DISSERTATION

COMPARING PRECIPITATION ESTIMATES, MODEL FORECASTS, AND RANDOM FOREST BASED PREDICTIONS FOR EXCESSIVE RAINFALL

Submitted by

Eric James

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

Advisor: Russ Schumacher

Michael Bell
Peter Jan Van Leeuwen
Ryan Morrison

Copyright by Eric Paul James 2023

All Rights Reserved

ABSTRACT

COMPARING PRECIPITATION ESTIMATES, MODEL FORECASTS, AND RANDOM FOREST BASED PREDICTIONS FOR EXCESSIVE RAINFALL

Flash flooding is an important societal challenge, and improved tools are needed for both real-time analysis and short-range forecasts. We present an evaluation of threshold exceedances of quantitative precipitation estimate (QPE) and forecast (QPF) datasets in terms of their degree of correspondence with observed flash flood events over a seven-year period. We find that major uncertainties persist in QPE for heavy rainfall. In general, comparison with flash flood guidance (FFG) thresholds provides the best correspondence, but fixed thresholds and average recurrence interval thresholds provide the best correspondence in certain regions of the contiguous US (CONUS). QPF threshold exceedances from the High-Resolution Rapid Refresh (HRRR) generally do not correspond as well as QPE exceedances with observed flash floods, except for the 1-h duration in the southwestern CONUS; this suggests that high-resolution model QPF may be a better indicator of flash flooding than QPE in some poorly observed regions.

Subsequently, we describe a new random forest (RF) based excessive rainfall forecast system using predictor information from the 3-km operational HRRR. Experiments exploring the use of spatial predictor information reveal the importance of averaging HRRR predictor fields across a spatial radius rather than using only information from sparse input grid points for regimes with small-scale excessive rain events. Tree interpreter results indicate that the forecast benefits of spatial aggregation stem from greater contributions provided by storm attribute predictors. Forecasts are slightly degraded when there is a mismatch between the trained RF model and the daily HRRR forecasts to which the model is applied, both in terms of initialization

time and HRRR model version. Use of FFG as an additional predictor leads to forecast improvements, highlighting the potential of hydrologic information to contribute to forecast skill. In addition, averaging predictor information across several HRRR initializations leads to a statistically significant improvement in forecasts relative to using predictor fields from a single HRRR initialization. The HRRR-based RF has been evaluated at the annual Flash Flood and Intense Rainfall Experiment (FFaIR) over the past three years, with year-over-year improvements stemming from the results of sensitivity experiments. The HRRR-based RF represents an important baseline for future machine learning based excessive rainfall forecasts based on convection-allowing models.

ACKNOWLEDGEMENTS

I would like to acknowledge the encouragement and support of my colleagues and supervisors at the NOAA Global Systems Laboratory over the past five years, giving me flexibility to complete coursework and work on PhD research while still working full time through the Cooperative Institute for Research in Environmental Sciences at the University of Colorado. In particular, I would like to thank Steve Weygandt, Curtis Alexander, Terra Ladwig, Ming Hu, and Guoqing Ge.

A special thank you is due to my mentor, colleague, and close friend Stan Benjamin for encouraging me to pursue this degree, and for his exemplary passion for science for the sake of improving weather forecasts.

I would also like to thank all the past and present members of the Schumacher research group, for creating a supportive environment for a non-traditional student like me, and for all the fruitful discussions on various topics related to meteorology and machine learning.

And of course, none of this work would have been possible without the support and incredible patience of my wife, Sarah.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Flash flooding, defined by the United States (US) National Weather Service (NWS) as "a rapid rise in water levels along rivers, creeks, normally dry washes, arroyos, or even normally dry land areas, generally occurring within 6 h of the causative rainfall or other event", is a critical societal challenge globally, and particularly in the US. The National Centers for Environmental Information (NCEI) estimates more than $72 billion (U.S. dollars) in damage and 212 deaths resulting from flooding and excessive rainfall in the U.S. during 2010-2019 (NCEI 2020). The majority of flood-related fatalities in the U.S. over the last ~50 y can be attributed to flash flooding (Ashley and Ashley 2008). Flash floods kill more people per year in the U.S. than any other weather hazard except for heat waves (NOAA 2022). Notable recent flash flooding disasters in the U.S. include Hurricane Harvey in Texas in August 2017 (Martinaitis et al. 2021), flooding associated with the remnants of Hurricane Ida in the northeastern U.S. in Sep 2021 (Smith et al. 2023), and repeated mesoscale convective systems (MCSs) in Kentucky and Missouri in Jul 2022 (Wix 2023). Many other parts of the world have also been subjected to severe flash flooding in recent years; for example, in Pakistan and Iran in Jul 2022 (at least 1800 people killed; Ghasabi et al. 2023; Pakistan National Disaster Management Authority 2022), and in central Europe in Jul 2021 (at least 180 people killed; Lehmkuhl et al. 2022).

Recent research has suggested that extreme precipitation in the U.S. may become more frequent with climate change, both in terms of volumetric rainfall from MCSs (Prein et al. 2017) and in terms of the upper tail of the wet day precipitation distribution (Harp and Horton 2022). Hydrologic model results forced with high-resolution climate simulations also indicate increased flashiness of U.S. floods under a high emission scenario (Li et al. 2022). Potential resulting

future increases in the frequency of flash flooding events, as well as increasing vulnerability due to population growth (Pielke and Downton 2002), underscores the importance of improved flash flood predictions.

Despite the societal importance of accurate flash flood forecasts, the rate of improvement of quantitative precipitation forecasts (QPF) for heavy rainfall events lags behind other forecast quality metrics (Barthold et al. 2015; Novak 2023). Accurate prediction of heavy rainfall depends not just upon an accurate representation of the convective environment, and not just upon an accurate representation of instantaneous storm structure, but also upon accurate storm morphology and evolution over time, including treatment of successive storms and training. Excessive rainfall can also arise from a much broader spectrum of precipitation systems than the spectrum that can produce, for example, tornadoes.

The Weather Prediction Center (WPC) issues daily Excessive Rainfall Outlooks (EROs) to highlight regions of concern for flash flooding in the coming days (Burke et al. 2023). The ERO is formulated as a probability of exceeding flash flood guidance (FFG) over a 24-h period within 25 miles of a point, and is issued operationally out to day five; in this work, we focus only on the day-one period. One of the major challenges associated with the ERO is how to define an excessive rainfall event for the purposes of forecast verification. Erickson et al. (2021) present quantitative verification of the WPC ERO during 2015-21 against a Unified Flooding Verification System (UFVS), which consists of Stage IV quantitative precipitation estimates (QPE) exceeding FFG, Stage IV QPE exceeding the 5-year average recurrence interval (ARI), U.S. Geological Survey (USGS) river gauge observations, and NWS local storm reports.

While many tools have been developed for flash flood analysis and forecasting, these tools have outstanding issues and challenges. In terms of precipitation analyses, QPE datasets

often do not agree with each other (e.g., Bytheway et al. 2020) and have substantial quality control issues (e.g., Nelson et al. 2016). Numerical weather prediction (NWP), which is the backbone of prediction beyond a few hours, struggles with capturing convective evolution relevant to flash flooding (e.g., Nielsen and Schumacher 2020). Nowcasting approaches struggle with realism (e.g., Radhakrishnan and Chandrasekar 2020).

Machine learning (ML) holds promise for providing improved tools for flash flood analysis and forecasting. Particular applications of ML relevant for flash flood analysis and forecast include ML-based QPE in complex terrain (Osborne et al. 2023), novel ML architectures for precipitation nowcasting (e.g., Zhang et al. 2023), post-processing of ensemble QPF for improved forecasts (Loken et al. 2019), and ML-based excessive rainfall prediction systems (e.g., Herman and Schumacher 2018a). ML-based excessive rainfall prediction has not received as much attention as ML-based prediction of the other convective hazards of tornadoes, hail, and severe wind (McGovern et al. 2023).

In this work, we first examine existing QPE and QPF datasets to shed light on their ability to highlight potential flash flooding. We adopt the framework of Herman and Schumacher (2018c) to evaluate how well exceedances of various types of precipitation thresholds in QPE and QPF datasets correspond with observed flash floods. Understanding the strengths and weaknesses of these datasets in terms of representing impactful precipitation provides important context for examining excessive rainfall prediction skill variability around the US.

