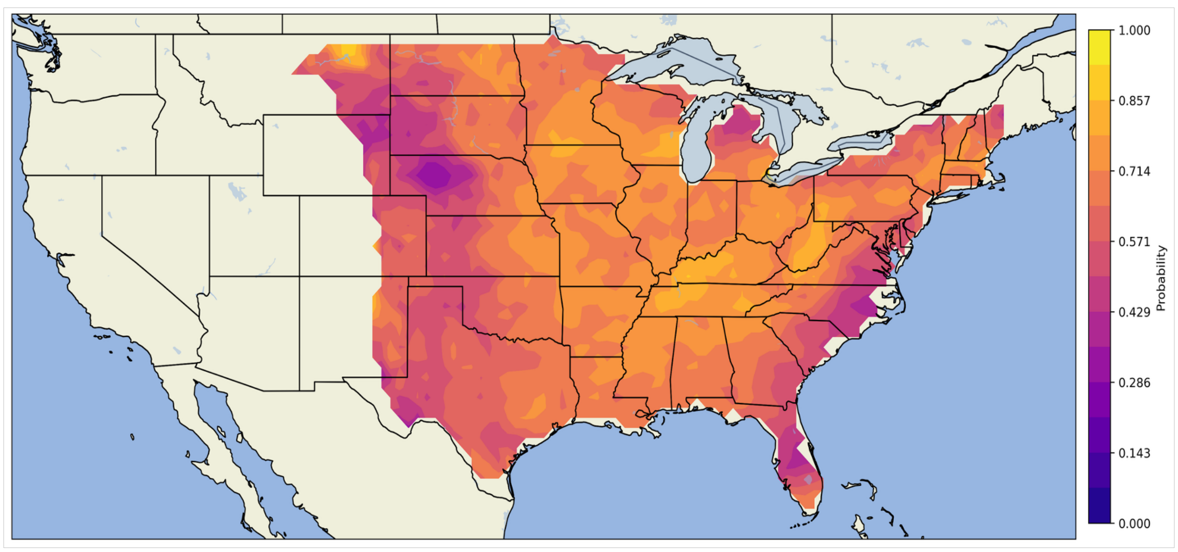


Figure 3.12: Conditional probability of a Day 1 WPC ERO issued given a Day 1 SPC TO issued ($P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$)(shaded) for each grid point with $\geq 10$ overlapping outlook days between 2017-2024. Grid points with insufficient data are unshaded.

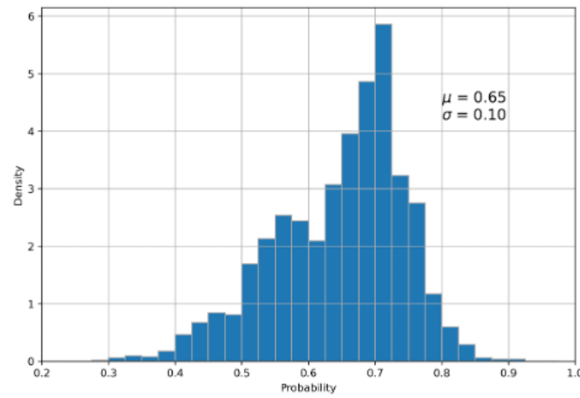


Figure 3.13: As in Figure 3.9 but for the probabilities seen in Figure 3.12

$P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$ is relative high throughout much of the east-central US, with probabilities in the 0.5-0.8 range (Figure 3.12). Locally higher probabilities are observed in the Appalachian valley, western Kentucky/Tennessee, and northeast Montana, and lowest probabilities are generally located along the east coast south of Maryland, the western half of the great plains, and northern Michigan. Similar to $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$, a region of considerably low probability is located in the northwestern part of Nebraska (∼0.3). The distribution of probabilities (Figure 3.13) is roughly uni-modal and has a negative (left) skew, indicating a concentration toward higher probabilities. The mean (0.65) and standard deviation (0.1) suggest that a small number of probabilities below 0.35 are outliers ($\pm 3\sigma$), however the skewness of the data makes it difficult to determine if these outliers are truly significant (Meropi et al. 2018).

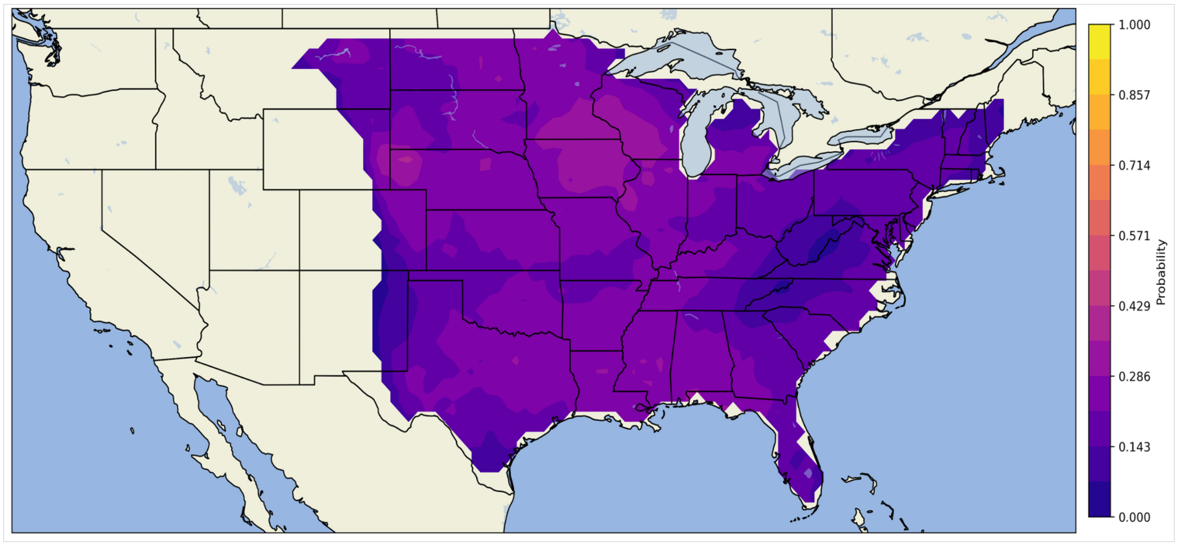


Figure 3.14: As in Figure 3.12 but for the conditional probability of a Day 1 SPC TO issued given a Day 1 WPC ERO issued ($P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Otlk}}$)).

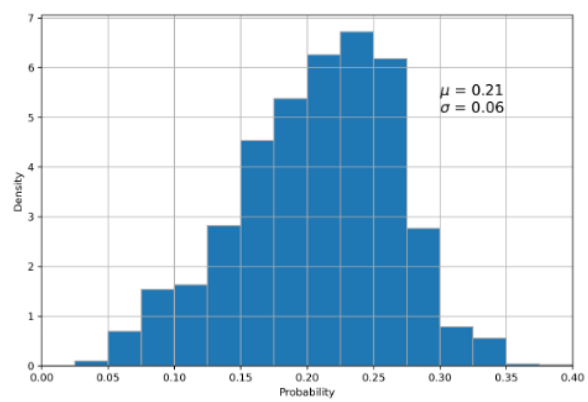


Figure 3.15: As in Figure 3.9 but for the probabilities seen in Figure 3.14

Finally, $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Otlk}}$ is maximized in a region covering parts of Iowa, southern Minnesota, and southwestern Wisconsin where probabilities are approximately 0.3-0.35 (Figure 3.14). Another region of locally high probability is seen in the northwest corner of Nebraska ($\sim$0.35). The lowest probabilities are observed in the Appalachian valley, eastern New Mexico, northern Michigan, and New England ($\leq$ 0.1). In all other regions, probabilities are roughly uniform about the mean (0.21). The distribution of probabilities (Figure 3.15) is uni-modal but has a slight negative (left) skew. The standard deviation (0.06) suggests that a small amount of probabilities below 0.03 ($\pm 3\sigma$) could be outliers, but once again the skewness makes this difficult to determine.

Overall, the average value for $P(\mathrm{FF|TOR})_{\mathrm{Otlk}}$ is ~102% higher than $P(\mathrm{TOR|FF})_{\mathrm{Otlk}}$ which is considerably larger than the difference between the warning conditional probabilities (45%). Comparing $P(\mathrm{FF|TOR})_{\mathrm{Warn}}$ with $P(\mathrm{FF|TOR})_{\mathrm{Otlk}}$, the difference in the average probability is ~52%, whereas the difference in average probability between $P(\mathrm{TOR|FF})_{\mathrm{Warn}}$ and $P(\mathrm{TOR|FF})_{\mathrm{Otlk}}$ is only ~13%. This result suggests that one of the two flash flood conditional methods is a significant over/underestimation of the true conditional probabilities, whereas the tornado conditional methods have better agreement. Therefore, a larger difference should be expected in the performance of flash flood conditional forecasts (FWC and FOC) compared to tornado conditional forecasts (TWC and TOC).

### 3.3.3 Discussion

None of the conditional probability maps showed a robust spatial pattern that aligned with the results of the linear regression model (Figure 3.6) or the intersecting warning frequencies (Figure 3.3). It was hypothesized that there would be locally higher (lower) conditional probability in regions with a strong positive (negative) correlation or where intersecting warnings were more (less) frequent, but this was not robustly indicated in any of the figures. This finding initially indicates that the results of the linear regression model are not consistent with where tornadoes and flash floods environments are conditionally related. Furthermore, since conditional probabilities do not appear to correlate well with TORFF frequency, it may suggest that treating tornadoes and flash floods as independent events will likely result in better forecast performance. There is, however, a consistency with the Memphis CWA which had the highest $P(\mathrm{FF|TOR})_{\mathrm{Warn}}$ (Figure 3.8) and the strongest correlation from the linear regression model (Figure

3.5). Memphis also has the highest rate of intersecting warnings (Figure 3.3) which may explain this locally high probability. Furthermore, $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$ had a maxima in the western Tennessee region which supports this (Figure 3.12). However, similar to what was found with the linear regression model, Memphis is the only region where these patterns are consistent, suggesting it may be an outlier in regards to the number of intersecting tornado and flash flood warnings issued by its WFO.

## 3.4 Regional Forecast Differences (RQ3)

Now that conditional probability has been defined, conditional joint forecasts can be created and subsequently verified alongside the IJO forecast to asses performance. One of the main research questions in this work is if/how regions of significant correlation affect forecast performance and whether the difference in performance is significant. It's hypothesized that the conditional joint forecast methods will perform better in significantly correlated regions, and that the IJO forecast will perform better outside these regions. To test this, forecast reliability and BSS are assessed for each forecast method in each region (Correlated CWAs and Eastern US) (Figure 3.7) to determine if any significant differences exist. In addition to testing the hypothesis, reliability diagrams will be used to characterize the general trends in forecast calibration between methods.

### 3.4.1 Reliability and Sharpness

To determine if there is a significant difference in forecast calibration, reliability diagrams with confidence intervals are presented for each forecast method in both regions. If confidence intervals do not overlap for the majority of forecast probabilities, then there is said to be a significant difference in forecast calibration between the two regions.

Sharpness diagrams are presented along with each reliability diagram to characterize the relative frequency of forecasts within each probability bin. Although the SPC TO and WPC ERO forecasts are not calibrated for TORFF events, reliability diagrams are presented for these forecasts to illustrate the difference in their probabilities and relative frequencies to assist in the interpretation of joint forecast performance.

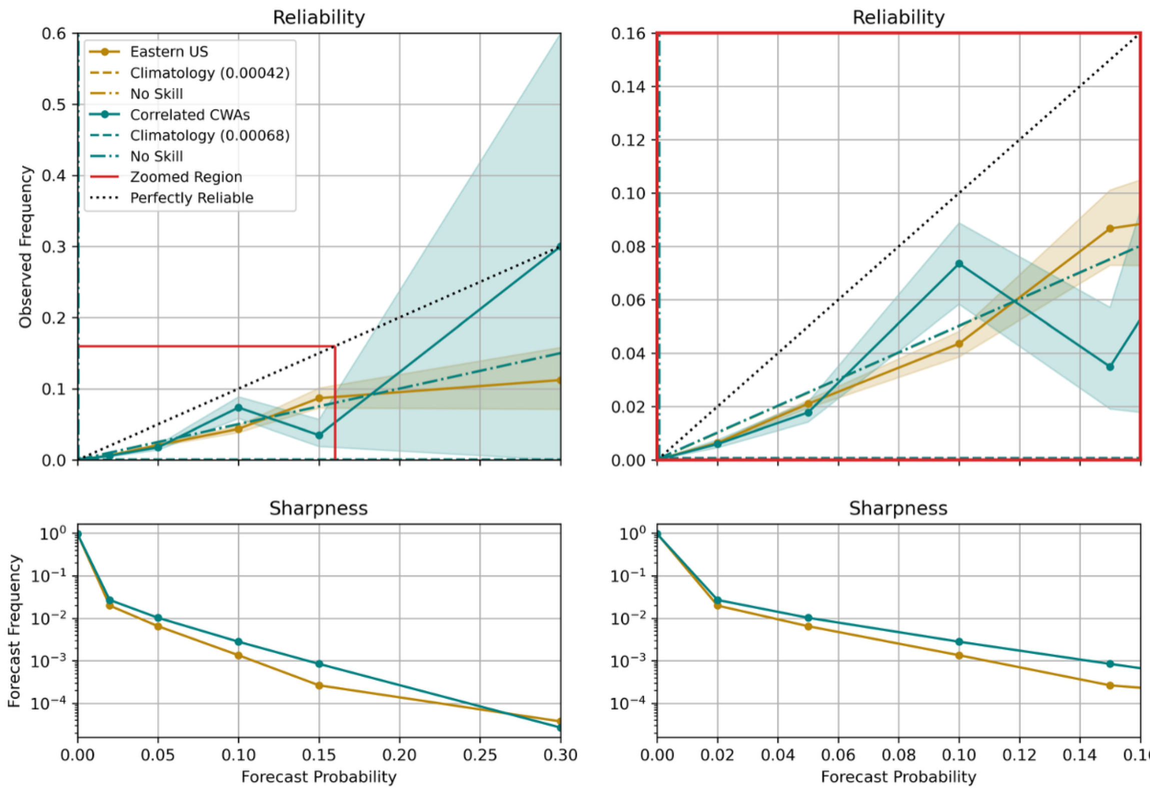


Figure 3.16: Reliability (Top) and sharpness (bottom) diagrams for the SPC TO in the Eastern US (tan) and Correlated CWAs (teal). 95% confidence intervals for observed frequency are shaded. The black dotted line indicates perfectly reliable, the dotted-dashed lines indicate the no-skill cutoffs for each region, and the dashed lines indicate the climatological frequency of TORFFs in each region (no resolution). The red box in the top right diagram indicates where the reliability diagram is zoomed in to show calibration at low-end probabilities. Reliability and sharpness diagrams on the left show the full range of probabilities, and diagrams on the right show only the zoomed-in probabilities.

The frequency of SPC TO probabilities (Figure 3.16) decrease monotonically as probability increases, and frequencies are slightly higher at all probabilities in the Correlated CWAs region except 0.3. Unsurprisingly, SPC TO significantly over-forecasts at all probabilities in both regions except at 0.3 in the Correlated CWAs region. At probabilities 0.02 and 0.05, the forecast is below the skillful forecast threshold for both regions. At 0.1, there is a significant difference in reliability between the two regions; reliability for the Correlated CWAs region improves and is within the skillful forecast threshold, whereas in the Eastern US reliability remains below the threshold. This then flips at 0.15 where reliability for Correlated CWAs falls below the threshold but in the Eastern US reliability improves but the confidence interval overlaps with the no-skill line. Given that confidence intervals overlap for the majority of thresholds, there is no significant difference in the calibration.