We then build upon previous work to describe a new random forest (RF) based system for excessive rainfall prediction based on inputs from a deterministic convection-allowing model (CAM). In particular, we focus on how predictors from a CAM should be treated differently

from those from a coarse global ensemble. ML in general, and RFs in particular, continue to grow in importance within the forecasting process for high-impact weather events, so understanding how to optimally construct such RF systems is of paramount importance for the future of these prediction systems. We present a variety of sensitivity experiments addressing the spatial and temporal aggregation of meteorological predictors for the RF, as well as the importance of the RF training period length and the impact of model changes during the training period. We also present an objective evaluation of real-time forecasts from the RF system over a three-year period, with a comparison against operational WPC EROs. The deterministic CAM-based RF described here represents an important baseline for future, more sophisticated ML approaches which use high-resolution predictor information from deterministic CAMs, as well as those based on future formal CAM ensemble systems.

CHAPTER 2: PRECIPITATION PROXIES FOR FLASH FLOODING: A SEVEN-YEAR ANALYSIS OVER THE CONTIGUOUS UNITED STATES

Flash flooding remains a difficult prediction problem, and one of high societal importance due to the projected increases in flash flood related losses due to population growth (e.g., Downton et al. 2005) and climate change (Prein et al. 2017). One of the key challenges with flash flood forecasting is the lack of a universally accepted definition of the phenomenon. The National Weather Service (NWS) defines a flash flood as "a rapid rise in water levels, along rivers, creeks, normally dry washes, arroyos, or even normally dry land areas, generally occurring within 6 h of the causative rainfall or other event" (NWS 2023). Even beyond the formal definition, the likelihood of flash flooding resulting from a given intensity and duration of heavy rain is strongly dependent upon a hydrologic response, which dramatically varies regionally and in time. Due to these complications, it is helpful for forecasters to have a quick way to see what magnitude of rainfall accumulation or rate is climatologically anomalous or would cause a flood response given other hydrologic factors. One way to sift the available information is to filter out events that are climatologically or hydrologically not as likely to cause flooding. A given rainfall accumulation over a given duration could have vastly different impacts depending on location or time (i.e., antecedent conditions, land surface type, vegetation, topography, etc.). In this regard, it is important to have accurate estimates of the precipitation threshold beyond which flash flooding may occur, at high spatial and temporal resolutions. From a climatological perspective, NOAA Atlas 14 (Bonnin et al. 2006; Perica et al. 2011, 2013a,b, 2015, 2018) is intended to reflect the average amount of precipitation that corresponds to a given recurrence interval, highlighting statistically "rarer" precipitation events. Flash flood guidance (FFG), on the other hand, is intended to reflect hydrologic capacity given soil

information and antecedent conditions. These thresholds are available down to the 1-h temporal scale, but there is some indication that precipitation accumulations at even finer temporal scales (e.g., 15 min) could be important for some types of flooding events (e.g., landslides; Kean et al. 2019).

Ultimately, treatment of the hydrological factors influencing the probability of flash flooding is appropriately handled only with advanced hydrologic models. The Flooded Locations and Simulated Hydrographs (FLASH) system (Gourley et al. 2017) now provides gridded comparisons of radar-based quantitative precipitation estimates (QPE) with recurrence intervals and FFG, as well as an ensemble hydrologic model which provides high-resolution and frequently-updated streamflow predictions. Recent work has also coupled the FLASH system with an experimental short-range ensemble forecast system (Yussouf et al. 2020; Martinaitis et al. 2022), enabling improved meteorological forcing and therefore improved streamflow forecasts. However, these novel applications are restricted to shorter lead times (3-6 hours), and there remains a need for comparison of longer-range convection-allowing model (CAM) forecasts with precipitation thresholds of interest.

In response to the somewhat ambiguous nature of flash flood events, and because of issues with the flash flood report (FFR) dataset, the Weather Prediction Center (WPC) has developed a dataset, known as the Unified Flooding Verification System (UFVS), which combines FFRs with "proxy" flood events derived from gridded comparisons of QPE vs. several thresholds (Erickson et al. 2019). This dataset builds upon earlier efforts to create a flash flood dataset which merges several data sources (Gourley et al. 2013), and is used to examine the performance of WPC's operational excessive rainfall outlooks (EROs; Erickson et al. 2021), as

well as other potential forecast guidance products such as the Colorado State University random forest (RF) systems (Herman and Schumacher 2018a; Schumacher et al. 2021).

Previous work has evaluated the correspondence between QPE compared with precipitation thresholds potentially of interest for the onset of flash flooding with reported flash flood events. Herman and Schumacher (2018c; HS18 hereafter) compared several common QPE datasets against fixed precipitation thresholds, average recurrence intervals (ARIs) from NOAA Atlas 14 and other sources, and FFG, examining their correspondence with both flash flood reports (FFRs) and NWS-issued flash flood warnings during a 2.5-year period. They found the best correspondence for 2.5 inches of precipitation in 24 h considering the CONUS as a whole, with regionally varying results for ARI exceedances. Gourley and Vergara (2021; GV21 hereafter) carried out a similar analysis using a more recent version of the multi-radar multi sensor (MRMS) product, finding best agreement with FFRs at shorter accumulation periods and much higher fixed thresholds, and also better performance for more sophisticated approaches such as ARI and FFG comparisons. Schumacher and Herman (2021) demonstrated that most of the differences in results between HS18 and GV21 were due to more frequent temporal sampling by GV21. There have also been a few studies of correspondence for smaller regions (e.g., Lincoln and Thomason 2018; Hammond 2018).

The purpose of this study is to extend the analysis of HS18 to a longer time period (seven years vs. 2.5 years), and to include, in the same analysis context, forecasts from a state-of-the-art convection-allowing modeling system. Comparing model quantitative precipitation forecasts (QPF) with various QPE products in this framework provides some guidance for forecasters seeking to use gridded model-based threshold exceedances in their forecasting operations. Although it would be instructive to include running totals to quantify agreement when including

overlapping time periods (as done by GV21), we focus on non-overlapping 1-h and 6-h periods in order to allow comparison with QPE datasets such as Stage IV. It is anticipated, as demonstrated by Schumacher and Herman (2021), that this will reduce the relative number of exceedances. Thus, the results presented here are likely to be more instructive for forecasting applications, rather than for real-time warning operations.

In the following section, the datasets used in the analysis are described. Section 2.2 outlines the methodology used for the analysis. Section 2.3 presents results, and section 2.4 provides a discussion and conclusions.

2.1. Datasets

As described by HS18, there are large uncertainties associated with defining the occurrence of a flash flood event, even within the US. They propose a simple contingency table framework for evaluating correspondence between QPE exceedances of different thresholds, and FFRs, keeping in mind all the uncertainties associated with FFRs. We adopt this framework herein to examine the frequency of these so-called "proxy" flash flood events in both QPE and QPF. In this section, we describe the datasets used to set a threshold for flash flooding. Table 1 shows the datasets evaluated in this study, in addition to the time periods included, and data availability. The remainder of this section describes the datasets included in this study.

2.1.1. Flash flood reports

In this study, we verify against flash flood reports (FFRs) obtained from the Iowa Environmental Mesonet (https://mesonet.agron.iastate.edu/lsr/). As documented in prior studies (e.g., Calianno et al. 2013; Clark et al. 2014; HS18), FFRs are subject to significant reporting biases related to population density and time of day, as well as biases related to NWS WFO reporting procedures. HS18 additionally compared against NWS flash flood warnings (FFWs),

Table 1. Datasets included in this study, with the corresponding analysis period. Percent complete indicates the fraction of times in which the CONUS grid is at least 90% spatially complete. See text for more description on treatment for FFG.