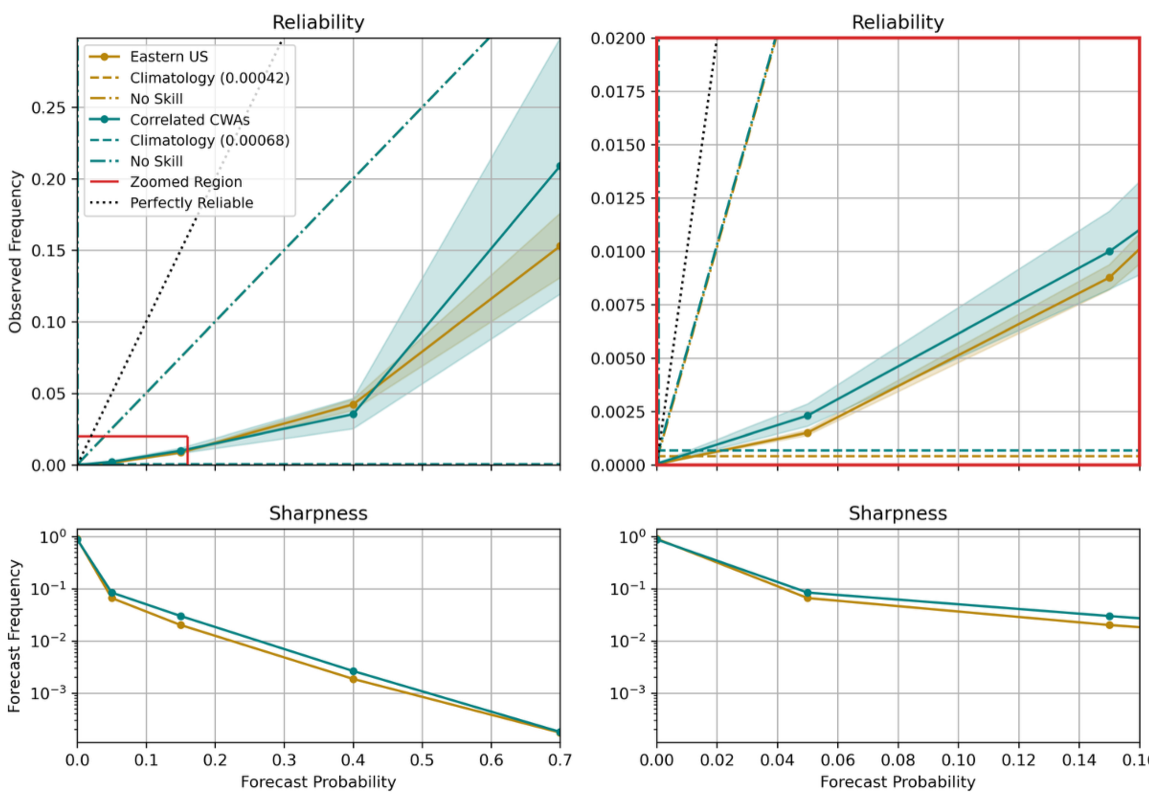


Figure 3.17: As in Figure 3.16 but for the WPC ERO.

Similar to SPC TO, the frequencies of WPC ERO probabilities decrease monotonically, and frequencies are slightly higher at all probabilities in the Correlated CWAs region except at 0.7 where they appear nearly the same (Figure 3.17). As expected, WPC ERO significantly over-forecasts at all probabilities and is significantly below the no-skill thresholds in both regions. At all probabilities except 0.05, there is no significant difference in reliability between the two regions. At 0.05, reliability is significantly better in the Correlated CWAs region, however given the incredibly small difference in observed frequencies (∼0.001), the effect on forecast performance would be negligible. Compared to SPC TO, observed frequencies and thus reliability are considerably lower due to the larger range of probabilities that the WPC ERO issues for flash flood guidance exceedance and the relative frequency at which these forecasts are issued.

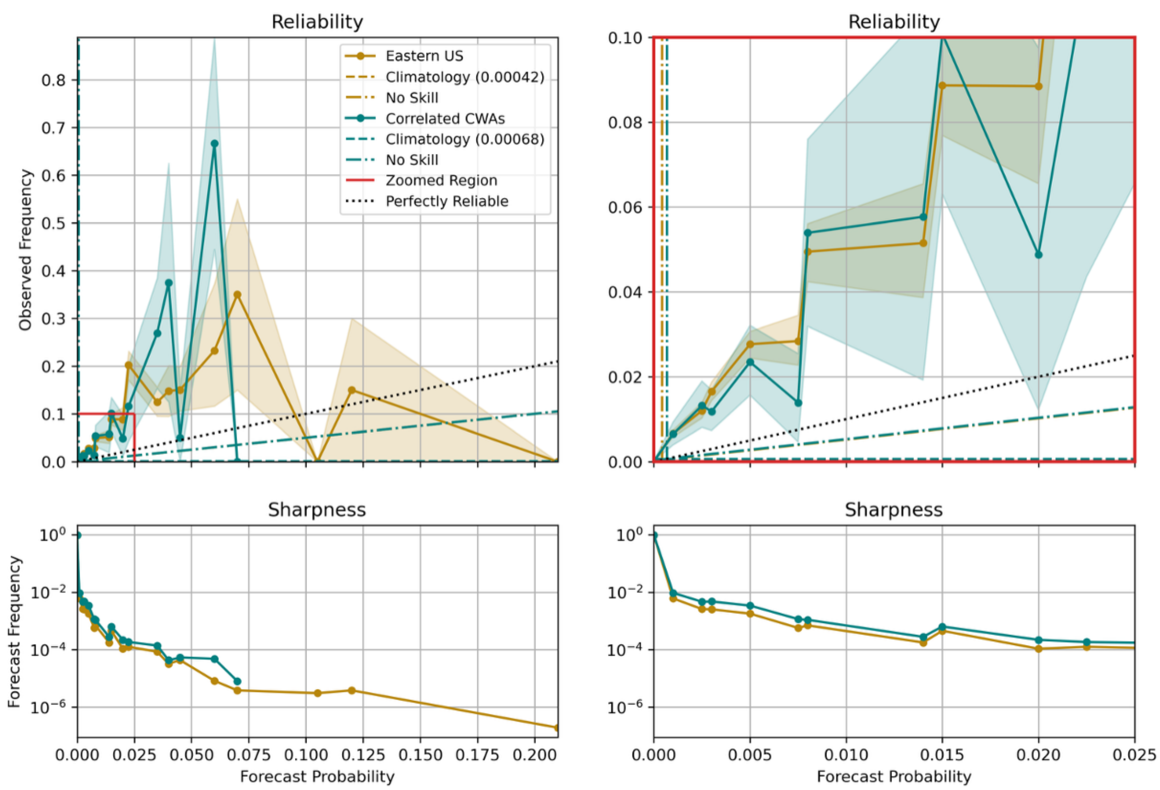


Figure 3.18: As in Figure 3.16 but for the IJO forecast.

The frequency of IJO forecast probabilities generally decrease with increasing probability except at 0.015 and 0.04, and the majority of probabilities are clustered between 0 and 0.025 (Figure 3.18). Between 0 and ∼0.04, the distribution of frequency in both regions is relatively the same except for slightly higher frequencies in the Correlated CWAs. At ∼0.06, frequency in the Correlated CWAs is noticeably larger, however the this could be due to sample size given the relatively low frequency at this probability for both regions. Furthermore, the maximum probability in the Eastern US (0.21) is more than double the maximum probability in Correlated CWAs (0.07). This difference is explained by the relative frequencies of SPC TO and WPC ERO forecast probabilities in which it becomes increasingly rare for higher probability contours to overlap with each other. It was determined that only a single day in the forecast period, 20 May, 2019, produced a forecast with a probability greater than 0.07 which occurred outside the Correlated CWAs region. This demonstrates the incredible rarity for higher end probabilities of the IJO forecast to occur and also explains the spikes in reliability which occur due to very low sample size of events within the forecast bins. In general, the IJO forecast significantly under-forecasts in both regions as a result of the high density of lower-end probabilities. At probabilities below 0.04, reliability for both regions is within the skillful forecast area. Above 0.04, reliability begins to spike in the Correlated CWAs, likely due to low sample size in the region, but generally remains significantly under-forecasted. Of the probabilities with forecasts in both regions (0.07 or lower), the majority of confidence intervals overlap, so no significant difference in calibration is observed.

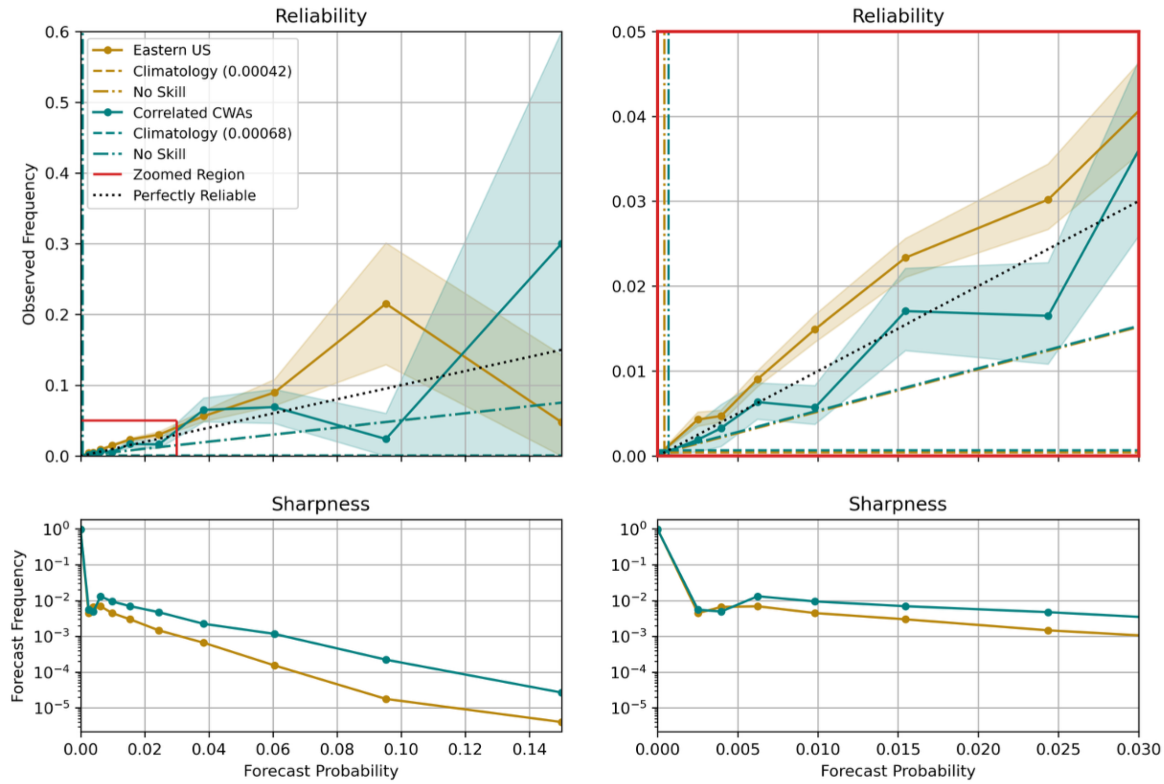


Figure 3.19: As in Figure 3.16 but for the FWC forecast.

At probabilities below ∼0.007, forecast frequencies for the FWC forecast are roughly equal between the two regions and frequency increases with probability until 0.007 where frequency decreases monotonically and is greater in the Correlated CWAs (Figure 3.19). Recall that the FWC forecast is the result of multiplying the SPC TO with $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ (Figure 3.8) which had an average probability of 0.38. Multiplying this average by the lowest SPC TO probability (0.02) yields 0.0076 which is roughly where the sharpness diagram indicates a peak in forecast frequency (above 0). In the Eastern US, FWC tends to slightly under-forecast at all probabilities except for 0.15 where low sample size likely skews observed frequency. In Correlated CWAs, reliability is somewhat improved with 6 out of 10 probabilities overlapping the perfectly reliable line. However, confidence intervals for these probabilities overlap with the Eastern US, so a significant difference in calibration overall is not observed overall. However, where confidence intervals do not overlap, there is significant over-forecasting in the Correlated CWAs that also overlaps with the no-skill line, suggesting that $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ may overestimate the conditional relationship in the correlated regions. Despite this, the FWC forecast has better overall calibration in the Correlated CWAs.

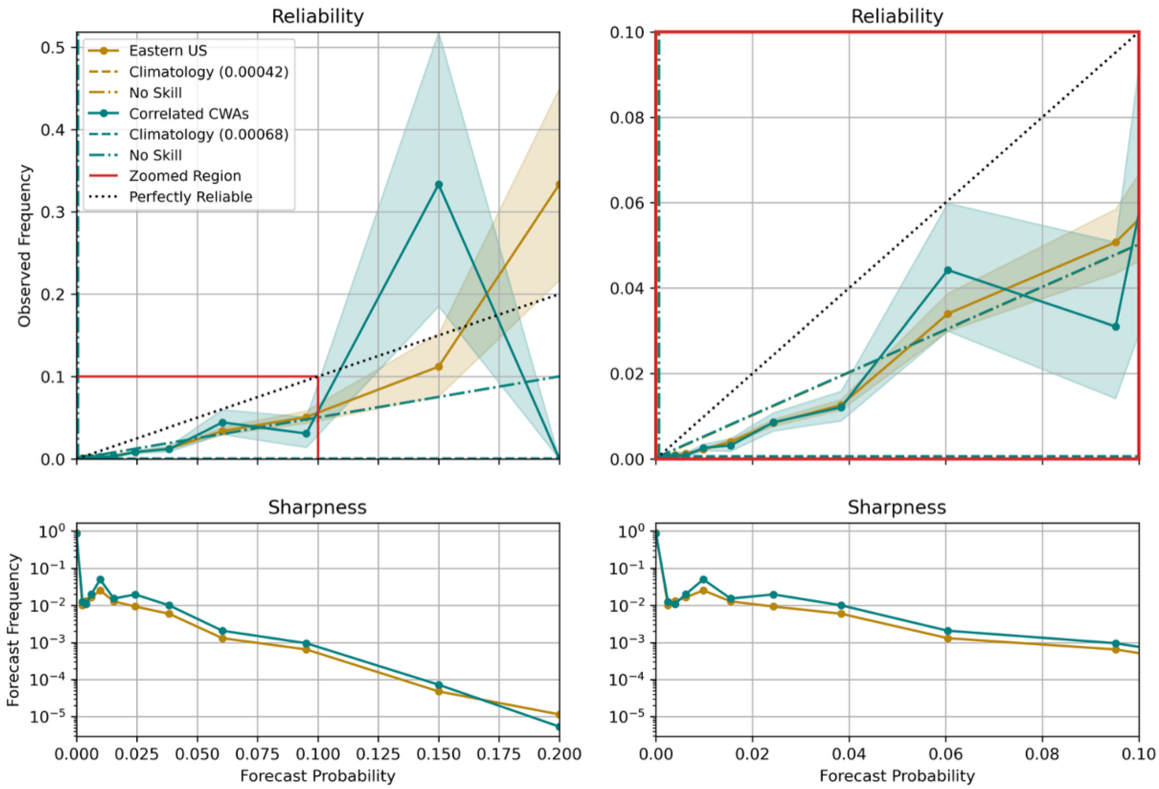


Figure 3.20: As in Figure 3.16 but for the TWC forecast.

For the TWC forecast, forecast frequencies initially increase until ∼0.01 then monotonically decrease, with slightly higher values for Correlated CWAs except at 0.2 (Figure 3.20). In a similar fashion to FWC, the peak in TWC forecast frequency at ∼0.1 is roughly equal to the lowest WPC ERO probability (0.05) multiplied by the average $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Warn}}$ (0.24) (Figure 3.11). For both regions at probabilities below 0.4, there is significant over-forecasting that is also below the no-skill thresholds. Between 0.04 and 0.1, over-forecasting is observed in both regions but slightly improved, however confidence intervals still overlap the no-skill line. There is a significant transition to under-forecasting at 0.15 in the Correlated CWAs and 0.2 in the Eastern US; however, confidence intervals are quite large, indicating low sample size. Below 0.1, all confidence intervals overlap, therefore there is no significant difference in calibration.