Dataset	Analysis period	Percent complete (days)	
		1 h	6 h
Stage IVv2 QPE	1 Jan 2015 – 13 Dec 2017 (1077 days)	100.0%	98.7%
Stage IVv3 QPE	13 Dec 2017 – 28 Jul 2020 (958 days)	100.0%	99.5%
Stage IVv4 QPE	28 Jul 2020 – 1 Jan 2022 (522 days)	100.0%	95.4%
CCPAv3 QPE	1 Jan 2015 – 18 Jul 2018 (1294 days)	-	98.9%
CCPAv4 QPE	18 Jul 2018 – 1 Jan 2022 (1263 days)	-	99.8%
MRMSv10 radar-only QPE	23 Feb 2015 – 1 Dec 2016 (647 days)	61.7%	81.9%
MRMSv11 radar-only QPE	1 Dec 2016 – 14 Oct 2020 (1413 days)	84.0%	94.8%
MRMSv12 radar-only QPE	14 Oct 2020 – 1 Jan 2022 (444 days)	80.6%	94.4%
HRRRv1 QPF	1 Jan 2015 – 23 Aug 2016 (600 days)	92.8%	95.7%
HRRRv2 QPF	23 Aug 2016 – 12 Jul 2018 (688 days)	94.6%	98.4%
HRRRv3 QPF	12 Jul 2018 – 2 Dec 2020 (874 days)	97.0%	99.0%
HRRRv4 QPF	2 Dec 2020 – 1 Jan 2022 (395 days)	95.7%	98.5%
FFG	1 Jan 2015 – 1 Jan 2022	97.7%	99.1%
ARIs	-	-	-

but demonstrated similar results when comparing against either FFRs or FFWs. As a result, we focus only on FFRs.

Figure 1 shows a map of the spatial distribution of FFRs during the seven-year period of record included in this study. Consistent with HS18's 2.5-year analysis, and with the 20-year

analysis of Ahmadalipour and Moradkhani (2019), FFRs are more common in the south-central and east-central US, with a secondary maximum in the southwestern US (along the lower Colorado River valley). In Texas, population influences and/or impacts of urbanized areas which are more prone to flash flooding are evident with the concentrations of FFRs in the Houston, Austin, and Dallas-Fort Worth areas, as well as near smaller cities such as Corpus Christi and Midland – Odessa. Similar urban effects are also evident elsewhere around the US.

The climatological frequency of "true" flash flood events is likely somewhat higher than reflected by the FFR dataset, and would likely benefit from a bias correction procedure similar to that developed by Potvin et al. (2019) for the tornado report dataset. However, such a procedure cannot correct bias by introducing FFRs for individual events. As a result, despite its deficiencies, we proceed with using FFRs as the "ground truth" data for our analysis.

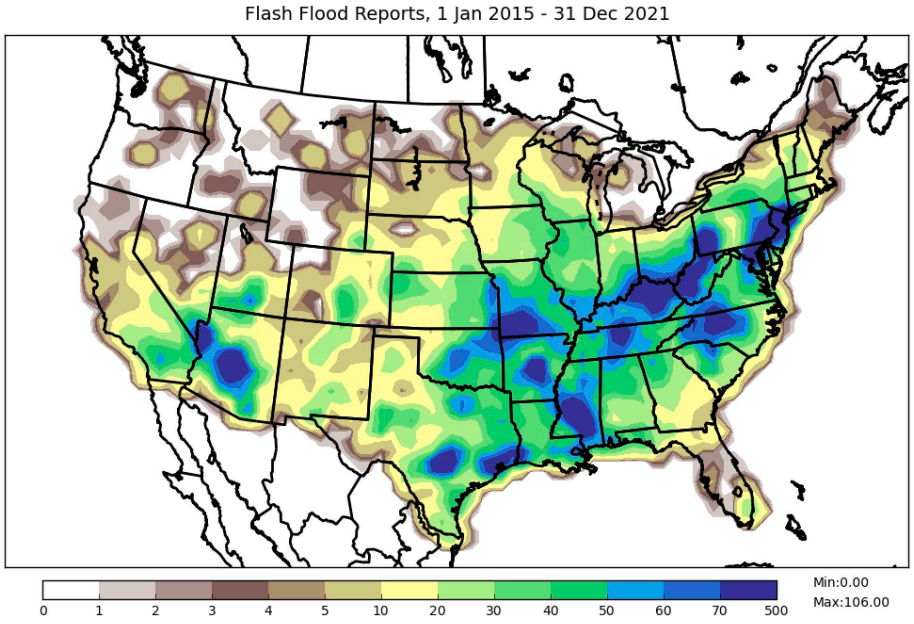

Fig. 1. Number of flash flood reports received during 2015 – 2021, on a 60 x 60 km grid.

2.1.2. Precipitation thresholds

Defining a precipitation threshold beyond which flash flooding occurs is a hydrologic problem. In this section, we describe the various approaches to defining a precipitation threshold in order to analyze correspondence with flash flood events.

2.1.2.1. FIXED PRECIPITATION THRESHOLDS

The simplest approach to defining a precipitation threshold is to use a fixed threshold. In this paper, we use several fixed thresholds in conjunction with different accumulation periods, based on previous studies.

2.1.2.2. AVERAGE RECURRENCE INTERVALS

As described by HS18, the use of ARIs is more complicated. Since the study of Herman and Schumacher (2018c), NOAA Atlas 14 (Bonnin et al. 2006; Perica et al. 2011, 2013a,b, 2015) has been updated for Texas (Perica et al. 2018), but still does not include the Pacific Northwest. For this reason, we use the approach of HS18 (described in their Appendix B) to estimate recurrence intervals in the Pacific Northwest for these accumulation periods. The ARIs are constructed from rain gauge observations with long records, using spatial statistics to estimate frequencies in regions of sparse observations.

Figure 2 shows the resulting ARIs for the 1-h and 3-h accumulation periods; these maps may be compared with HS18 Fig. 1. By definition, ARI values increase monotonically with increasing rarity. There is a spatial pattern with higher values in the southeastern US and lower values to the north, and especially in the interior western US. Comparing Fig. 2g with HS18's Fig. 1g illustrates the changes over Texas associated with the Atlas 14 update there, with more physical detail evident in the revised results (Fig. 2g).

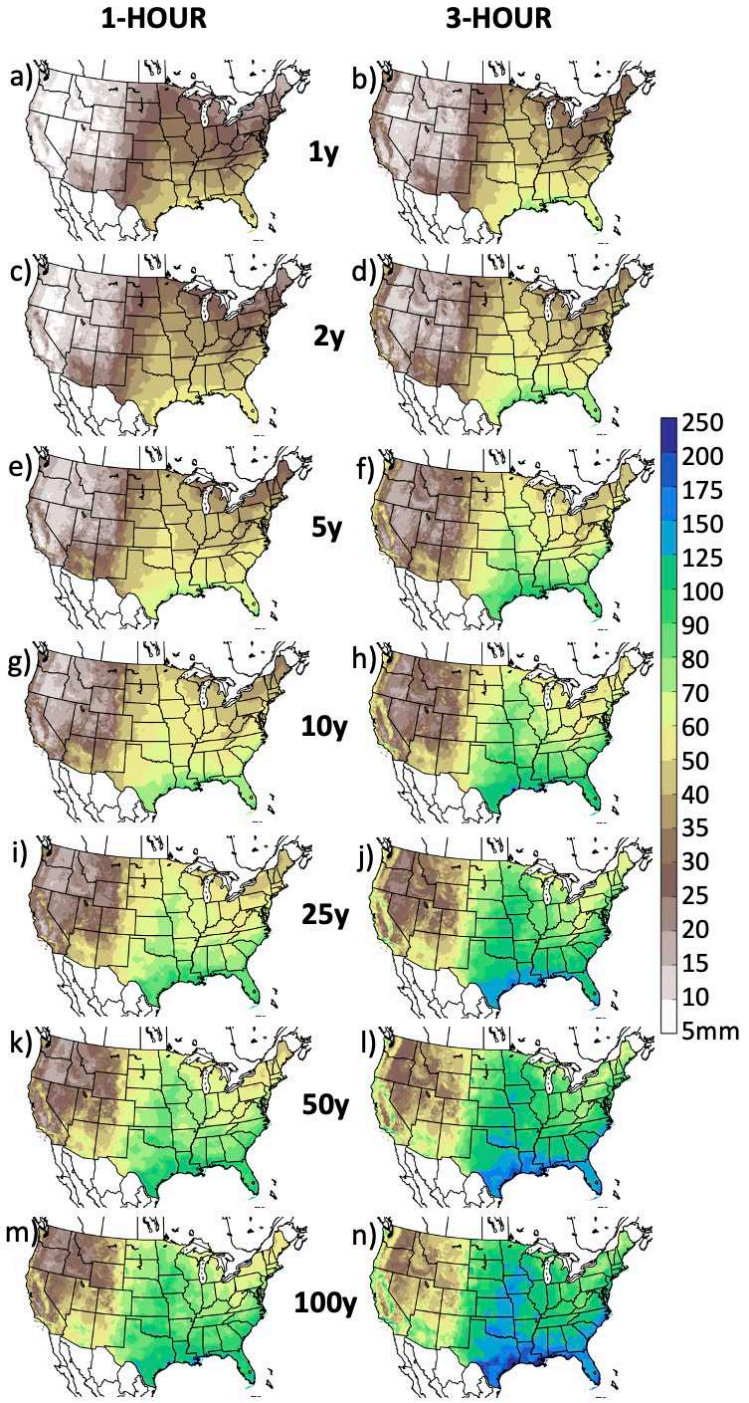

Fig. 2. Average recurrence intervals (ARIs), derived primarily from NOAA Atlas 14, for (left column) 1-h and (right column) 3-h durations, in mm. Shown are (a-b) 1-y, (c-d) 2-y, (e-f) 5-y, (g-h) 10-y, (i-j) 25-y, (k-l) 50-y, and (m-n) 100-y ARIs. Additional details on the derivation of the ARIs are provided in the text.