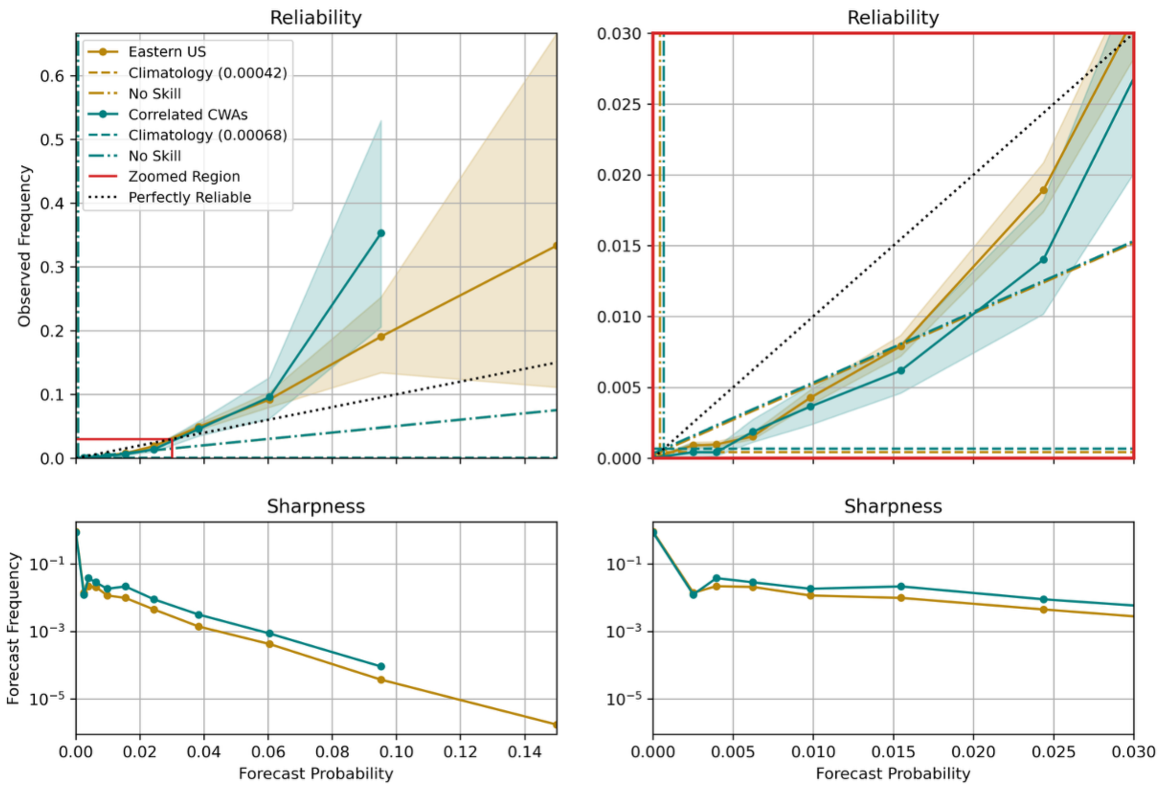


Figure 3.21: As in Figure 3.16 but for the WCA forecast.

The frequency of WCA forecast probabilities peak at roughly 0.003 and frequencies are slightly higher in Correlated CWAs at all probabilities except ∼0.0025 (Figure 3.21). At probabilities below 0.02 for both regions, there is significant over-forecasting that also falls below the no-skill line. Above 0.02, there is positive skill, and at ∼0.04, there is a shift to under-forecasting in both regions. Confidence intervals for both regions overlap at all probabilities except for 0.15, so there is no significant difference in calibration.

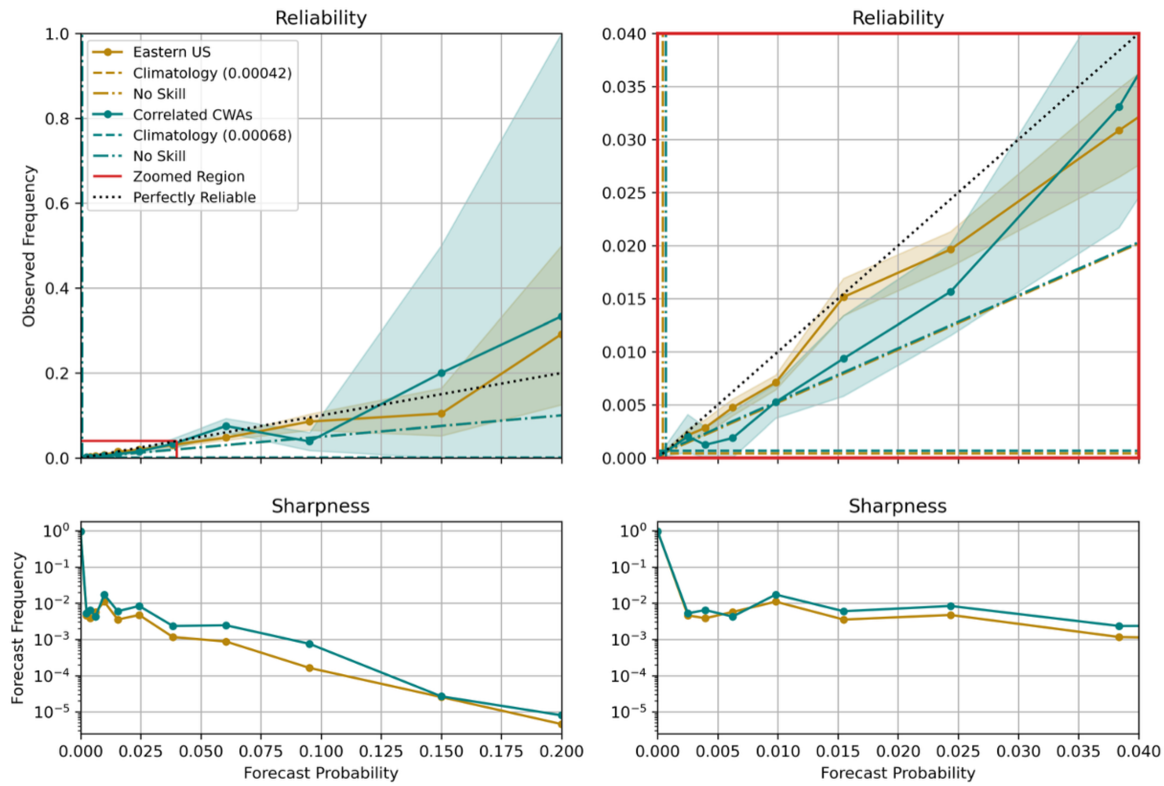


Figure 3.22: As in Figure 3.16 but for the FOC forecast.

The FOC forecast (Figure 3.22) has a peak in frequency at 0.01 which is approximately the lowest SPC TO value (0.02) multiplied by the average $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$ probability (0.65) (Figure 3.13). Frequencies are generally higher in the Correlated CWAs region except for ∼0.07, and there are relatively larger differences at ∼0.06 and ∼0.09. Reliability diagrams indicate a relatively well-calibrated forecast in the Eastern US, with confidence intervals overlapping the perfectly reliable line at 5 out of 11 probabilities. Furthermore, reliability is entirely withing the skillful forecast area except at 0.15 where the confidence interval slightly overlaps with the no-skill line. In Correlated CWAs, there is significant under-forecasting at probabilities below 0.03 and at ∼0.9, and over-forecasting at 0.15 and above. In general, the FOC forecast is better calibrated in the Eastern US, but confidence intervals for each region overlap at all probabilities, so no significant difference is observed.

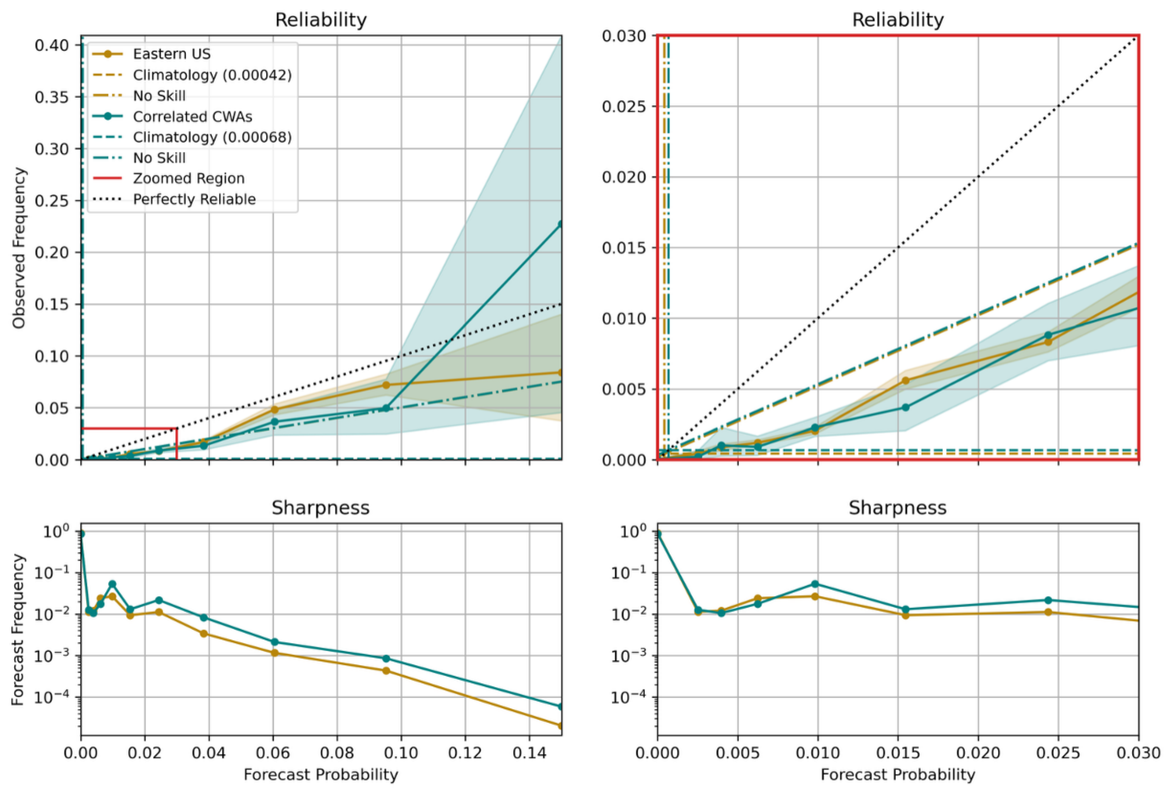


Figure 3.23: As in Figure 3.16 but for the TOC forecast.

Similar to the FOC forecast, the peak in frequency for the TOC forecast is at ∼0.01 (Figure 3.23). This is roughly equal to the minimum WPC ERO probability (0.05) multiplied by the average $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Otlk}}$ probability (0.21) (Figure 3.15). At probabilities greater than ∼0.07, frequencies are slightly greater in the Correlated CWAs. At probabilities below 0.1, there is significant over-forecasting in both regions, and at probabilities below 0.04, reliability is significantly below the skillful forecast thresholds in both regions. Above 0.1, under-forecasting is observed in Correlated CWAs, however there is very high uncertainty which makes this inconclusive. In general, forecast calibration is slightly better in the Eastern US, but confidence intervals for both regions overlap at all probabilities so no significant difference is observed.

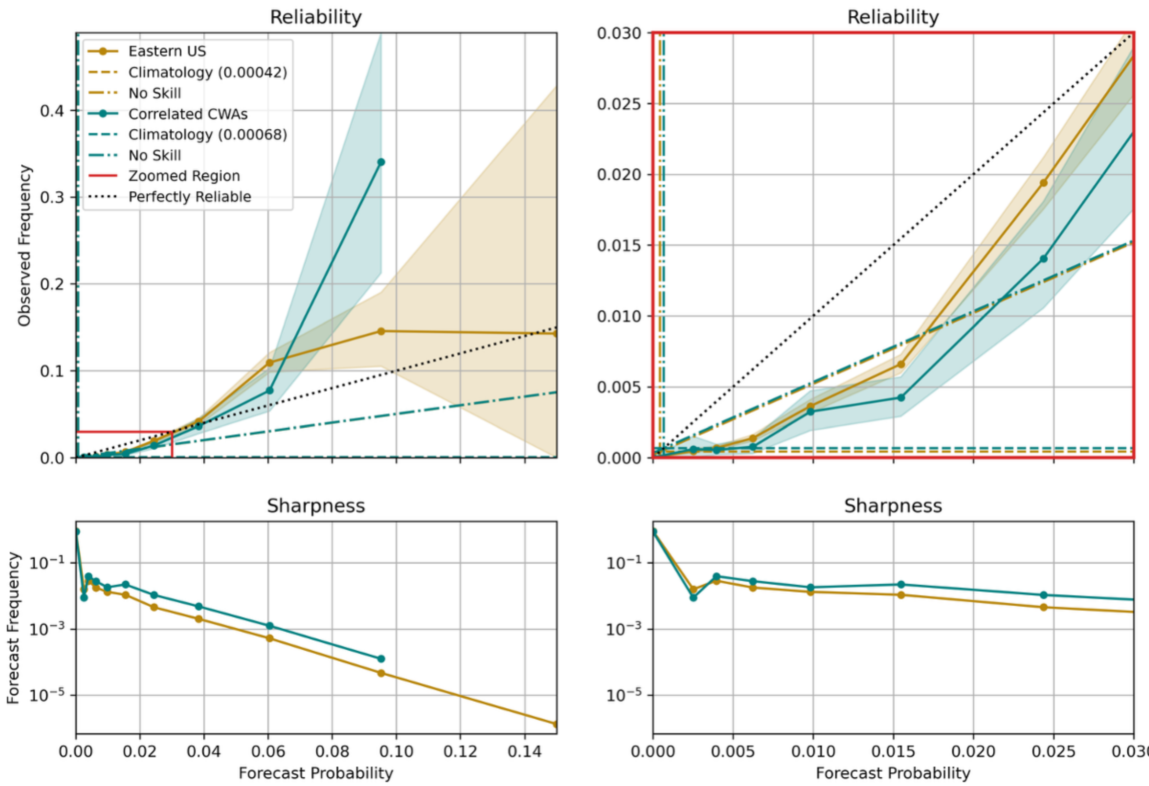


Figure 3.24: As in Figure 3.16 but for the OCA forecast.

Similar to the WCA forecast, the probability with peak frequency for the OCA forecast is ∼0.003 (Figure 3.24). Forecast frequencies are slightly higher for the Correlated CWAs at all probabilities except ∼0.025 where it is slightly lower. At probabilities below ∼0.016, there is significant over-forecasting, and reliability is below the no-skill threshold for both regions. Above 0.04, there is a shift to under-forecasting in both regions and reliability is within the skillful forecast area. Notably, this shift was also seen in the warning conditional average forecast at approximately the same probability. Confidence intervals overlap at the majority of probabilities, so no significant difference in calibration is observed.

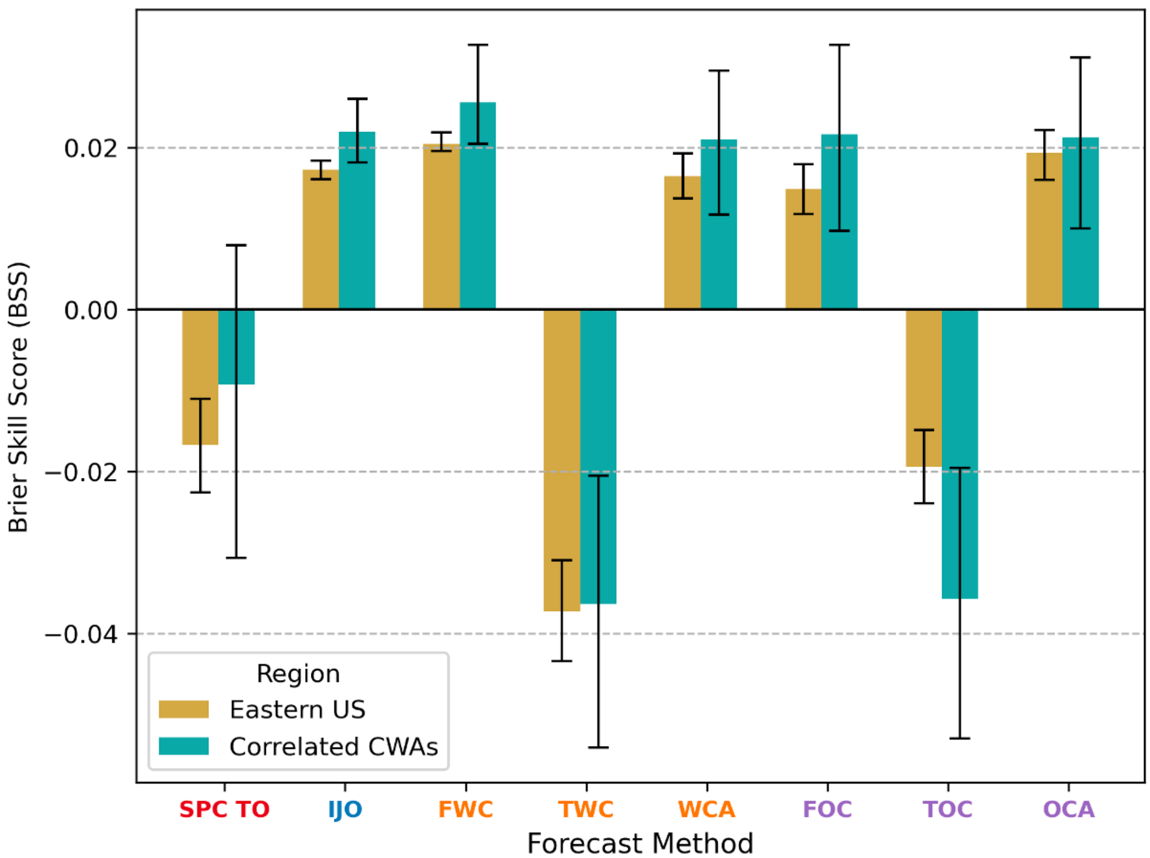


Figure 3.25: BSS for each forecast method in the Eastern US (tan) and Correlated CWAs (teal). Black bars indicate 95% confidence intervals.

### 3.4.2 Brier Skill Score

To further asses if the region of correlation has an impact on forecast performance, BSS is calculated with 95% confidence intervals for each region to determine if a statistically significant difference is observed. Figure 3.25 shows the results of these tests for each forecast method except for the WPC ERO which had a BSS of approximately -2 in both regions, skewing the axes of the figure and making forecast comparison unnecessary. All forecasts except for TOC have a higher BSS in the Correlated CWAs region, however none of the differences are significant. The TOC forecast has the largest difference in BSS between the two regions, while the TWC and OCA forecasts have the smallest difference. In both regions, the TWC and TOC forecasts have significantly negative BSSs that are also considerably lower than the SPC TO . Given that TWC and TOC significantly over-forecasted at nealy all probabilities, this result is not surprising.

### 3.4.3 Discussion

Relative forecast frequencies in the Correlated CWAs were higher at nearly all probabilities across all forecasts, but this is likely due to the smaller size of the region and its location in the central US where tornado and flash flood forecasts are more likely to be issued relative to the entire Eastern US. The best calibrated forecasts across both regions were the FWC and FOC which were within the skillful forecast threshold and closest to perfectly reliable at all probabilities. Notably, these two forecasts are the only ones derived exclusively from the SPC TO. Those derived from only the WPC ERO (TWC and TOC) exhibited significant over-forecasting at all probabilities and no skill at lower-end probabilities. This finding is likely explained by the higher frequency and larger range of WPC ERO forecast probabilities (Figure 3.17) which had much worse calibration than SPC TO on its own. Despite $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Warn}}$ and $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Otlk}}$

being much lower than $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ and $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$) on average, the range of WPC ERO probabilities and frequency with which they are issued are simply much too high and outweigh the lower conditional probabilities, resulting in significant over-forecasting. Interestingly, the conditional average forecasts were also sensitive to this over-forecasting at lower probabilities.

It was shown in section 3.3 that the difference between the average of $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ and $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$ (52%) was much greater than the difference between the average of $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Warn}}$ and $P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Otlk}}$ (13%). This explains why there was a greater difference in overall reliability between the FWC and FOC forecasts when compared to the TWC and TOC forecasts. Furthermore, the difference in reliability for the FWC forecast was notably larger between the two regions than any other forecast. This is likely due to the larger variance in $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ across CWAs (Figure 3.9), therefore resulting in over-forecasting in the Correlated CWAs where $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ was slightly higher than the rest of the Eastern US (Figure 3.8). The conditional average forecasts (WCA and OCA) showed very similar trends in reliability (Figures 3.21, 3.24) which suggests that averaging the conditional probabilities results in a similar forecast regardless of whether warning or outlook-scale probabilities are used.

BSSs were higher, albeit not significantly, in the Correlated CWAs region for all but the TOC forecast. This result is likely due to the higher climatological base rate of TORFF events in the Correlated CWAs region (0.00068) compared to the Eastern US (0.00042) which will inflate BSS regardless of whether the forecast was more skillful since BSS is shown to be quite sensitive to sample size and climatology (Murphy 1973; Bradley et al. 2008). The uncertainty in BSS is clearly reflected in the larger confidence intervals for the Correlated CWAs, making it challenging to draw any other conclusions from these results other than the lack of a significant difference.

Overall, none of the forecast methods exhibited significant differences in performance between the two regions, therefore disproving the hypothesis that there would be improved performance by conditional joint forecasts in the Correlated CWAs and by the IJO forecast in the Eastern US. There are a few theories as to why these results were observed. First, it was discussed in section 3.2 that regions of significant correlation may not be robust given the caveats with warning data and lack of correspondence with TORFF frequency or conditional probabilities. As a result, the region of correlation analyzed in this section may not fully represent where significant correlations truly exist. It may also be true that the Correlated CWAs region is a result of significant outliers in the warning data which was also mentioned as caveat of the linear regression model. A lack of truly significant correlation would therefore explain the lack of a significant difference in forecast performance. Therefore, future work is needed to asses tornado-flash flood correlation with more robust methods to validate the relationship. Nonetheless, these initial results demonstrate that regions of significant tornado-flash flood correlation are not significant enough to be considered in developing the TORFF framework moving forward.

Despite the disproven hypothesis, the FWC and FOC forecasts were considerably better calibrated in both regions compared to the IJO forecast which suggests that conditional probability is better at highlighting the true probability of TORFF occurrences. This also supports the idea that tornadoes and flash floods may be correlated on a larger scale than the results of the linear regression model initially identified given the successful calibration of FWC and FOC in both regions. Since FWC and FOC were better calibrated than all other conditional joint forecasts, it's evident that TORFF environments are better represented by typical tornado environments with regionally higher flash flood risk as given by the conditional probabilities. This agrees with previous work demonstrating the strong similarities of TORFF environments to typical tornadic environments with higher levels of moisture or where flash flooding is simply more common (Nielsen et al. 2015; Burow et al. 2025; Ellis et al. 2023). More discussion on the differences in forecast performance will be presented in the following sections.

## 3.5 Comparing Forecast Performance (RQ4)

Since no significant difference was found in forecast performance between the two regions, the remaining forecast analysis in this study will consider only a single collective region that combines the Eastern US with the Correlated CWAs to compare overall performance. Although it was already determined that FWC and FOC forecasts had better calibration, further analysis into forecast skill across a collective region can provide further insight as to why this may be the case and if it holds true across seasons and time periods.

### 3.5.1 Brier Skill Scores

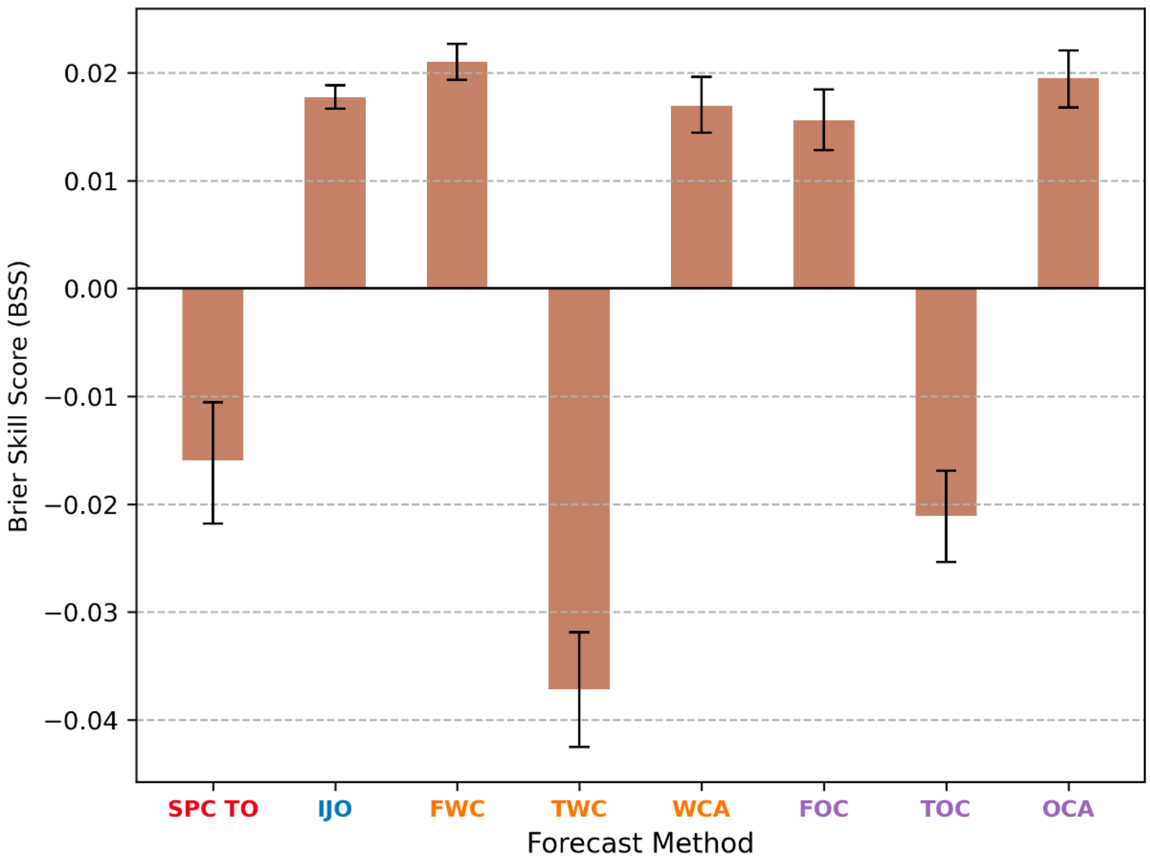


Figure 3.26: As in Figure 3.25 but for the entire Eastern US (including Correlated CWAs)

Similar to what was seen in Figure 3.25, the SPC TO, TWC, and TOC forecasts have significant negative BSS in the entire Eastern US (Figure 3.26), indicating no improvement over forecasting climatology. The remaining forecast methods all have a significant positive BSS indicating improvement over climatology, although quite small. The FWC forecast has the highest overall BSS of 0.021. The next highest is the OCA forecast with a BSS of 0.0195, and third is the IJO forecast with a BSS of 0.018. Confidence intervals indicate that the BSS of the FWC forecast is significantly higher than the IJO forecast, but not the OCA forecast. Furthermore, confidence intervals for the WCA and FOC forecasts, which also have positive BSS, overlap with the top three methods. Comparing warning vs outlook conditional joint forecasts, the FWC forecast performs significantly better than FOC, but the TWC forecast performs significantly *worse* than TOC. The OCA forecast performs slightly better than WCA forecast, but not significantly. Given these results, further analysis into the seasonal and yearly performance of forecasts is warranted to provide further insight into performance patterns.

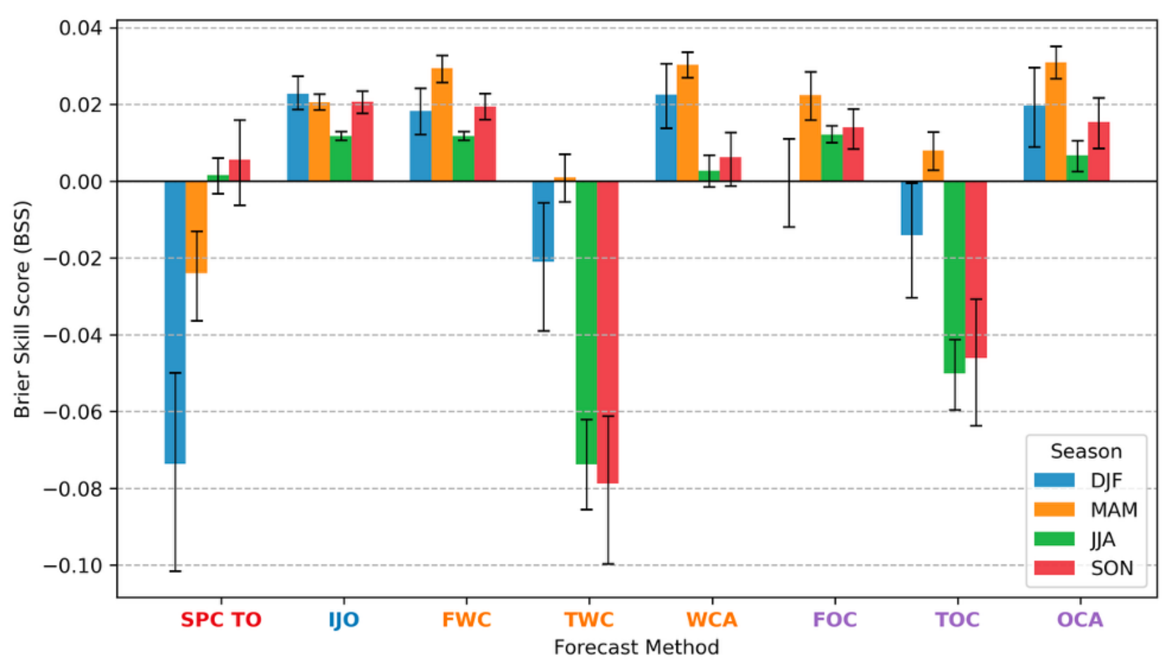


Figure 3.27: BSS by meteorological season for each forecast method in the entire Eastern US (including Correlated CWAs) with 95% confidence intervals (black bars).

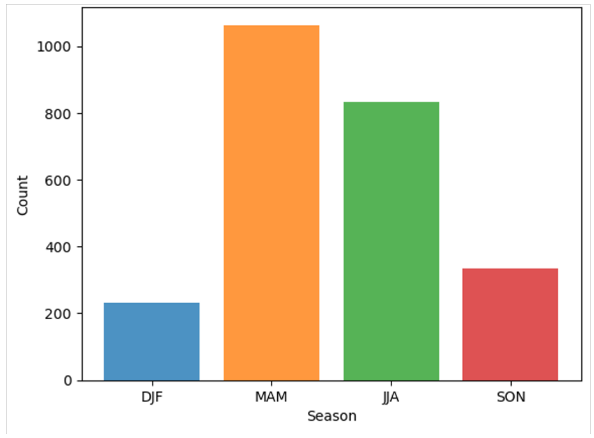


Figure 3.28: Distribution of TORFF events for each meteorological season

Figure 3.27 illustrates the difference in BSS for each forecast method in the entire Eastern US (including Correlated CWAs) as a function of meteorological season, and Figure 3.28 shows the distribution of TORFF events within each season. Unsurprisingly, the majority of TORFF events occur in the spring months (MAM) (∼1050) when tornadoes are most frequent (Figure 3.28). There are also a considerable amount of summer (JJA) TORFFs, but approximately 200 less than spring. Winter (DJF) has the fewest TORFFs at just above 200, and fall (SON) takes third with ∼300. In general, the best forecast season is spring with 6 out of 8 forecasts having a significantly positive BSS. For the FWC, TWC, and TOC forecasts, spring BSS is significantly higher than all other seasons. For the TWC and TOC forecasts, spring is the *only* season with a positive BSS. Notably, these two forecasts also have the lowest BSS of all forecasts for summer, fall, and, if excluding SPC TO, winter. The WPC ERO (not shown) also had the best BSS in spring by a significant margin. In the winter, the IJO, FWC, WCA, and OCA forecasts have positive BSS, none of which are significantly higher than all other seasons. The SPC TO notably has the worst performance in the winter which is significantly lower than all other seasons. A particularly interesting pattern is seen in summer as it is the lowest performing season for the IJO, FWC, WCA, TOC, and OCA forecasts. This finding is especially puzzling given that the preceding season has the best performance for all but one of these forecasts and there is no particular pattern in the methods used among these forecasts. Finally, fall has relatively average performance among forecasts, with 4/8 forecasts having significant positive BSS (IJO, FWC, FOC, OCA). A concluding insight from Figure 3.27 is that only the IJO, FWC, and OCA forecasts have positive BSS across all seasons, which is notable given that these were the top three forecasts in terms of overall BSS (Figure 3.26).