2.1.2.3. FLASH FLOOD GUIDANCE

Flash flood guidance (FFG) describes the fields produced by River Forecast Centers (RFCs), through various approaches, to provide guidance on probable amounts of precipitation over a given period of time required for the onset of bank full conditions on streams (Clark et al. 2014). The FFG construction methodology varies regionally around the US, leading to regionally varying performance characteristics. Figure 3 shows the median FFG value for the 7-year analysis period, as well as the 10th percentile, and the equivalent ARI recurrence interval; the figure may be compared with HS18's Fig. 2. Consistent spatial patterns emerge between Fig. 3 and HS18's Fig. 2, despite the difference in analysis period (7 years here vs. 2.5 years in HS18). Median 6-h FFG values range from less than 25 mm in the Pacific Northwest to greater than 150 mm in portions of the southern US. As shown by HS18, dramatic differences in FFG emerge across RFC boundaries. In particular, the Northwest RFC produces FFG that varies only slightly across accumulation interval, while the California Nevada and Colorado Basin RFCs' FFG increases dramatically from 1-h to 6-h accumulation (Fig. 3a-c). The same pattern is seen for the higher-risk 10th percentile FFGs (Fig. 3g-i). Finally, comparing the median and 10th percentile FFG values reveals that FFG is essentially constant in time in the western US (e.g., Fig. 3d-f, j-l); this is consistent with the use of the Flash Flood Potential Index in these regions, which is based on gridded physiographic information rather than soil moisture estimates (e.g., Clark et al. 2014).

Clark et al. (2014) carried out an analysis of FFG performance by evaluating Stage IV QPE vs. FFG and comparing against FFRs, finding critical success index (CSI) maximizing at 0.2 in the eastern CONUS. It is important to note that some of the low skill evident in the western CONUS in their analysis is likely due to shortfalls in the Stage IV QPE. They also

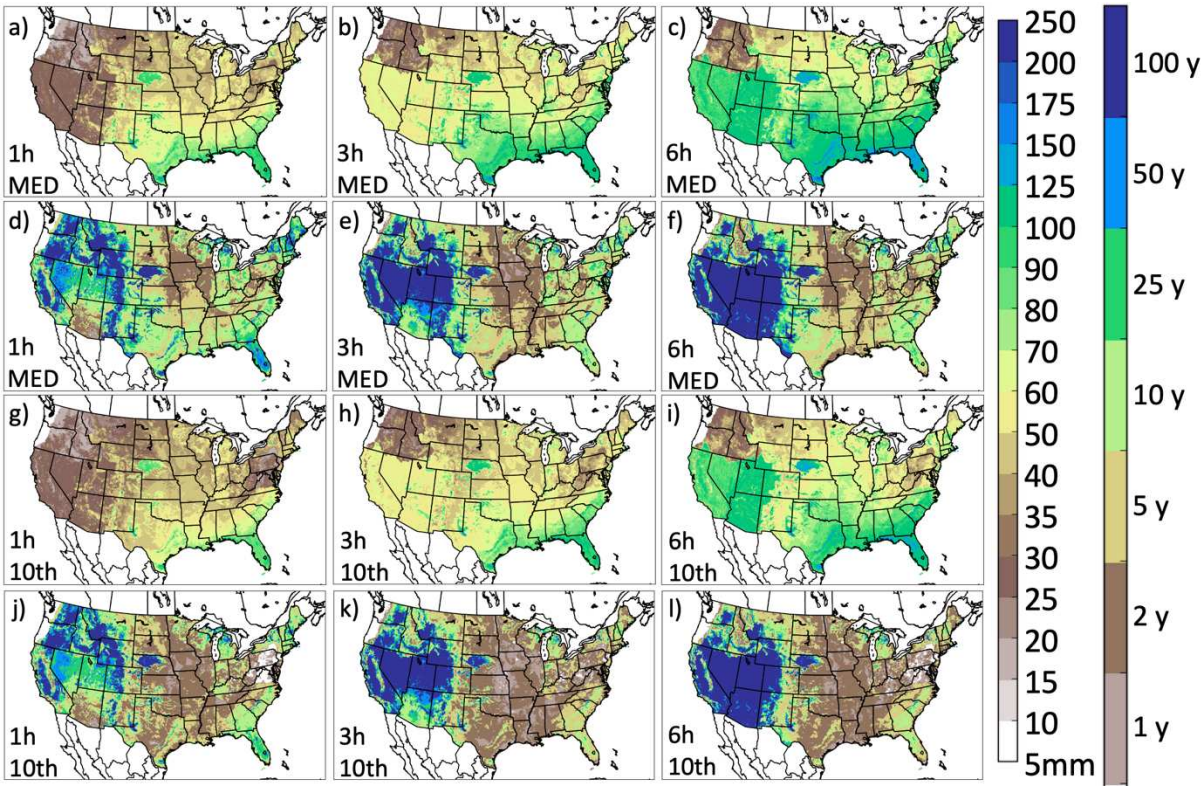

Fig. 3. (a-f) Median and (g-l) 10th percentile FFG estimates over the 7-yr period of record. Shown are (left column) 1-h, (middle column) 3-h, and (right column) 6-h FFG values. Panels (a-c) and (g-i) correspond to the actual threshold estimates, while (d-f) and (j-l) correspond to the equivalent ARIs to those thresholds to the particular gridpoint.

found that there was significant skill dependency upon the dataset of flash flood events used for verification.

For FFG in this study, special treatment was carried out to allow extension of the analysis back to 2015. Prior to July 2017, all FFG grids valid at 06 UTC were missing data for six RFCs covering the western, northern, and central CONUS, and FFG grids at 00 and 18 UTC were missing data for the three western RFCs. This was handled by using FFG values from the most recent valid time that provided values for the point in question, as long as it occurred during the previous 24 h. This allows us to achieve 99.1% data coverage for the 2015-21 period (see Table 1).

2.1.3. Quantitative precipitation estimates

In this section, we describe the various QPE datasets used in the analysis, building upon the findings of HS18.

2.1.3.1. STAGE IV

Stage IV is the RFC-produced precipitation analysis (Nelson et al. 2016). There are well-known Stage IV quality control (QC) issues, including discontinuities along RFC boundaries, and radar artefacts. The quality of Stage IV estimates varies by RFC, largely dependent upon the availability of radar and gauge observations. In addition, the hourly and 6-h Stage IV QPE values are not necessarily consistent. However, the Stage IV products are widely used in precipitation retrieval evaluation as well as model verification.

2.1.3.2. CLIMATOLOGY-CALIBRATED PRECIPITATION ANALYSIS (CCPA)

Because of the aforementioned weaknesses of Stage IV, particularly in the population of heavy to extreme precipitation events, another dataset has been developed which uses a simple linear regression model to adjust Stage IV towards the daily Climate Prediction Center (CPC) global gauge analysis. This dataset, referred to as the Climatology-Calibrated Precipitation Analysis (CCPA; Hou et al. 2014), corrects some of the biases of the Stage IV dataset, but retains the small-scale structure of precipitation events. Previous studies have documented how CCPA also tends to mute extreme values that are found in Stage IV (HS18).

2.1.3.3. MULTI-RADAR MULTI-SENSOR (MRMS) RADAR-ONLY QPE

The NSSL MRMS project aims to use data from ground-based radar and other sources to create a variety of user-focused analysis products, including QPE. The MRMS QPE, formerly entitled the National Mosaic and multi-sensor QPE (NMQ; Zhang et al. 2011), has undergone

extensive development over the past decade (e.g., Zhang et al. 2016; Qi et al. 2016). MRMS was implemented operationally in 2014.

The original radar-based MRMS QPE is described by Zhang et al. (2011). The QPE features four Z-R relationships, applied on a pixel-by-pixel basis. Since its original inception, improved QC measures have been applied, including use of polarimetric radar observations, and a vertical profile of reflectivity correction for bright banding (Zhang et al. 2016). Tang et al. (2020) describe more recent QC developments for radar-based QPE. In addition, Zhang et al. (2020) describe a dual-polarization radar synthetic QPE which has since been implemented as part of the operational MRMS radar-only QPE (GV21).

2.1.4. Quantitative precipitation forecasts

The HRRR is an hourly-updating convection-allowing model run operationally since Sep 2014, using community-supported data assimilation and model software (Dowell et al. 2022, hereafter D22; James et al. 2022). The HRRR produces hourly QPF, which has been evaluated against both QPE datasets and rain gauge observations in certain regions and for limited time periods (e.g., Ikeda et al. 2013, Bytheway and Kummerow 2015, Bytheway et al. 2017, Dougherty et al. 2021, and English et al. 2021). A comprehensive evaluation of HRRR QPF, including how it has changed between HRRR versions, is beyond the scope of this study, but work is underway to document this in a peer-reviewed article.

The HRRR initialization procedure is described in detail by D22 (section 3), and consists of several steps. Radar data are ingested in the context of latent heat application in four 15-min windows during a 1-h "pre-forecast" for each HRRR simulation (Weygandt et al. 2022). Following the radar DA, conventional observations are assimilated using an approach that varies by HRRR version (D22). The assimilation step also carries out a non-variational stratiform

cloud hydrometeor analysis step (Benjamin et al. 2021), which allows for realistic analysis and short-term prediction of cloud cover. Short-range HRRR forecasts exhibit some dependence on radar observations due to the use of radar data for initialization.

2.2. Methodology

Analysis of model QPF exceedances of various thresholds, in the context of flash flood prediction, has been done for several years as part of the annual Flash Flood and Intense Rainfall (FFaIR) experiment (e.g., Barthold et al. 2015). In this section, we describe the methodology employed here to examine the correspondence of QPE/QPF threshold exceedances with FFRs, following the general approach of HS18.

In contrast to HS18, who evaluated on the ~4 km HRAP grid, here comparison is done on the 3km HRRR grid. QPE products are interpolated to the 3km HRRR grid using the National Centers for Environmental Prediction (NCEP) ipolates library (https://www.nco.ncep.noaa.gov/pmb/docs/libs/iplib/ipolates.html). We used neighborhood budget interpolation, preserving precipitation maxima; we tested sensitivity to using ipolates budget interpolation, as well as the impact of preserving maxima versus doing average interpolation, finding minimal sensitivity in the results. LSRs are put on the closest HRRR gridpoint, and then projected onto multiple nearby HRRR gridpoints using a 40 km radius of influence, as in HS18. Both the point QPE / QPF exceedances and the projected LSRs are then upscaled to a 60-km grid for evaluation. HS18 used a 0.5-degree latitude – longitude grid; we tested using a latitude-longitude grid, finding only minor sensitivity for the results. Contingency table statistics were then calculated relating the QPE / QPF exceedances to the occurrence of flash flood reports.

In order to remove clearly erroneous QPE values, we followed the approach of Herman and Schumacher (2016b; their Appendix). This approach uses the fact that recurrence interval (ARI) threshold exceedances should occur with some specified frequency. Time series of QPE (for 1-, 6-, or 24-h durations) at each gridpoint were used to assess time-lagged correlations, and thus to determine the approximate number of independent events in each time series; based on this, we can assess the statistical likelihood of observing different numbers of ARI exceedances based on the binomial distribution. We remove QPE values exceeding the 99.99% percentile for any of the ARIs shown in Fig. 2. All QPE datasets were subject to this QC; model QPF was not subjected to it.

2.3. Results

The volume of data involved in this study, especially the variety of QPE / QPF sources, precipitation thresholds, and accumulation windows, is large, so we present here only a subset that is relevant to telling the story of comparison with HS18, as well as evaluating HRRR QPF in the same framework.

2.3.1. CONUS-wide results

In this section, we summarize our results in terms of CONUS-wide performance. We begin with some "heat maps", showing the frequency of exceedance of various thresholds; these spatial patterns can be compared with Fig. 1, which shows the frequency of FFRs during the period.

2.3.1.1. HEAT MAPS

Figure 4 shows exceedance counts of a single QPE dataset (6-h CCPA) against various fixed, ARI, and FFG ratios. Fixed thresholds, as the simplest formulation, exhibit the well-known climatology of heavy precipitation across the CONUS. The high precipitation thresholds