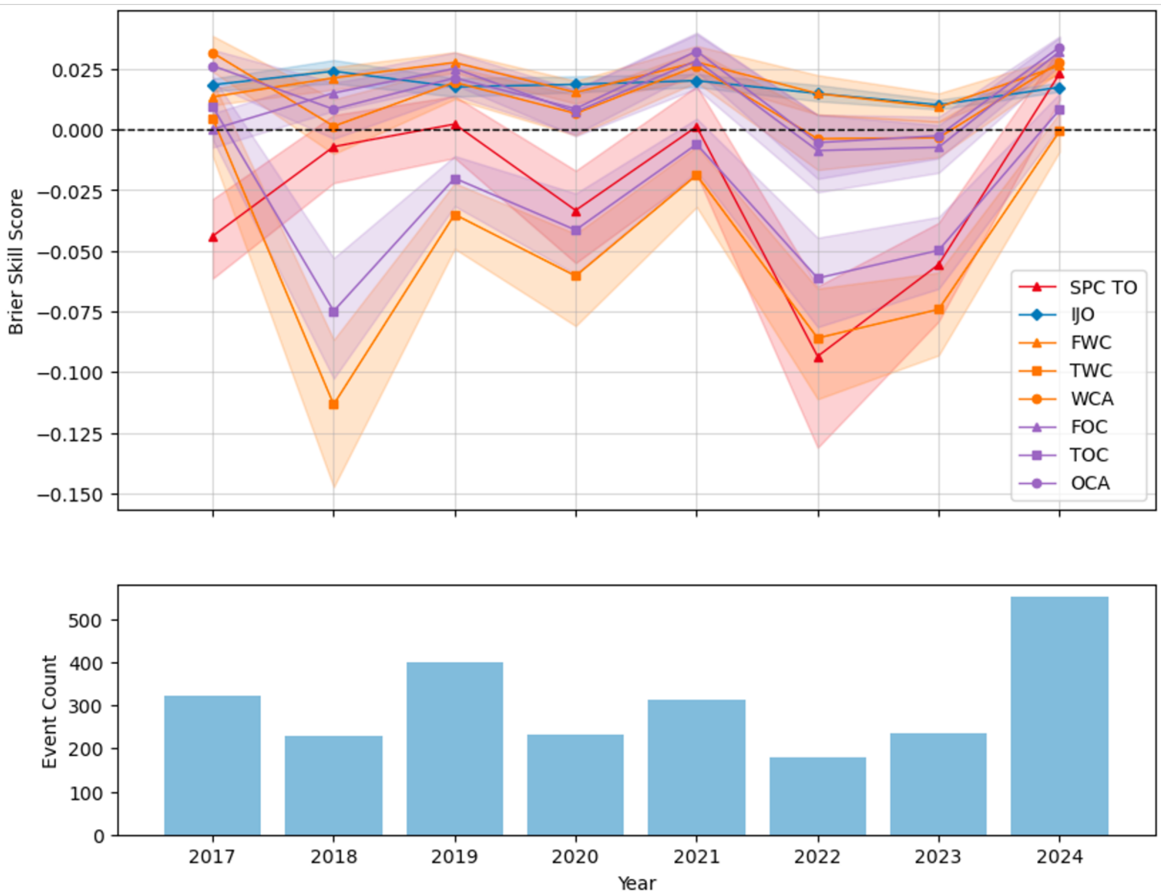


Figure 3.29: BSS by year for each forecast method (top) with 95% confidence intervals (shaded) and yearly TORFF event count (bottom)

Figure 3.29 shows the annual BSS for each forecast method as well as the distribution of TORFF events for each year. In general, BSS is higher among all forecasts when there is a peak in TORFF events, especially in 2024 which has the highest TORFF rate and nearly all forecasts have a positive BSS. BSSs for the IJO and FWC forecasts vary the least across the forecast period relative to the other forecasts and are the only two forecasts to maintain a positive BSS across all years. The TWC and TOC forecasts appear to have the most sensitivity to sample size given the considerably sharp decline in BSS in 2018 and 2022 which have the lowest event counts. During the years with the top three highest TORFF counts (2019, 2021, and 2024), all conditional joint forecasts besides TWC and TOC perform better than the IJO forecast (although not always significantly), and the opposite occurs (IJO is best) during the years with the top three lowest TORFF counts (2018, 2020, 2022).

### 3.5.2 Performance Diagrams and CSI

As a final method of forecast evaluation, performance diagrams for each forecast are shown in Figure 3.30 along with the CSI at each probability threshold in Figure 3.31. Due to the extreme rarity of TORFF events and the outlook-scale forecast area, a significant amount of false positives occur across forecast grids. Consequently, the success ratio is very low and performance diagrams indicate very low skill. Despite this, performance diagrams are still useful to compare the performance of forecasts relative to each other to determine which have a better tradeoff between missed event and false alarms.

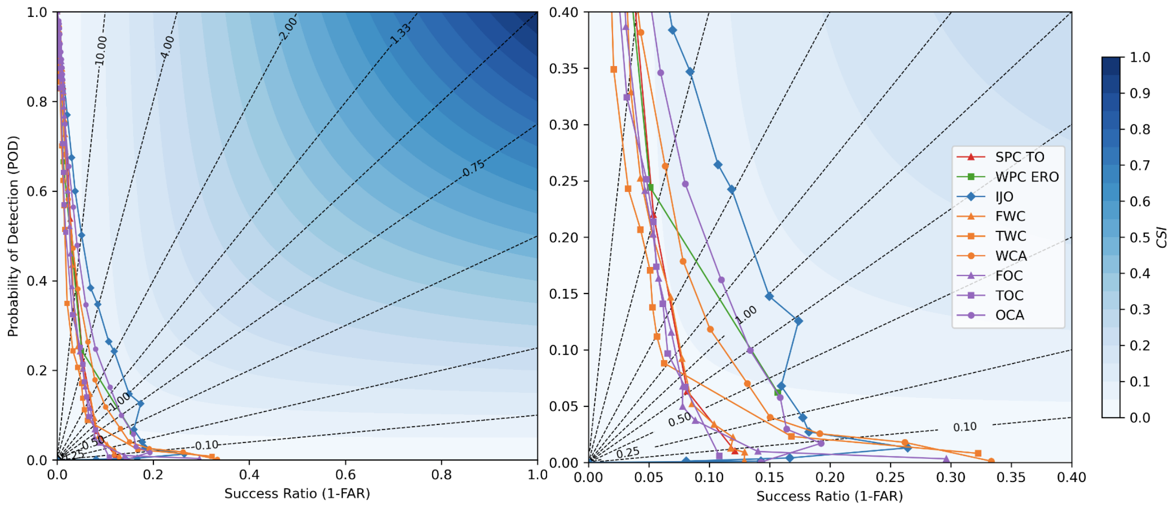


Figure 3.30: Performance diagrams for each forecast method in the entire Eastern US (including Correlated CWAs). Black dashed lines indicate forecast bias values, and blue shading indicates CSI. The right diagram is a zoomed-in version of the left diagram focusing on POD and SR from 0-0.4.

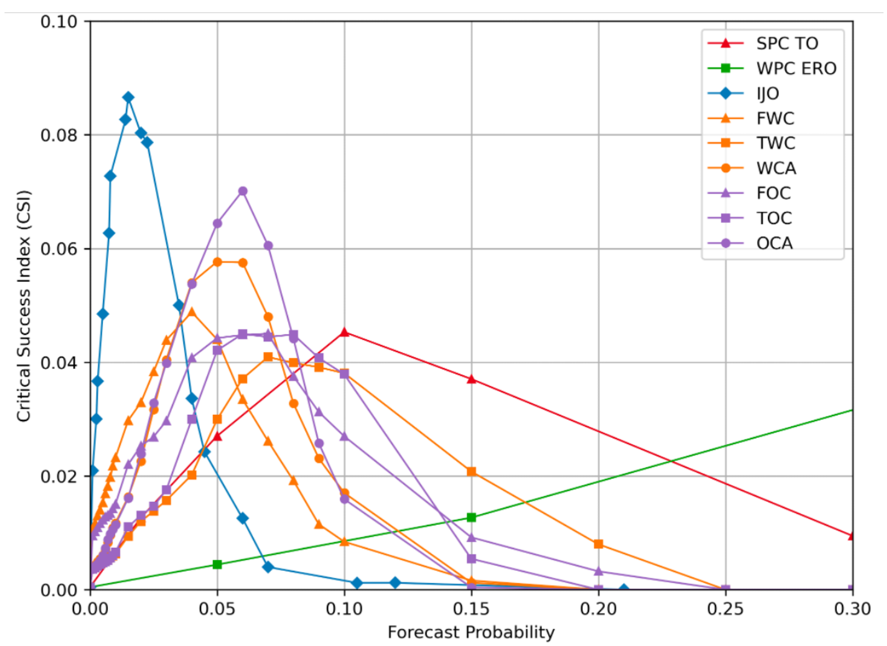


Figure 3.31: CSI values for each forecast method in the entire Eastern US

As expected, the majority of SRs in Figure 3.30 are limited to the 0.35 and below which indicates high false alarm rates among all forecasts. Furthermore, all forecasts demonstrate a general increase in SR as the probability of detection (POD) decreases. This indicates that as the probability threshold increases, the size of the forecast area significantly decreases which reduces false alarms while also increasing the number of missed events. Performance is generally maximized at PODs between 0.05 and 0.4 for all forecasts which will be the main focus of the analysis here. At these PODs, the IJO forecast has the highest overall SRs between ∼0.07 and ∼0.17 (Figure 3.30) indicating it has the best overall spatial discrimination. The IJO forecast also has the highest overall CSI value of ∼0.09 (Figure 3.31). The OCA forecast has the second best overall performance with SRs between ∼0.06 and ∼0.16 in the relevant POD range with a maximum CSI of ∼0.07. The WCA forecast follows behind with SRs ranging from ∼0.05 to ∼0.14 and a maximum CSI of ∼0.058. The TWC forecast has the worst overall performance with SRs ranging from ∼0.025 to ∼0.06 in the relevant POD range, and a maximum CSI of ∼0.04. The FWC, FOC, TOC, and SPC TO forecasts are all somewhat clustered with SRs between ∼0.03 and ∼0.07 and have maximum CSI values ranging from ∼0.045 to ∼0.05. The WPC ERO is also clustered with these forecasts with the exception of its maximum probability (0.7) which has an SR of ∼0.16 and a maximum CSI of 0.047 (not shown, occurring at maximum probability). With the exception of the WPC ERO, the distributions of CSI (Figure 3.31) for each forecast method appear to follow a log-normal or gamma distribution given the long tails extending to the right. In general, it appears that forecasts with a higher density of lower-end probabilities have higher peaks in CSI. This trend is supported by the sharpness diagrams for the IJO, WCA, and OCA forecasts (Figures 3.18, 3.21, 3.24) which generally showed more clustering and peaks in relative frequency at lower-end probabilities.

### 3.5.3 Discussion

Overall, BSSs for the collective Eastern US (Figure 3.26) indicated that the IJO, FWC, FOC, WCA, and OCA forecasts all had a significant improvement over climatology (BSSs 0.015-0.02). Expectedly, the SPC TO and WPC ERO had significantly negative BSS which makes sense given they are not calibrated to TORFF events. The WPC ERO (not shown) had significantly lower BSS than the SPC TO, likely due to its larger ranger of probabilities and the frequency at which it is issued compared to SPC TO, making it much less calibrated. As a result, the conditional joint forecast methods derived from it (TWC and TOC) had significant negative BSS (Figure 3.26) indicating no improvement over climatology. These results align with the reliability diagrams in the previous section which showed the TWC and TOC forecasts to have the most significant over-forecasting in both regions (Figures 3.20, 3.23). Interestingly, overall BSS for the TWC forecast was significantly lower than the TOC forecast (Figure 3.26), suggesting that warning-scale conditional probability for tornadoes ($P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Warn}}$) is less accurate than outlook-scale ($P(\mathrm{TOR}|\mathrm{FF})_{\mathrm{Otlk}}$). The opposite trend was observed for the overall BSS of flash flood conditional forecasts (FWC and FOC), suggesting the warning-scale conditional probability for flash floods ($P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$) is generally more accurate than outlook-scale ($P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Otlk}}$).

When BSS was analyzed by season, spring was shown to have the highest BSS across the majority of forecast methods (Figure 3.27). Given the frequency of TORFFs in spring, this result makes sense as it simply more likely for any TORFF forecast to overlap with events when they are most frequent. This finding also explains why the TWC and TOC forecasts only have a positive BSS in spring; the WPC ERO these forecasts are derived from is likely to over-forecast in the less-tornadic seasons where flash flood risk remains high such as late summer and early fall (Schumacher and

Johnson 2006; Dougherty and Rasmussen 2019), whereas in spring, it is more likely to overlap with tornadic environments by chance. The TWC and TOC forecasts also have the lowest overall BSSs for summer and fall which supports this theory. Notably, this pattern is not observed for the FWC or FOC forecasts (which are derived from the SPC TO), further supporting the conclusion that TORFF environments are better represented by forecasts for typical tornadic environments. Of the three forecasts that had significant positive BSS across all seasons (IJO, FWC, and OCA), summer was the worst performing season. This result indicates that summertime TORFFs are typically not as well represented by any of the forecast methods. The reason for this is not immediately clear, but there could be a particular convective mode of TORFFs that's more common in the summer that forecasts are not as skillful in highlighting. Previous work has not explicitly shown the distribution of TORFF storm mode across seasons, so this could be an avenue for future research.

Results of analyzing BSS by year demonstrated that the majority of conditional joint forecasts are sensitive to event count/sample size (Figure 3.29). Only the IJO and FWC forecasts maintained a positive BSS across all years even when event count was low, suggesting they are much less sensitive to smaller sample sizes and have better forecast performance overall. This finding also reinforces what was found in the seasonal analysis. The TWC and TOC forecasts showed the greatest variance and lowest overall BSS across years (Figure 3.29) which further confirms that the WPC ERO poorly represents TORFF environments during periods with low event count. It was also determined that the WPC ERO is issued with uniform frequency across all years in the forecast period, demonstrating a lack of consistency with the yearly fluctuations in TORFF frequency. Most interesting is the fact that the FWC, FOC, WCA, and OCA forecasts performed better than the IJO forecast in the three years with the highest TORFF count. This result may suggest that conditional joint forecasts (except for TWC and TOC) tend to have a more accurate *distribution* of probabilities that are capable of highlighting TORFF probability with better discrimination, but the frequency with which these probabilities are issued and/or the size of the forecast area is not as well calibrated and is therefore more sensitive to sample size.

Analysis of performance diagrams for each forecast revealed that the IJO forecast best minimized false alarms and maximized CSI (Figures 3.30, 3.31), therefore having the best spatial discrimination of any forecast. The WCA and OCA forecasts also demonstrated considerable skill relative to the other forecasts. Notably, these three forecast methods are the only which utilize both operational outlooks to derive the forecast, suggesting a better overall tradeoff between missed events and false alarms when both outlooks are integrated in some way. This finding may be explained by the average size of each forecast whereby a larger forecast area will naturally have a larger rate of false alarms, decreasing SR. The three methods that combine both operational outlooks likely have increasingly smaller forecasts at each threshold since there will typically be less spatial coverage in the outlook overlap as opposed to the individual outlooks used by conditional joint forecasts. The size of forecasts was not investigated in this study, so this remains speculative. It may also be true that the spatial discrimination of these forecasts is simply better due to the combined skill of the operational outlooks, not necessarily due to smaller area. Nonetheless, these results support that combining the SPC and WPC outlooks results in the best spatial discrimination for TORFF events, albeit not the best calibration/reliability which was shown to be relatively poor for these three forecasts (Figures 3.18, 3.21, 3.24).

In summary, the best performing TORFF forecast methods in the entire Eastern US in terms of BSS are the IJO and FWC forecasts which had significantly positive BSS across all seasons and years within the forecast period. The 3 other conditional joint forecasts that had significantly positive BSS overall (FOC, WCA, and OCA) were shown to have more sensitivity to sample size and were therefore less skillful. The IJO forecast also had the best spatial discrimination, but was shown to significantly under-forecast, while the FWC forecast was the opposite with poor spatial discrimination but much better calibration. Given that these two forecasts had very similar overall BSS, a potential user of these forecasts must decide whether to prioritize forecast calibration (FWC is best) or minimizing false alarms (IJO is best). The middle ground option would be the OCA forecast which, although had negative BSS in 2022 and 2023 (Figure 3.29), had the second best spatial discrimination and had positive BSS across all seasons. This method demonstrated significant over-forecasting at lower-end probabilities and under-forecasting at higher-end which suggests strong forecast skill at the mid-range which may be beneficial to some users. Previous TORFF research has largely emphasized situational awareness and hazard localization (Ellis et al. 2023; First et al. 2022; Henderson et al. 2020), suggesting that improvements in spatial discrimination may be more operationally meaningful than modest improvements in calibration alone.

The most significant caveat with the analysis of TORFF forecasts is the extreme rarity of the events, even given that the intersecting warning dataset is used for verification over LSRs which overestimates TORFF occurrences. CSI and BSS values were very low relative to what is seen in forecast verification for the individual hazards (Hitchens and Brooks 2017; Herman et al. 2018; Erickson et al. 2021), however, both CSI and BSS are shown to be sensitive to small sample sizes which explains these low values. As a result, it's very difficult to determine a baseline for what a skillful forecast for TORFFs would be using these statistics. Once such method that's been used in pervious literature is the practically perfect hindcast (Hitchens et al. 2013) which has been used in the verification of SPC convective outlooks to develop a baseline for CSI. Unfortunately, when this method was tested in this work, it was very difficult to tune to the extreme rarity of TORFFs and did not generate meaningful forecast contours for comparison. Therefore, it remains unclear whether the CSI and/or BSS values generated by these forecasts are significant relative to the broader field of forecast verification for convective hazards. Despite this, joint forecast methods generally showed significant improvement over using the SPC TO and WPC ERO alone in terms of reliability, improvement over climatology, and spatial discrimination. This suggests there is still significant utility in these forecasts to highlight where TORFF conditional are favorable rather than relying on previously existing methods.