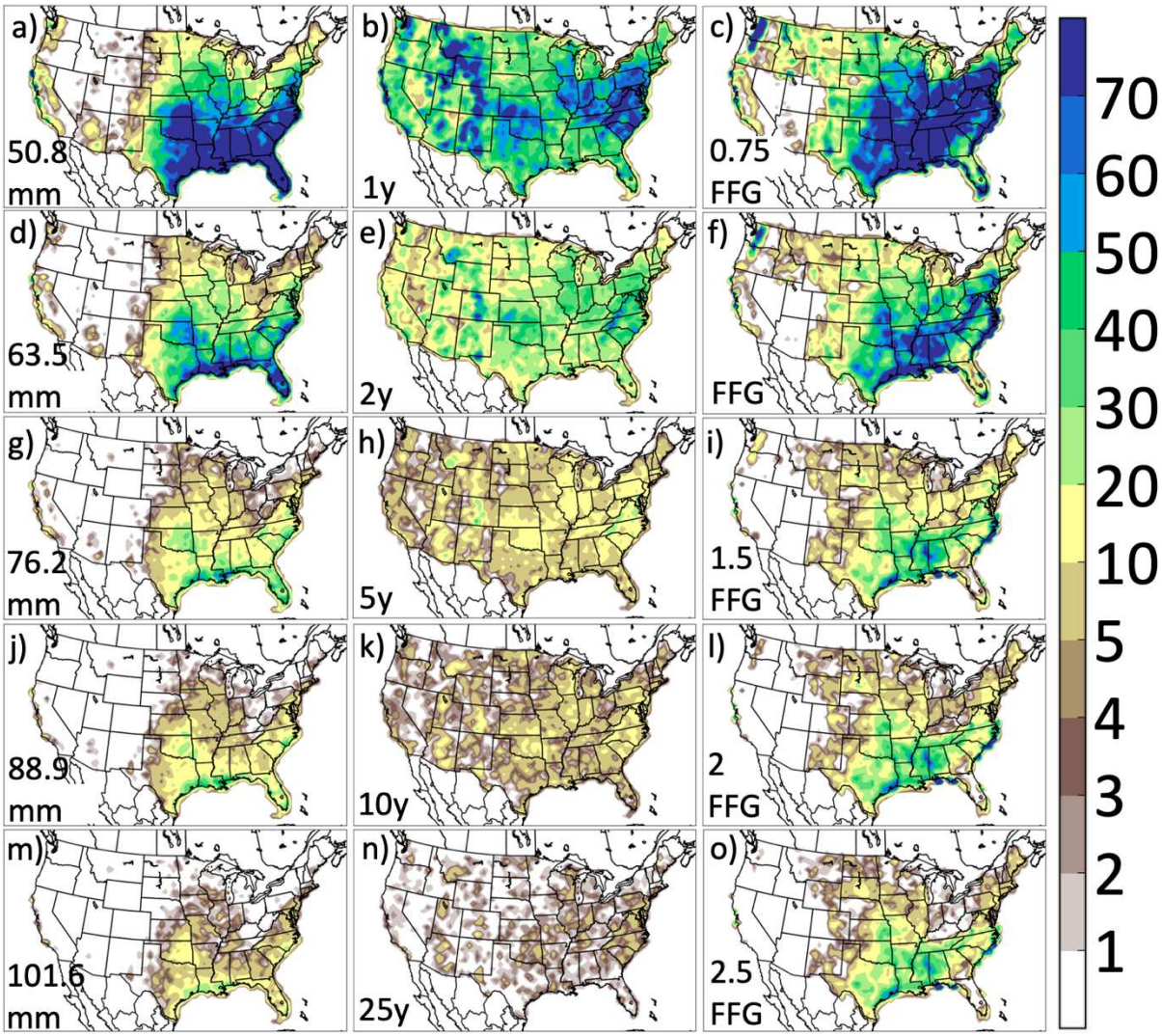

Fig. 4. Exceedance counts of 6-h CCPA during 2015 – 2021. Shown are (left column) fixed thresholds, (middle column) ARI thresholds, and (right column) FFG ratio thresholds. Specifically, panels (a, d, g, j, m) show 50.8, 63.5, 76.2, 88.9, and 101.6 mm $(6\ h)^{-1}$, respectively, (b,e,h,k,n) show 1, 2, 5, 10, and 25 y ARI thresholds for the 6-h duration, respectively, and (c,f,i,l,o) show 0.75, 1.0, 1.5, 2.0, and 2.5 FFG for the 6-h duration. 2519 days are included in the analysis.

are mostly confined to the Gulf coast region with tongues of higher probability of exceeding, for example, 76.2 mm $(6\ h)^{-1}$, extending northward along the Atlantic coast and into eastern Oklahoma (Fig. 4g). CCPA estimates exceeding 50.8 mm $(6\ h)^{-1}$ occur occasionally along the

Pacific coast and in the Sierra Nevada and Cascade ranges, and also in the Sonoran desert of Arizona (Fig. 4a). The ARI thresholds accomplish their purpose by somewhat normalizing frequency across the CONUS (e.g., Fig. 4b). However, maxima and minima are still evident, due to departures from climatology during this period, errors in CCPA, and/or biases in the ARIs. Exceedances of FFG (Fig. 4f) exhibit a somewhat different spatial pattern, due to the intended physical variability of the thresholds required for flash flooding, as well as artificial regional differences in FFG methodology. In general, FFG exceedances are more frequent, relatively, in the Midwest and Appalachians, and less frequent in Florida / southern Georgia and the Nebraska sandhills, than exceedances of 63.5 mm $(6\ h)^{-1}$ (Fig. 4d,f). Six-hour FFG is almost never exceeded in much of the southwestern CONUS (particularly Nevada, Utah, and Arizona; Fig. 4f).

Figure 5 shows comparisons of 6-h heat maps for different QPE / QPF datasets compared against representative fixed, ARI, and FFG thresholds. Striking differences emerge among the different datasets in this analysis. The Stage IV high bias in the New Mexico through Montana Front Range area discussed by HS18 is seen (Fig. 5a-c), particularly in the ARI exceedances (Fig. 5b), compared with CCPA exceedances (Fig. 5d-f) which may be expected to be close to reality in the eastern US due to the climatological correction. Yet the radar-only MRMS product shows even more ARI exceedances in this region than the Stage IV (Fig. 5h), and in fact the radar-only MRMS product has much more precipitation than the other datasets across most of the CONUS (Fig. 5g-i). The counts of HRRR forecasted 76.2 mm $(6\ h)^{-1}$ precipitation events appear spatially similar to the QPE datasets, although HRRR predicts more events than captured by CCPA or Stage IV in the eastern US, instead more closely matching the number of radar-only

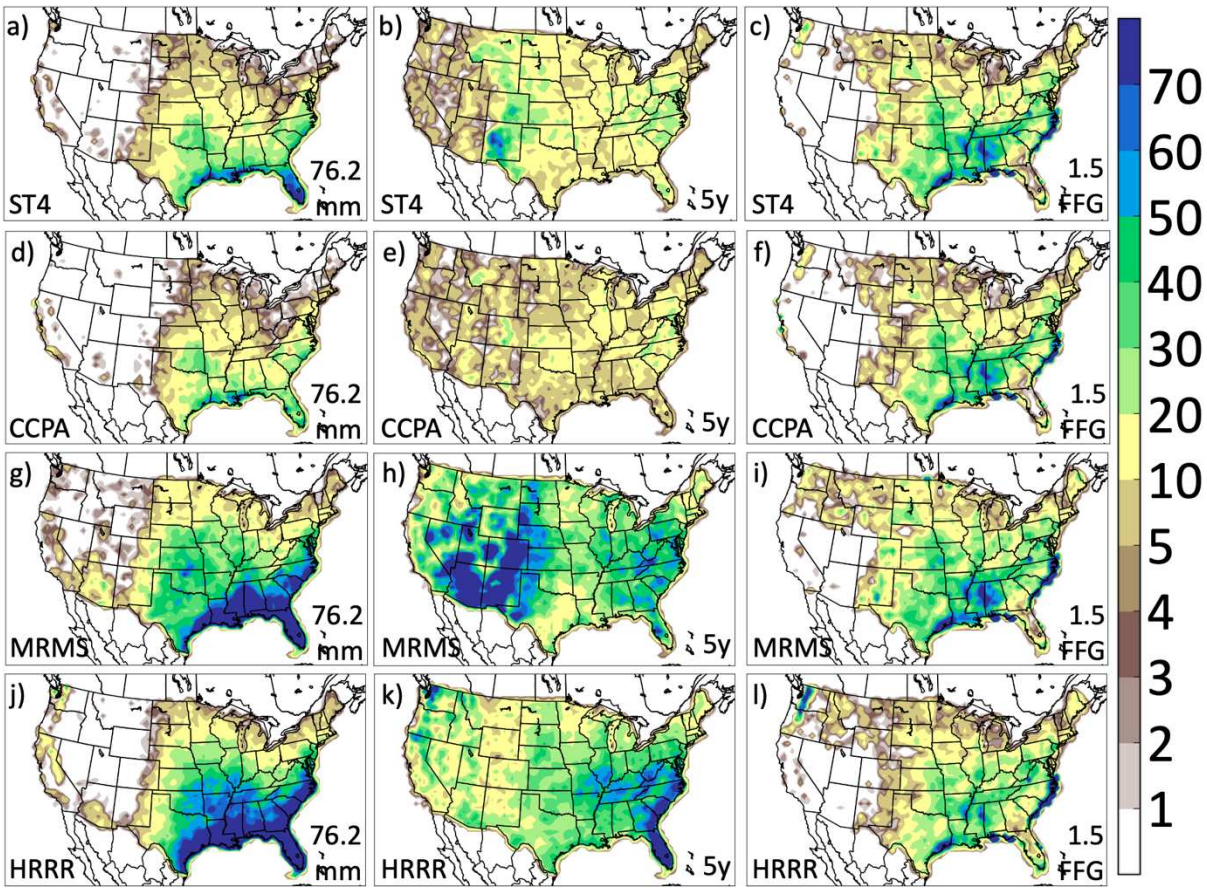

Fig. 5. Exceedance counts of 6-h QPE and QPF during 2015 – 2021. Shown are exceedances of (left column) 76.2 mm $(6\ h)^{-1}$, (middle column) 6-h 5 y ARI, and (right column) 1.5 6-h FFG. Products shown are (a-c) Stage IV, (d-f) CCPA, (g-i) radar-only MRMS, and (j-l) 0-6-h HRRR QPF. 2204 days are included in the analysis.

MRMS events (Fig. 5g,j). However, the HRRR predicts fewer 76.2 mm $(6\ h)^{-1}$ events than indicated by MRMS over the western US, with the exception of the Sierra Nevada and Cascade Range (Fig. 5g,j). The pattern and magnitude of FFG exceedances is similar among the datasets, indicating that FFG variability outweighs the importance of QPE / QPF differences (Fig. 5, right column). This suggests the correspondences of different QPE / QPF datasets vs. FFG thresholds against FFRs will vary more depending on the FFG ratio used, rather than on the precipitation dataset. HS18 evaluated QPE datasets against an FFG ratio of 1 only, while GV21 evaluated

MRMS only, so it is difficult to determine if this is consistent with prior studies. HRRR QPF has dramatically fewer ARI exceedances than MRMS in the western US (compare Fig. 5h,k).