## 3.6 RFML Model (RQ5)

An RFML model was designed and tested in this study to discern whether it is capable of highlighting TORFF events with sufficient calibration and how it's spatial discrimination compares to the joint forecasts across test cases. Cross validation segment 3 (Figure 2.4) produced the highest overall BSS (0.04), so it is used as the final model to generate test case forecasts with operational GEFS data, and its testing period is used to asses model calibration.

### 3.6.1 Calibration

Reliability and sharpness diagrams for the testing/evaluation period of segment 3 are presented to asses model calibration. Since this evaluation period is different from the period used to evaluate the joint forecasts in the previous sections, reliability and sharpness cannot be directly compared with the joint forecasts. Regardless, these results are useful in assessing the general performance of the model and to assist in the interpretation of test cases.

Relative forecast frequencies for the RFML model decrease monotonically and rel-

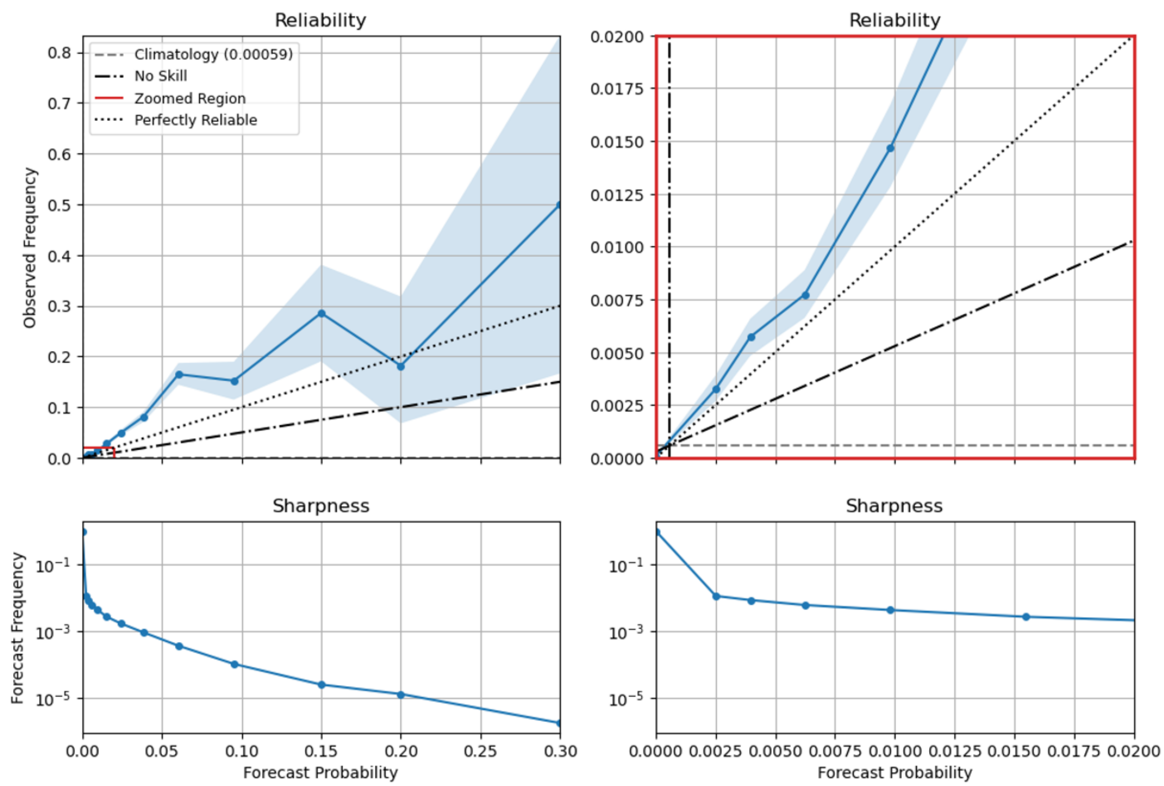


Figure 3.32: As in Figure 3.16 but for cross validation segment 3 of the RFML model.

atively smoothly with increasing probability (Figure 3.32). The model exhibits significant under-forecasting for all probabilities at and below 0.15. At very low probabilities (0 to ∼0.006) the model is somewhat better calibrated, but there is still significant under-forecasting. At ∼0.2, the model slightly over-forecasts, but the low frequency ($\sim 10^{-5}$) of this probability and large confidence interval suggest low sample size. At 0.3, the forecast once again under-forecasts, but with an even larger confidence interval due to low sample size. Despite significant under-forecasting, reliability at all probabilities except for 0.15 is within the skillful forecast area.

### 3.6.2 Test Cases

#### 3.6.2.1 Test Case 1: 1 September, 2021

On 1 September, 2021, the remnants of hurricane Ida struck the northeast coast, bringing significant rain and multiple tornadoes to the region (Zhu et al. 2022). The majority of TORFF events occurred in a clustered line from northeast Maryland into central New Jersey (Figure 3.33). There was also a smaller cluster located along the southeast coast of Massachusetts. The SPC TO forecast area (Figure 3.33a) covers nearly all of the events but trends slightly south of the main cluster. As a result, the FWC and FOC forecasts (Figure 3.33d,g) also trend slightly south of the cluster but have higher resolution that better highlight the area of northeast Maryland, north Delaware, and southwest New Jersey. The WPC ERO (Figure 3.33b) similarly covers the majority of TORFF events but trends too far north such that the majority of events boarder the south end of the 70% contour. This trend is reflected in the TWC and TOC forecasts (Figure 3.33e,h) which have forecast maxima north of the cluster in east-central Pennsylvania. As expected, the TWC and TOC forecasts also have significantly higher probability than the other conditional outlooks. The WCA and OCA forecasts (Fig-

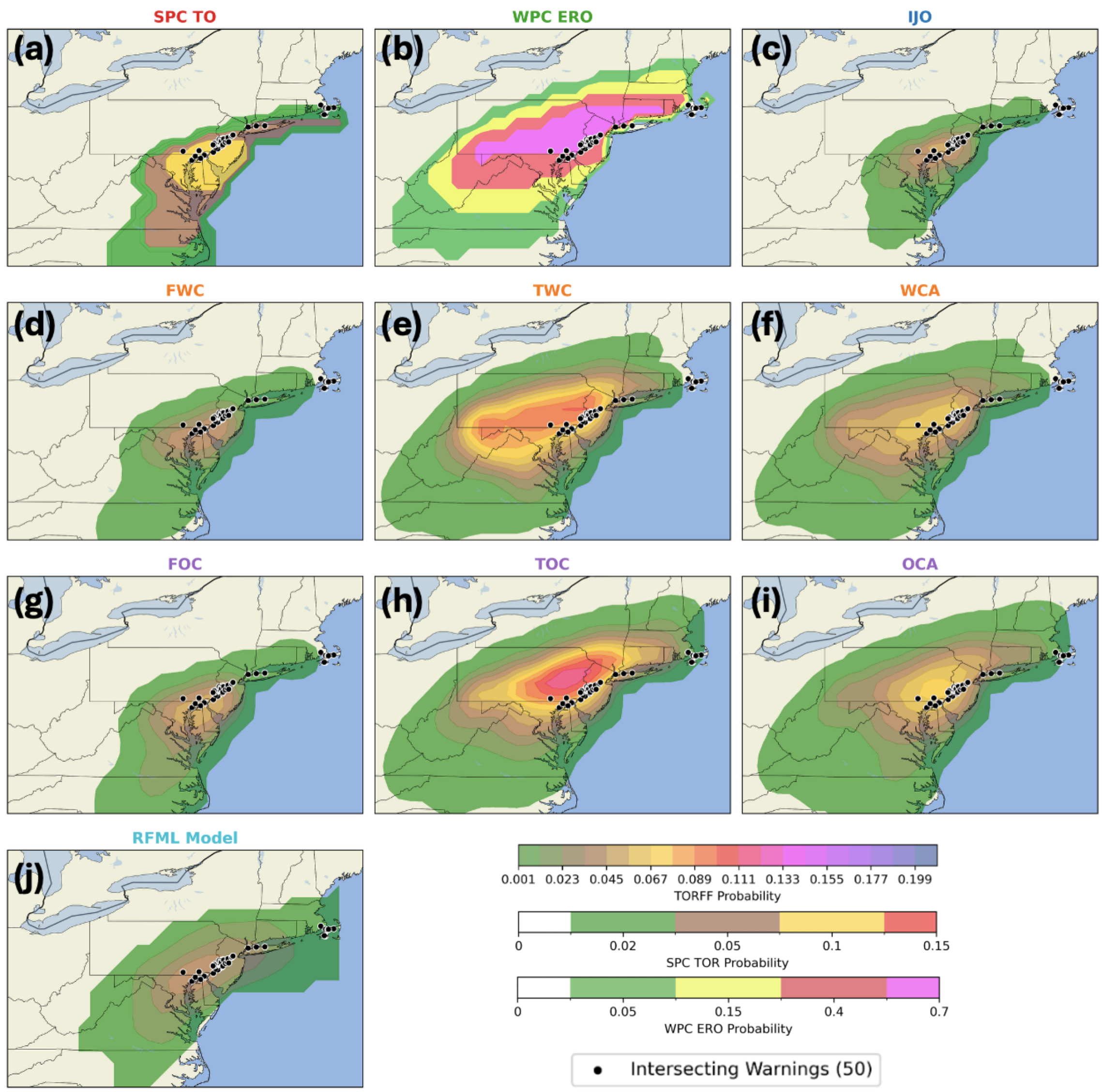


Figure 3.33: Forecast probabilities for each forecast method (shading) and coordinates of intersecting warnings (black dots) valid 1200 UTC September 1, 2021 to 1200 UTC September 2, 2021

ure 3.33f,i) have the best spatial discrimination of the conditional joint forecasts, with forecast maxima just to the north of the event cluster in southeast Pennsylvania. The range of probabilities between the WCA and OCA forecasts is roughly the same, however the lowest probability contour for OCA extends to the cluster in Massachusetts. The IJO forecast (Figure 3.33c) has the most conservative forecast area and still covers most events except those in Massachusetts. The IJO forecast also has the best spatial discrimination with a forecast maximum well aligned with the cluster over northeast Maryland/Southeast Pennsylvania, although it is fairly small at the highest probability. Finally, the RFML model (Figure 3.33j) has comparable spatial discrimination to the IJO forecast, with a large forecast maximum over northeast Maryland/Southeast Pennsylvania that captures a large majority of the cluster. The forecast area is slightly larger than the IJO forecast so it is not as precise, but the area extends to the cluster in Massachusetts at the lowest probability capturing more events. Probabilities of the machine learning model are also the lowest overall which aligns with the model's tendency to under-forecast (Figure 3.32).

#### 3.6.2.2 Test Case 2: 27 April, 2024

For the 27 April, 2024 case, TORFF events were scattered from north Texas into west-central Missouri (Figure 3.34). There was a notable cluster of events in south-central Oklahoma where a strong MCS/supercells produced significant flash flooding and an EF4 tornado in Marietta, OK (NOAA 2024). The SPC TO forecast area (Figure 3.34a) covers all events with only a few outside the 15% contour in Missouri and Nebraska. The WPC ERO (Figure 3.34b) similarly covers all TORFF events but is narrower and more slanted to the right. Unlike test case 1, there is a significant difference in the FWC and FOC forecasts (Figure 3.34d,g); the FWC forecast has considerably lower probabilities than FOC overall, especially in Kansas where $P(\mathrm{FF}|\mathrm{TOR})_{\mathrm{Warn}}$ was

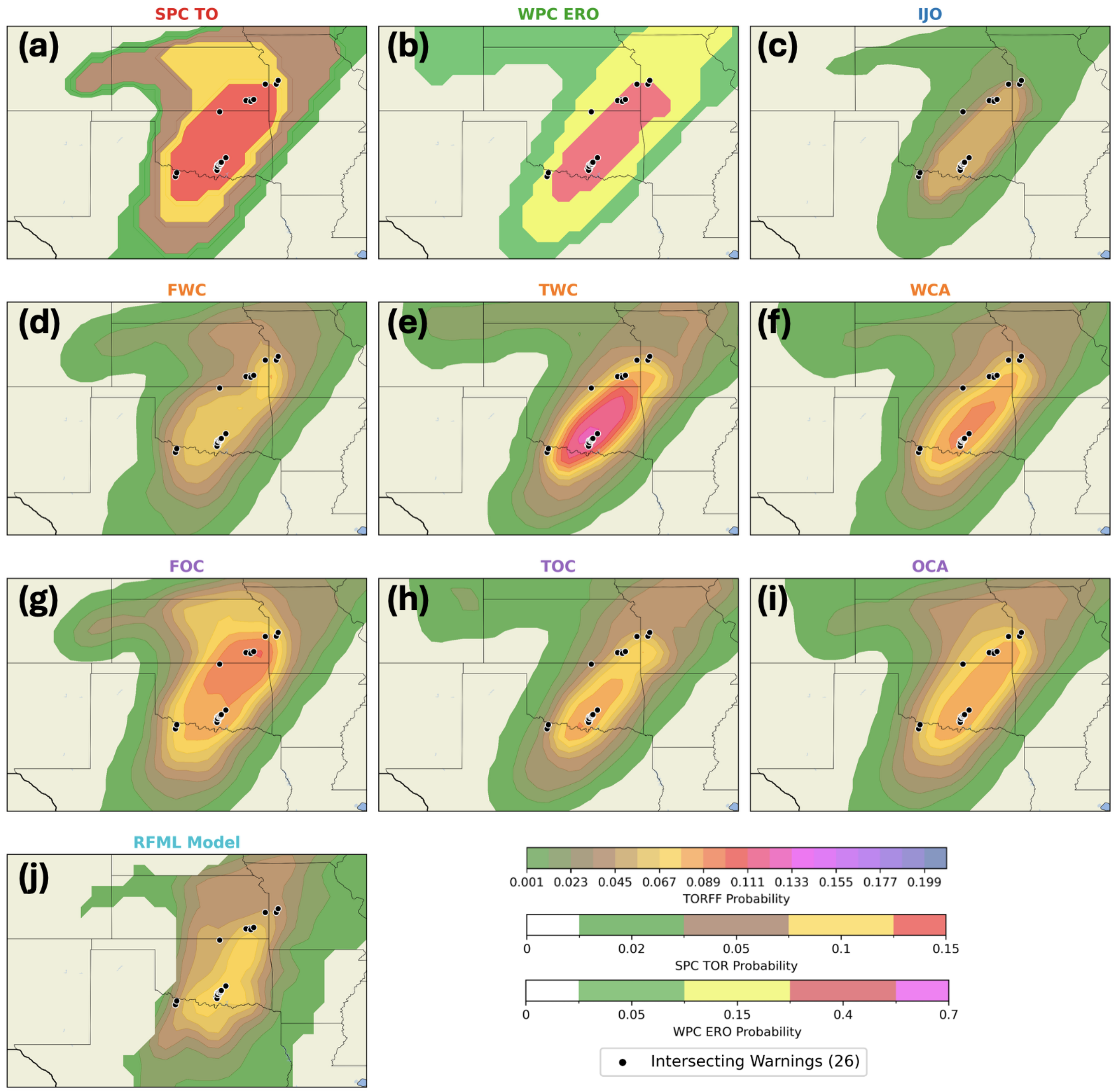


Figure 3.34: As in Figure 3.33 but valid 1200 UTC April 27, 2024 to 1200 UTC April 28, 2024

shown to be lowest (Figure 3.8). Despite this difference, both forecasts have probability maxima along the southern Kansas/Missouri boarder which somewhat highlight the small cluster in southeast Kansas. The TWC and TOC forecasts (Figure 3.34e,h) both have probability maxima from the central Oklahoma/Texas border into northeast Oklahoma that highlight the Marietta cluster, indicating better spatial discrimination than the FWC and FOC forecasts. The WCA and OCA forecasts (Figure 3.34f,i) have similar shape and probability maxima to the TWC and TOC forecasts, however the WCA forecast has much lower average probability than TWC forecast. Furthermore, the WCA forecast best highlights the cluster in Marietta while capturing the majority of events within relatively higher probability contours. Similar to test case 1, the IJO forecast (Figure 3.34c) has the most conservative forecast area, however its probability maximum is relatively large and does not highlight the individual clusters as well as the conditional joint forecasts. Finally, the RFML model forecast area (Figure 3.34j) captures all events, with the maximum probability contour highlighting the Marietta cluster but also extending further southeast than the joint forecasts. The shape of the forecast is wider and somewhat less linear than the joint forecasts which indicates less confidence in the forecast region, especially compared to its performance in test case 1.

#### 3.6.2.3 Test Case 3: 17 March, 2021

For the 17 March, 2021 case, TORFF events were mainly clustered in Mississippi and Alabama, with a particularly dense cluster in southwest Alabama and a smaller cluster in north-central Alabama (Figure 3.35). There are also two outlier events in southwest Missouri. The SPC TO forecast (Figure 3.35a) captures all events, with the majority occurring within the 15% and 30% contours. The WPC ERO (Figure 3.35b) also captures all events within the very large 15% contour. This case is unique in that the

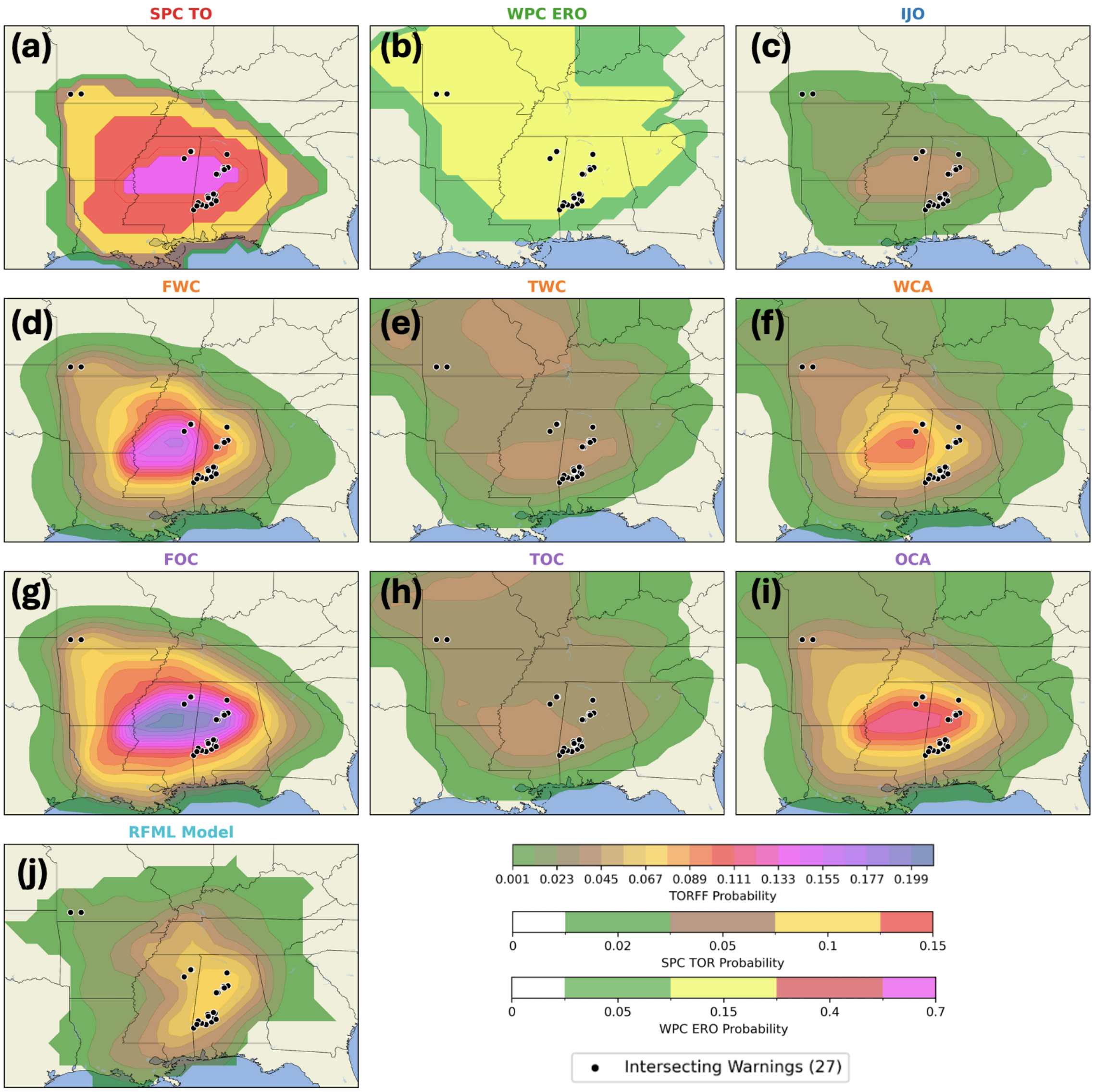


Figure 3.35: As in Figure 3.33 but valid 1200 UTC March 17, 2021 to 1200 UTC March 18, 2024

SPC TO probabilities are much higher than the WPC ERO, so the FWC and FOC forecasts have higher probabilities than the TWC and TOC forecasts. The FWC and FOC forecasts (Figure 3.35d,g) have similar shape and probability maxima in central Mississippi, however the FOC forecast has a higher maximum probability that also extends further east. The spatial discrimination of these two forecasts is decent given most events occur within relatively high probability contours; however, they fail to specifically highlight the cluster in Alabama. The TWC and TOC forecasts (Figure 3.35e,h) are similar with very low overall probabilities and poor spatial discrimination due to the large and relatively uniform WPC ERO forecast. The WCA and OCA forecasts (Figure 3.35f,i) largely reflect the shape and probability maxima of the FWC and FOC forecasts but with slightly lower probabilities, demonstrating similar spatial discrimination. Once more, the IJO forecast (Figure 3.35c) has roughly the smallest forecast area, but doesn't accurately highlight the cluster of events in southwest Alabama given the large probability maximum that ranges from central Mississippi to central Alabama. The RFML model in this case (Figure 3.35j) clearly shows the best spatial discrimination with a probability maximum from southwest to central Alabama which highlights both clusters in Alabama. The RFML model also extends into southwest Arkansas to highlight the two events there.

### 3.6.3 Discussion

The RFML model shows very promising initial results in test cases by accurately highlighting areas where TORFFs occur and generally displaying better spatial discrimination than the joint forecasts. Unfortunately, the machine learning had poor overall calibration (Figure 3.32), significantly under-forecasting at nearly all probabilities. This under-forecasting was evident in test cases 1 and 2 for which the model had relatively low probabilities compared to the other forecasts. These trends in reliability are very similar to what was found by Hill et al. (2024) who analyzed RFML models for excessive rainfall events, indicating that RFML models may struggle to represent excessive rainfall frequencies accurately, thus limiting reliability for TORFF events. Despite this, under-forecasting for TORFF events is not entirely problematic given the incredible rarity of such events which inherently results in extremely low forecast probabilities. Poor reliability at such low frequencies is therefore somewhat expected and less operationally concerning than spatial discrimination, as the primary challenge is identifying where these events are most likely to occur rather than assigning perfectly calibrated probabilities.

The RFML model performed considerably better than the joint forecasts in test case 3, mainly as a result of poor WPC ERO resolution in highlighting where flash flooding may occur. This result initially indicates that the model is skillful in identifying areas where there is enhanced flood risk where the WPC ERO cannot, likely as a result of the inclusion of multiple predictors for atmospheric moisture and precipitation (APCP, PWAT, Q2M, Q500). The model had the worst performance in test case 2 where it appeared the least confident spatially. Given this case occurred over the central plains (Oklahoma), this could indicate degraded performance where moisture advection isn't as strong or where discrete storm modes are more common (Smith et al. 2012), but further research would be needed to validate these claims.

Ideally, the reliability, BSS, and CSI of the RFML model should be compared to joint forecasts across the entire 2017 to 2024 forecast period, but a lack of time and data availability made this unfeasible. Therefore, it's difficult to make definitive conclusions on the overall performance of the model relative to the joint forecasts. Despite this, performance of the model is very promising in these test cases and demonstrates that even with extremely rare events, RFML is capable of forecasting TORFFs with comparable spatial discrimination to joint forecasts.

In this study, the RFML model was developed only as an initial test with a single set of atmospheric variables to determine if there was improved performance in test cases. Consequently, there was not enough time to perform feature importance tests that would highlight which predictors are contributing most to the model's performance or to test other sets of predictors. However, previous work on RFML models with similar GEFS predictors have shown APCP, CAPE, and SRH to have relatively higher importance for tornado forecasts (Hill et al. 2020), and ACPC and PWAT for excessive rainfall (Herman and Schumacher 2018b). Therefore, it's likely that these predictors would have relatively high importance for TORFF events, but more work will be needed to verify this. Furthermore, Hill et al. (2020) also found 500 hPa wind shear, 850 hPa wind shear, and mean sea level pressure (MSLP) to have relative high importance for tornadoes, but these predictors were not used in this study. Given that shear was found to play an important role in the development of TORFF environments (Nielsen et al. 2015; Burow et al. 2025; Rogash and Racy 2002; Rogash and Smith 2000), there could be an appreciable increase in model performance with the inclusion of these variables in future work. Furthermore, additional predictors at the 250 hPa and 850 hPa levels could provide additional predictive skill given previous research has found lower-tropospheric moisture and upper-tropospheric forcing for ascent to play significant roles in the formation of TORFFs (Maddox and Dietrich 1982; Rogash and Smith 2000; Rogash and Racy 2002; Nielsen et al. 2015).

For the joint forecasts, the results of the test cases can be tied back to trends in performance highlighted in the previous sections. The IJO forecast had the smallest overall forecast area in all test cases which supports the findings from section 3.5.2 showing it to have the highest max CSI and best overall spatial discrimination. The IJO forecast also had relatively low probabilities across all cases which demonstrates the significant under-forecasting shown in Figure 3.18. The FWC, FOC, TWC, and TOC forecasts all had roughly the same size as the operational outlook used to derive them, but generally had enhanced spatial discrimination thanks to the localized conditional probabilities. The significant over-forecasting by the TWC and TOC forecasts (Figures 3.20, 3.23) is also evident in test cases 1 and 2 for which they had considerably higher overall probabilities than the other forecasts. Finally, the WCA and OCA forecasts generally had better spatial discrimination in the test cases than the other conditional forecasts, supporting the findings in section 3.5.2 which showed them having relatively higher CSI and SR values.

# Chapter 4

# Summary and Conclusion

TORFF events, although extremely rare, are an exceptionally dangerous weather hazard that can pose substantial risks to the public and present numerous challenges for forecasters, broadcasters, and EMs who must coordinate the communication of both threats simultaneously. Despite growing research into the climatology and meteorological setup of TORFFs, no research had attempted to establish a forecast framework that is capable of highlighting the risk of TORFF events with sufficient accuracy or calibration. This thesis addressed that gap by evaluating multiple forecast methods for TORFF events including independent joint probabilities derived from operational SPC tornado outlooks and WPC excessive rainfall outlooks, conditional probability methods that account for statistical relationships between tornadoes and flash floods, and finally a random forest machine learning model trained on GEFS reforecast data.

TORFF events were defined as spatially and temporally intersecting tornado-flash flood warnings which were shown to occur most frequently in the lower Mississippi River valley, particularly in the Memphis, TN CWA. This area was roughly where relatively high tornado and flash flood warnings frequencies intersected. Overlaps of SPC TOs and WPC EROs also occurred with high frequency in these areas, demonstrating that overlaps in synoptic-scale tornado and flash flood environments can be used to estimate where TORFF environments are favorable.

Given previous research that has demonstrated similarities in tornado, flash flood, and TORFF environments, it was hypothesized that a significant statistical relationship between tornadoes and flash floods would be observed across the CONUS such that forecast probabilities for the two hazards could not be assumed independent. The correlation between tornado and flash flood warnings was analyzed with a linear regression model which indicated 7 out of 59 NWS CWAs (11.86%) with sufficient data to have a significant positive correlation. The lack of significance in the majority of CWAs made it difficult to prove the hypothesis on a broad scale. Although there was a general trend towards positive correlation, these results alone are insufficient to demonstrate a robust physical relationship between the two events, and future research is necessary to explore the relationship in more detail. Despite this, a region of 5 adjacent CWAs from Memphis to Des Moines was particularly notable and was used to delineate a single region (Correlated CWAs) where forecasts would be evaluated separately to determine if a significant difference in forecast performance exists. Furthermore, this motivated the use of conditional probability as an alternative to independent probabilities, with a hypothesis that joint forecasts using conditional (independent) probability would perform better (worse) in the correlated region. Additionally, it was hypothesized that spatial distributions of conditional probability would exhibit patterns similar to the linear regression results and TORFF frequency.

Conditional probabilities were estimated at the warning-scale and outlook-scale using tornado/flash flood warnings and SPC/WPC outlooks, respectively, to determine the probability of a tornado given a flash flood ($P(\mathrm{TOR}|\mathrm{FF})$) and the probability of a flash flood given a tornado ($P(\mathrm{FF}|\mathrm{TOR})$). Analyses of conditional probability distributions revealed that the $P(\mathrm{FF}|\mathrm{TOR})$ is greater than $P(\mathrm{TOR}|\mathrm{FF})$ in nearly all regions, and that outlook-scale estimates for $P(\mathrm{FF}|\mathrm{TOR})$ were much higher on average than the warning-scale. Conversely, averages for $P(\mathrm{TOR}|\mathrm{FF})$ were much closer for the two scales. Spatial distributions did not show robust patterns that matched the linear regression model or TORFF frequency, suggesting that the statistical relationship between tornadoes and flash floods is complex, and conditional probability does not necessarily reflect TORFF climatology or where strong correlation is observed. As a result, future work could explore alternative methods in evaluating the conditional relationship between tornadoes and flash floods to asses if similar conclusion can be made on the relationship to correlation or TORFF climatology.

Reliability diagrams and BSS were used to determine if significant differences in forecast performance exist between the correlated and uncorrelated CWA regions. Results indicated no significant difference in reliability or BSS for any forecast method, suggesting that regional differences in tornado-flash flood correlation do not significantly impact TORFF forecast performance and are thus unimportant in the development of the framework. Several caveats were recognized in this analysis such as small sample size in the Correlated CWAs region and uncertainty in the region itself due to limitations with the linear regression model. As a result, these findings should be interpreted with caution and do not necessarily imply that tornado-flash flood dependence is unimportant for TORFF forecasting. Future studies utilizing larger datasets and alternative methods for identifying regions of hazard dependence may provide additional insight into this relationship.

Given that no significant difference in forecast performance was observed between correlated and uncorrelated regions, the two regions were merged, and forecast performance in this single region was compared among all joint forecasts. Forecast verification indicated that the IJO and FWC forecasts were the most skillful overall; both maintained significantly positive BSS across all seasons and years, demonstrating consistent improvement over climatology. In contrast, the TWC and TOC forecasts generally exhibited negative BSS due to their reliance on the WPC ERO, which tended to overforecast TORFF environments and showed poor calibration. Seasonal analysis revealed that forecast skill was highest during spring, when TORFF events are most frequent, and lowest during summer, suggesting that summertime TORFF environments may be less effectively represented by the forecast methods. Interannual analysis further showed that most conditional forecasts were sensitive to sample size, whereas the IJO and FWC forecasts maintained positive skill even during years with relatively few events.

Performance diagrams demonstrated that the IJO forecast provided the best spatial discrimination, minimizing false alarms and maximizing CSI. The WCA and OCA forecasts also performed well, suggesting that forecast methods incorporating both SPC TO and WPC ERO information achieve a more favorable balance between missed events and false alarms. However, these methods generally exhibited poorer calibration than the FWC forecast. Consequently, a tradeoff exists between calibration and spatial discrimination; the FWC forecast was best calibrated, while the IJO forecast offered the greatest ability to localize TORFF risk. Despite the low absolute values of CSI and BSS resulting from the extreme rarity of TORFF events, nearly all joint forecast methods substantially outperformed the SPC TO and WPC ERO alone, demonstrating the value of dedicated TORFF forecasting approaches for identifying regions of enhanced dual-hazard risk.

Finally, an RFML model was trained with GEFS reforecast data and intersecting warnings to asses if it was capable of highlighting TORFF events with sufficient calibration and if there was improved spatial discrimination across test cases compared to the joint forecasts. The model demonstrated promising potential for TORFF forecasting by accurately identifying regions of enhanced TORFF risk and generally exhibiting spatial discrimination that was comparable to and often better than the joint forecast methods. The model was particularly effective in situations where operational outlooks struggled to localize flash flood risk, suggesting that incorporating environmental predictors related to moisture and precipitation can provide additional forecast value.

However, the model exhibited poor calibration and a tendency to under-forecast, and its overall performance could not be comprehensively evaluated due to limited data availability. Consequently, while the results indicate that machine learning is a viable approach for forecasting these extremely rare events, additional verification, predictor testing, and feature importance analyses are needed to fully assess and improve its operational utility.