For 1-h accumulations, results are broadly consistent with those seen for the 6-h duration (Fig. 6). In the western US, pronounced circles of higher frequency of MRMS exceedances of the 5-year ARI are seen around each WSR-88D (Fig. 6e); such radar artefacts are not as evident for the 6-h duration (Fig. 5h). We again see the relative high bias of MRMS compared to Stage IV (Fig. 6a,d). For the 1-h duration, we see relatively more events in the southwestern US, potentially reflecting the prevalence of short-duration extreme rainfall events in this region (left column of Fig. 6 vs. Fig. 5). The pattern of 1-h FFG exceedances (right column of Fig. 6) is somewhat different from the pattern of 6-h FFG exceedances (right column of Fig. 5), with 1-h exceedances appearing more uniformly distributed across the southern US in the QPE datasets, including in the desert southwest. The 1-h FFG exceedances seem to be in better agreement with the spatial pattern of flash flood reports than the 6-h FFG exceedances (cf. Fig. 1).

2.3.1.2. Correspondence Metrics

Figure 7 shows equitable threat score (ETS) for the dataset / threshold combinations shown in Fig. 5, illustrating the changes in correspondence with varying datasets and precipitation thresholds. ETS is calculated in a contingency table framework, with FFRs functioning as the observed events. ETS is formulated similarly to CSI, but is compared with a reference random set of events, such that positive values indicate better correspondence than a random set of events, and negative values indicate worse correspondence than a random set of events. Greater skill is evident in the east for all thresholds (Fig. 7). Note that ETS cannot be calculated at gridpoints where forecasted events never occur; these are evident as white

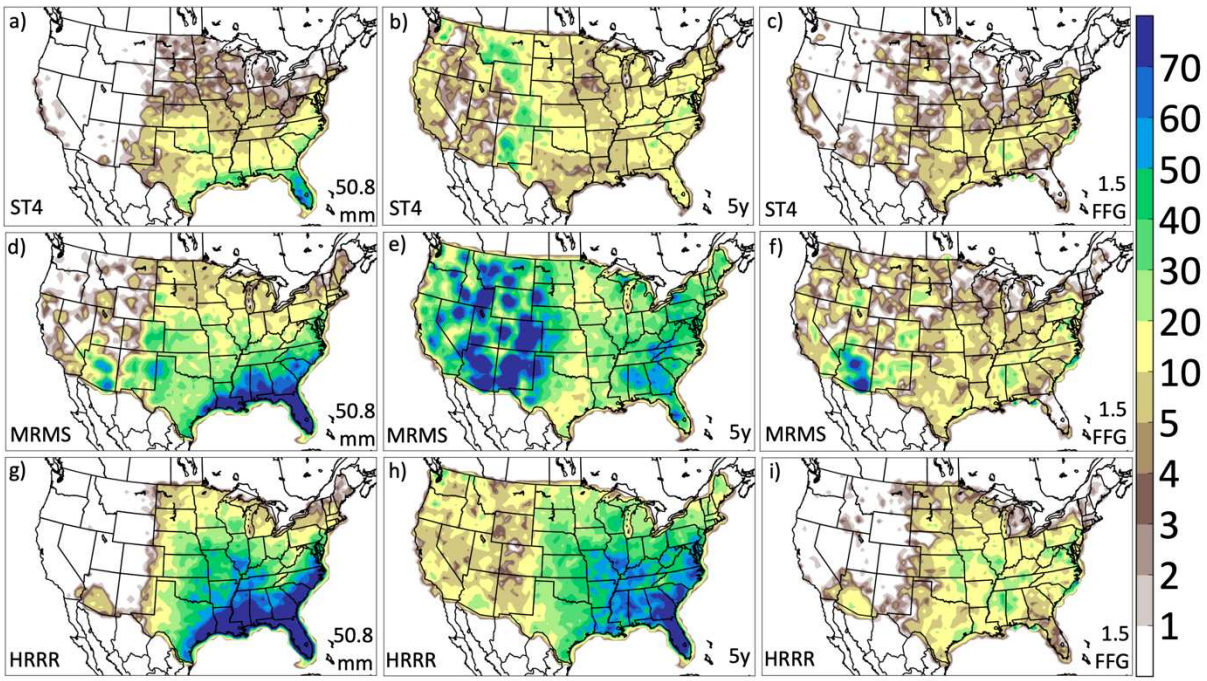

Fig. 6. Exceedance counts of 1-h QPE and QPF during 2015 – 2021. Shown are exceedances of (left column) 50.8 mm $(1\ h)^{-1}$, (middle column) 1-h 5 y ARI, and (right column) 1.5 1-h FFG. Products shown are (a-c) Stage IV, (d-f) radar-only MRMS, and (g-i) 0-1-h HRRR. 1227 days are included in the analysis.

gridpoints in Fig. 7. For example, Stage IV never exceeded the 76.2 mm $(6\ h)^{-1}$ threshold during 2015-21 in many places in the northwestern US (Fig. 7a). Overall, comparison with the static 76.2 mm $(6\ h)^{-1}$ threshold corresponds best with FFRs in the southern US for Stage IV and CCPA (Fig. 7a,d), while the 1.5FFG threshold appears to have the best correspondence in the northern US (Fig. 7c,f). For the southwestern US, MRMS and HRRR exceedances of ARIs appear to have the best correspondence with FFRs (Fig. 7h,k). Interestingly, HRRR exceedances of the 5-year ARI (Fig. 7k) have higher ETS than Stage IV or CCPA exceedances of the 5-year ARI in this region (cf. Fig. 7a,b,d,e). ARI thresholds appear to provide the best correspondence in the northwestern US, for all datasets (middle column of Fig. 7).