In summary, the relative success of the IJO, FWC, and RFML model forecasts prove that there are multiple effective methods to model TORFF environments with sufficient lead time in a probabilistic framework. The IJO forecast and RFML model demonstrated the best spatial discrimination in test cases which proves that TORFF environments can be accurately distinguished from tornado and flash flood only environments by combining operational outlooks for the respective hazards or with a statistical approach using atmospheric data, respectively. The FWC forecast demonstrated the best overall calibration which indicates that conditional probability in conjunction with the SPC TO is best at highlighting the true probability of TORFF events. Despite this, modest improvements in calibration are generally less important to the forecast framework than accurate spatial discrimination. Therefore, the IJO and RFML forecasts should be the main focus of future work in developing forecast methods that accurately distinguish TORFF environments. The meteorological characteristics of TORFFs remain fairly complex and difficult to predict with confidence, but the results of this thesis demonstrate strong potential for forecast products specific to TORFFs to highlight areas of enhanced risk. Therefore, forecasters, broadcasters, and EMs could use such forecasts in the future to improve their understanding of dual-hazard weather events, increasing lead time for necessary protective action and enhancing communication with the public when threats may overlap.

# Reference List


Ashley, W. S., A. M. Haberlie, and J. Strohm, 2019: A Climatology of Quasi-Linear Convective Systems and Their Hazards in the United States. https://doi.org/10.1175/WAF-D-19-0014.1, URL https://journals.ametsoc.org/view/journals/wefo/34/6/waf-d-19-0014_1.xml.

Bradley, A. A., S. S. Schwartz, and T. Hashino, 2008: Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score. https://doi.org/10.1175/2007WAF2007049.1, URL https://journals.ametsoc.org/view/journals/wefo/23/5/2007waf2007049_1.xml.

Breiman, L., 2001: Random Forests. *Machine Learning*, **45 (1)**, 5–32, https://doi.org/10.1023/A:1010933404324, URL https://doi.org/10.1023/A:1010933404324.

Brier, G. W., 1950: VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Brooks, H. E., C. A. Doswell, and M. P. Kay, 2003: Climatological Estimates of Local Daily Tornado Probability for the United States. URL https://journals.ametsoc.org/view/journals/wefo/18/4/1520-0434_2003_018_0626_ceoldt_2_0_co_2.xml.

Bröcker, J., and L. A. Smith, 2007: Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*, **22 (3)**, 651–661, https://doi.org/10.1175/WAF993.1, URL https://journals.ametsoc.org/doi/10.1175/WAF993.1.

Burke, P. C., A. Lamers, G. Carbin, M. J. Erickson, M. Klein, M. Chenard, J. McNatt, and L. Wood, 2022: The Excessive Rainfall Outlook at the Weather Prediction Center: Operational Definition, Construction, and Real-Time Collaboration. URL https://repository.library.noaa.gov, https://doi.org/10.1175/BAMS-D-21-0281.1.

Burow, D., K. Ellis, and L. Tran, 2021: Simultaneous and collocated tornado and flash flood warnings associated with tropical cyclones in the contiguous United States. *International Journal of Climatology*, **41 (8)**, 4253–4264, https://doi.org/10.1002/joc.7071, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.7071, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.7071.

Burow, D., K. Kressler, and Z. Searcy, 2025: On the Convective Environments, Modes, and Warning Verifications of Tornado- and Flash Flood-Warned Storms in the Southeast United States. *Meteorological Applications*, **32 (3)**, e70 062, https://doi.org/10.

1002/met.70062, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/met.70062, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.70062.

Chaney, P. L., and G. S. Weaver, 2010: The Vulnerability of Mobile Home Residents in Tornado Disasters: The 2008 Super Tuesday Tornado in Macon County, Tennessee. https://doi.org/10.1175/2010WCAS1042.1, URL https://journals.ametsoc.org/view/journals/wcas/2/3/2010wcas1042_1.xml.

Charilaou, P., and R. Battat, 2022: Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, **28 (5)**, 605–607, https://doi.org/10.3748/wjg.v28.i5.605, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8905023/.

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

Dougherty, E., and K. L. Rasmussen, 2019: Climatology of Flood-Producing Storms and Their Associated Rainfall Characteristics in the United States. https://doi.org/10.1175/MWR-D-19-0020.1, URL https://journals.ametsoc.org/view/journals/mwre/147/11/mwr-d-19-0020.1.xml.

Ellis, K. N., J. M. First, S. M. Strader, N. S. Grondin, D. Burow, and Z. Medley, 2023: The Climatology, Vulnerability, and Public Perceptions Associated with Overlapping Tornado and Flash Flood Warnings in a Portion of the Southeast United States. https://doi.org/10.1175/WCAS-D-23-0018.1, URL https://journals.ametsoc.org/view/journals/wcas/15/4/WCAS-D-23-0018.1.xml.

Erickson, M. J., B. Albright, and J. A. Nelson, 2021: Verifying and Redefining the Weather Prediction Center's Excessive Rainfall Outlook Forecast Product. https://doi.org/10.1175/WAF-D-20-0020.1, URL https://journals.ametsoc.org/view/journals/wefo/36/1/WAF-D-20-0020.1.xml.

Erickson, M. J., J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification Results from the 2017 HMT–WPC Flash Flood and Intense Rainfall Experiment. https://doi.org/10.1175/JAMC-D-19-0097.1, URL https://journals.ametsoc.org/view/journals/apme/58/12/jamc-d-19-0097.1.xml.

Ernst, S. R., and Coauthors, 2024: A Tale of Two Hazards: Studying Broadcast Meteorologist Communication of Simultaneous Tornado and Flash Flood (TORFF) Events. National Weather Association, URL https://repository.library.noaa.gov, https://doi.org/10.15191/nwajom.2024.1201.

First, J. M., K. Ellis, and S. Strader, 2022: Double trouble: Examining public protective decision-making during concurrent tornado and flash flood threats in the U.S. Southeast. *International Journal of Disaster Risk Reduction*, **81**, 103 297, https://doi.org/10.1016/j.ijdrr.2022.103297, URL https://www.sciencedirect.com/science/article/pii/S2212420922005167.

Gallus, W. A., N. A. Snook, and E. V. Johnson, 2008: Spring and Summer Severe Weather Reports over the Midwest as a Function of Convective Mode: A Preliminary Study. https://doi.org/10.1175/2007WAF2006120.1, URL https://journals.ametsoc.org/view/journals/wefo/23/1/2007waf2006120_1.xml.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. https://doi.org/10.1175/BAMS-D-12-00014.1, URL https://journals.ametsoc.org/view/journals/bams/94/10/bams-d-12-00014.1.xml.

Harrison, D. R., and C. D. Karstens, 2017: A Climatology of Operational Storm-Based Warnings: A Geospatial Analysis. https://doi.org/10.1175/WAF-D-15-0146.1, URL https://journals.ametsoc.org/view/journals/wefo/32/1/waf-d-15-0146_1.xml.

Henderson, J., E. R. Nielsen, G. R. Herman, and R. S. Schumacher, 2020: A Hazard Multiple: Overlapping Tornado and Flash Flood Warnings in a National Weather Service Forecast Office in the Southeastern United States. https://doi.org/10.1175/WAF-D-19-0216.1, URL https://journals.ametsoc.org/view/journals/wefo/35/4/wafD190216.xml.

Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic Verification of Storm Prediction Center Convective Outlooks. https://doi.org/10.1175/WAF-D-17-0104.1, URL https://journals.ametsoc.org/view/journals/wefo/33/1/waf-d-17-0104_1.xml.

Herman, G. R., and R. S. Schumacher, 2018a: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. https://doi.org/10.1175/MWR-D-17-0250.1, URL https://journals.ametsoc.org/view/journals/mwre/146/5/mwr-d-17-0250.1.xml.

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. https://doi.org/10.1175/MWR-D-17-0307.1, URL https://journals.ametsoc.org/view/journals/mwre/146/6/mwr-d-17-0307.1.xml.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049, https://doi.org/10.1002/

qj.3803, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

Herzman, D., 2024: IEM :: NWS Watches Warnings Advisories Download. URL https://mesonet.agron.iastate.edu/request/gis/watchwarn.phtml.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. https://doi.org/10.1175/MWR-D-19-0344.1, URL https://journals.ametsoc.org/view/journals/mwre/148/5/mwr-d-19-0344.1.xml.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting Excessive Rainfall with Random Forests and a Deterministic Convection-Allowing Model. https://doi.org/10.1175/WAF-D-21-0026.1, URL https://journals.ametsoc.org/view/journals/wefo/36/5/WAF-D-21-0026.1.xml.

Hill, A. J., R. S. Schumacher, and M. L. Green, 2024: Observation Definitions and Their Implications in Machine Learning–Based Predictions of Excessive Rainfall. https://doi.org/10.1175/WAF-D-24-0033.1, URL https://journals.ametsoc.org/view/journals/wefo/39/11/WAF-D-24-0033.1.xml.

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. https://doi.org/10.1175/WAF-D-22-0143.1, URL https://journals.ametsoc.org/view/journals/wefo/38/2/WAF-D-22-0143.1.xml.

Hitchens, N. M., and H. E. Brooks, 2017: Determining Criteria for Missed Events to Evaluate Significant Severe Convective Outlooks. https://doi.org/10.1175/WAF-D-16-0170.1, URL https://journals.ametsoc.org/view/journals/wefo/32/4/waf-d-16-0170_1.xml.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective Limits on Forecasting Skill of Rare Events. https://doi.org/10.1175/WAF-D-12-00113.1, URL https://journals.ametsoc.org/view/journals/wefo/28/2/waf-d-12-00113_1.xml.

James, E. P., and R. S. Schumacher, 2024: Precipitation Proxies for Flash Flooding: A Seven-Year Analysis over the Contiguous United States. https://doi.org/10.1175/JHM-D-23-0203.1, URL https://journals.ametsoc.org/view/journals/hydr/25/9/JHM-D-23-0203.1.xml.

Jiang, Q., D. T. Dawson II, F. Li, and D. R. Chavas, 2025: Classifying synoptic patterns driving tornadic storms and associated spatial trends in the United States. *Npj Climate and Atmospheric Science*, **8 (1)**, 7, https://doi.org/10.1038/s41612-025-00897-1, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11711087/.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Post-processing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests. https://doi.org/10.1175/WAF-D-19-0109.1, URL https://journals.ametsoc.org/view/journals/wefo/34/6/waf-d-19-0109_1.xml.

Maddox, R., and W. Dietrich, 1982: METEOROLOGICAL SETTINGS ASSOCIATED WITH SIMULTANEOUS OCCURRENCES OF SEVERE THUNDERSTORMS AND FLASH FLOODS. *National Weather Digest*, **7 (3)**.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. https://doi.org/10.1175/BAMS-D-16-0123.1, URL https://journals.ametsoc.org/view/journals/bams/98/10/bams-d-16-0123.1.xml.

Mercer, A., and A. Bates, 2019: Meteorological Differences Characterizing Tornado Outbreak Forecasts of Varying Quality. *Atmosphere*, **10 (1)**, 16, https://doi.org/10.3390/atmos10010016, URL https://www.mdpi.com/2073-4433/10/1/16.

Meropi, P., C. Bikos, and Z. George, 2018: Outlier detection in skewed data. *Simulation Modelling Practice and Theory*, **87**, 191–209, https://doi.org/10.1016/j.simpat.2018.05.010, URL https://www.sciencedirect.com/science/article/pii/S1569190X1830073X.

Murphy, A. H., 1973: A New Vector Partition of the Probability Score. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

Nielsen, E. R., G. R. Herman, R. C. Tournay, J. M. Peters, and R. S. Schumacher, 2015: Double Impact: When Both Tornadoes and Flash Floods Threaten the Same Place at the Same Time. https://doi.org/10.1175/WAF-D-15-0084.1, URL https://journals.ametsoc.org/view/journals/wefo/30/6/waf-d-15-0084_1.xml.

Nielsen, E. R., and R. S. Schumacher, 2020: Observations of Extreme Short-Term Precipitation Associated with Supercells and Mesovortices. https://doi.org/10.1175/MWR-D-19-0146.1, URL https://journals.ametsoc.org/view/journals/mwre/148/1/mwr-d-19-0146.1.xml.

NOAA, 2007: National Weather Service Storm Based Warnings. URL https://www.weather.gov/media/pah/WeatherEducation/stormbased.pdf.

NOAA, 2024: 2024 Oklahoma Tornadoes. NOAA's National Weather Service, URL https://www.weather.gov/oun/tornadodata-ok-2024.

NOAA, 2024: Weather Related Fatality and Injury Statistics. NOAA's National Weather Service, URL https://www.weather.gov/hazstat.

NOAA, 2025: County Warning Area Boundaries. NOAA's National Weather Service, URL https://www.weather.gov/gis/CWABounds.

NOAA, 2026a: About the WPC. URL https://www.wpc.ncep.noaa.gov/html/about2.shtml.

NOAA, 2026b: NOAA/NWS Storm Prediction Center. URL https://www.spc.noaa.gov/misc/about.php.

P dos Santos, R., 2016: Some Comments on the Reliability of NOAA's Storm Events Database. Social Science Research Network, Rochester, NY, URL https://papers.ssrn.com/abstract=2799273, https://doi.org/10.2139/ssrn.2799273.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. https://doi.org/10.1175/2008WAF2222159.1, URL https://journals.ametsoc.org/view/journals/wefo/24/2/2008waf2222159_1.xml.

Rogash, J. A., and J. Racy, 2002: Some Meteorological Characteristics of Significant Tornado Events Occurring in Proximity to Flash Flooding. *Weather and Forecasting*, **17 (1)**, 155–159, https://doi.org/10.1175/1520-0434(2002)017⟨0155:SMCOST⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0434(2002)017⟨0155:SMCOST⟩2.0.CO;2.

Rogash, J. A., and R. D. Smith, 2000: Multiscale Overview of a Violent Tornado Outbreak with Attendant Flash Flooding. *Weather and Forecasting*, **15 (4)**, 416–431, https://doi.org/10.1175/1520-0434(2000)015⟨0416:MOOAVT⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0434(2000)015⟨0416:MOOAVT⟩2.0.CO;2.

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. https://doi.org/10.1175/BAMS-D-20-0186.1, URL https://journals.ametsoc.org/view/journals/bams/102/9/BAMS-D-20-0186.1.xml.

Schumacher, R. S., and R. H. Johnson, 2006: Characteristics of U.S. Extreme Rain Events during 1999–2003. https://doi.org/10.1175/WAF900.1, URL https://journals.ametsoc.org/view/journals/wefo/21/1/waf900_1.xml.

SciKit-Learn Developers, 2026: RandomForestClassifier. URL https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Sherburn, K. D., and Coauthors, 2019: Partnering Research, Education, and Operations via a Cool Season Severe Weather Soundings Program. https://doi.org/10.1175/BAMS-D-17-0186.1, URL https://journals.ametsoc.org/view/journals/bams/100/2/bams-d-17-0186.1.xml.

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

Sobash, R. A., D. J. Gagne, C. L. Becker, D. Ahijevych, G. N. Gantos, and C. S. Schwartz, 2023: Diagnosing Storm Mode with Deep Learning in Convection-Allowing Models. https://doi.org/10.1175/MWR-D-22-0342.1, URL https://journals.ametsoc.org/view/journals/mwre/151/8/MWR-D-22-0342.1.xml.

White, E., and A. J. Hill, 2026: Severe Weather Forecasts from Artificial Intelligence Weather Prediction Models. https://doi.org/10.1175/AIES-D-25-0065.1, URL https://journals.ametsoc.org/view/journals/aies/5/1/AIES-D-25-0065.1.xml.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*. Elsevier Academic Press, Amsterdam; Boston, URL https://www.amazon.com/Statistical-Atmospheric-Sciences-International-Geophysics/dp/0123850223/ref=pd_bxgy_14_img_3?_encoding=UTF8&psc=1&refRID=ESPQQ0R2PB1TP1VJSGCZ.

Yuan, K.-H., and P. M. Bentler, 2001: Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, **54 (1)**, 161–175, https://doi.org/10.1348/000711001159366, URL https://onlinelibrary.wiley.com/doi/abs/10.1348/000711001159366, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1348/000711001159366.

Zhu, Y.-J., J. M. Collins, P. J. Klotzbach, and C. J. Schreck, 2022: Hurricane Ida (2021): Rapid Intensification Followed by Slow Inland Decay. https://doi.org/10.1175/BAMS-D-21-0240.1, URL https://journals.ametsoc.org/view/journals/bams/103/10/BAMS-D-21-0240.1.xml.