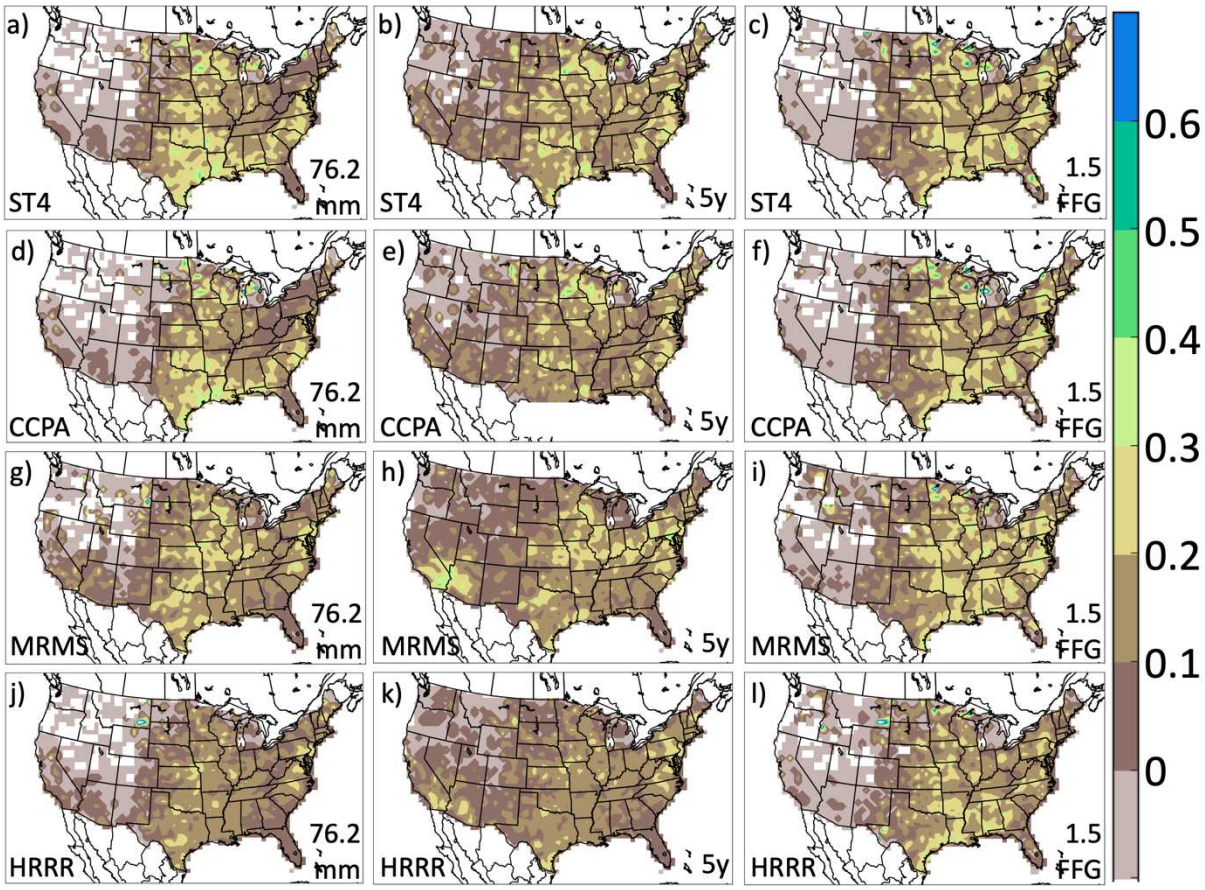

Fig. 7. Maps showing equitable threat score (ETS) for QPE/QPF exceedances of thresholds vs. observed flash flood reports (FFRs) during 2015 – 2021. Shown are exceedances of (left column) 76.2 mm $(6\ h)^{-1}$, (middle column) 6-h 5 y ARI, and (right column) 1.5 6-h FFG. Shown are (a-c) Stage IV, (d-f) CCPA, (g-i) radar-only MRMS, and (j-l) 0-6-h HRRR QPF. 2204 days are included in the analysis.

Figure 8 shows CONUS-wide ETS results for each dataset and thresholds. These results may be compared directly with Fig. 15 of HS18, and with Figs. 2-4 of GV21 (keeping in mind the more frequent temporal sampling by GV21). For the 6-h duration, ETS maximizes for the 50.8 – 63.5 mm $(6\ h)^{-1}$ for fixed thresholds (Fig. 8a), at the 2–5-year ARI (Fig. 8c), and an FFG ratio of 1-1.5 (Fig. 8e), with Stage IV providing slightly higher ETS than MRMS; these results agree well with HS18, although we find highest ETS for Stage IV exceedances of ARI thresholds and MRMS or HRRR exceedances of FFG thresholds (in contrast to HS18's finding of highest ETS for fixed threshold comparisons). For the 1-h duration, the highest ETSs are seen for FFG

a) 6-h Fixed Thresholds

Threshold	Stage IV	MRMS	HRRR 0-6h
25.4 mm (6 h)$^{-1}$	7.47	6.04	6.17
38.1 mm (6 h)$^{-1}$	10.16	8.68	7.69
50.8 mm (6 h)$^{-1}$	11.72	10.56	8.54
63.5 mm (6 h)$^{-1}$	11.69	11.22	8.61
76.2 mm (6 h)$^{-1}$	10.19	10.22	7.84
88.9 mm (6 h)$^{-1}$	7.98	9.29	6.57
101.6 mm (6 h)$^{-1}$	6.09	7.62	5.40
114.3 mm (6 h)$^{-1}$	4.63	6.27	4.33
127 mm (6 h)$^{-1}$	3.34	4.88	3.21

b) 1-h Fixed Thresholds

Threshold	Stage IV	MRMS	HRRR 0-1h
25.4 mm (1 h)$^{-1}$	9.02	6.96	7.79
38.1 mm (1 h)$^{-1}$	10.23	9.29	8.26
50.8 mm (1 h)$^{-1}$	7.71	9.51	7.60
63.5 mm (1 h)$^{-1}$	4.48	6.73	6.47
76.2 mm (1 h)$^{-1}$	2.28	4.26	4.93
88.9 mm (1 h)$^{-1}$	0.97	2.06	3.60
101.6 mm (1 h)$^{-1}$	0.36	1.00	2.11
114.3 mm (1 h)$^{-1}$	0.14	0.48	1.21
127 mm (1 h)$^{-1}$	0.05	0.19	0.51

c) 6-h ARI Thresholds

Threshold	Stage IV	MRMS	HRRR 0-6h
1y ARI	12.27	9.18	8.52
2y ARI	12.43	10.32	9.10
5y ARI	11.06	10.88	8.85
10y ARI	8.99	10.36	7.78
25y ARI	5.90	8.77	5.80
50y ARI	4.03	7.11	4.59
100y ARI	2.53	5.56	3.11

d) 1-h ARI Thresholds

Threshold	Stage IV	MRMS	HRRR 0-1h
1y ARI	10.33	6.95	7.95
2y ARI	10.66	8.00	8.62
5y ARI	9.08	8.73	8.86
10y ARI	7.27	8.48	8.60
25y ARI	4.68	7.17	7.51
50y ARI	3.00	5.97	6.31
100y ARI	2.07	4.77	5.21

e) 6-h FFG Ratio Thresholds

Threshold	Stage IV	MRMS	HRRR 0-6h
0.25 FFG	5.46	4.67	4.95
0.5 FFG	8.41	7.90	7.31
0.75 FFG	10.86	10.09	8.91
FFG	11.72	11.17	9.87
1.5 FFG	11.14	11.29	9.49
2 FFG	9.68	10.35	8.04
2.5 FFG	8.61	9.14	6.74
3 FFG	8.07	8.37	5.85
3.5 FFG	7.59	7.86	5.32
4 FFG	7.32	7.19	4.96
4.5 FFG	7.03	6.86	4.59
5 FFG	6.75	6.59	4.41

f) 1-h FFG Ratio Thresholds

Threshold	Stage IV	MRMS	HRRR 0-1h
0.25 FFG	5.38	3.92	4.96
0.5 FFG	9.62	6.76	7.96
0.75 FFG	12.43	9.73	9.82
FFG	12.04	11.43	10.78
1.5 FFG	8.24	10.32	9.51
2 FFG	6.11	7.84	6.93
2.5 FFG	4.92	6.01	5.15
3 FFG	4.25	4.87	4.07
3.5 FFG	3.56	4.02	3.31
4 FFG	3.12	3.34	2.76
4.5 FFG	2.81	2.94	2.37
5 FFG	2.48	2.62	2.08

Fig. 8. ETS (multiplied by 100) by dataset and threshold for (a,c,e) 6-h and (b,d,f) 1-h duration. Shown are (a-b) fixed, (c-d) ARI, and (e-f) ratios of FFG thresholds. Dataset / threshold combinations are color coded by ETS, with higher ETS being shaded darker green. Results are for the 2015 – 2021 period, with 1199 days included in the analysis.

exceedances for every dataset (Fig. 8f); this is in contrast to HS18 finding higher ETS for fixed thresholds at the 1-h duration (see their Fig. 15a-c,g). For fixed and ARI thresholds, we generally find higher scores for the 6-h duration (Fig. 8a,c) than for the 1-h duration (Fig. 8b,d), in agreement with both HS18 and GV21. In terms of ETS, Stage IV emerges with the highest CONUS-wide score for all types of thresholds and for both 1-h and 6-h durations, in general agreement with HS18. The highest ETS for Stage IV tends to be at lower thresholds than is seen in MRMS, for all threshold types and both durations; this is because of the generally higher QPE in MRMS. Higher QPE in MRMS overall leads to heavier QPE events matching the frequency of FFRs more closely than lighter QPE events.

To visualize correspondence between QPE / QPF exceedances and observed FFRs in a more wholistic fashion, Figure 9 shows performance diagrams (Roebber 2009) for all of CONUS, showing (left column) 6-h and (right column) 1-h durations. Pairs associated with a single QPE / QPF dataset are colored alike, with precipitation thresholds types grouped into the same panel. Before we discuss the differences between the datasets, there are some general characteristics of the performance diagrams worth noting. All results for a single dataset and threshold type exhibit a curve going from the upper left portion of the diagram (high probability of detection, POD, but also high false alarm ratio, FAR, for relatively light thresholds) to the lower middle portion of the diagram (low POD but with varying FAR by dataset for rare thresholds like the 100-year ARI). Fixed and ARI thresholds in general exhibit a similar appearance, with distinction in the slope of their performance diagram curves, going from minimal distinction between the various QPE/QPF datasets at the lowest precipitation thresholds (in terms of POD and success ratio, SR), but a greater distinction at the highest thresholds (in terms of SR). The different slopes represent the different datasets' climatologies of heavy

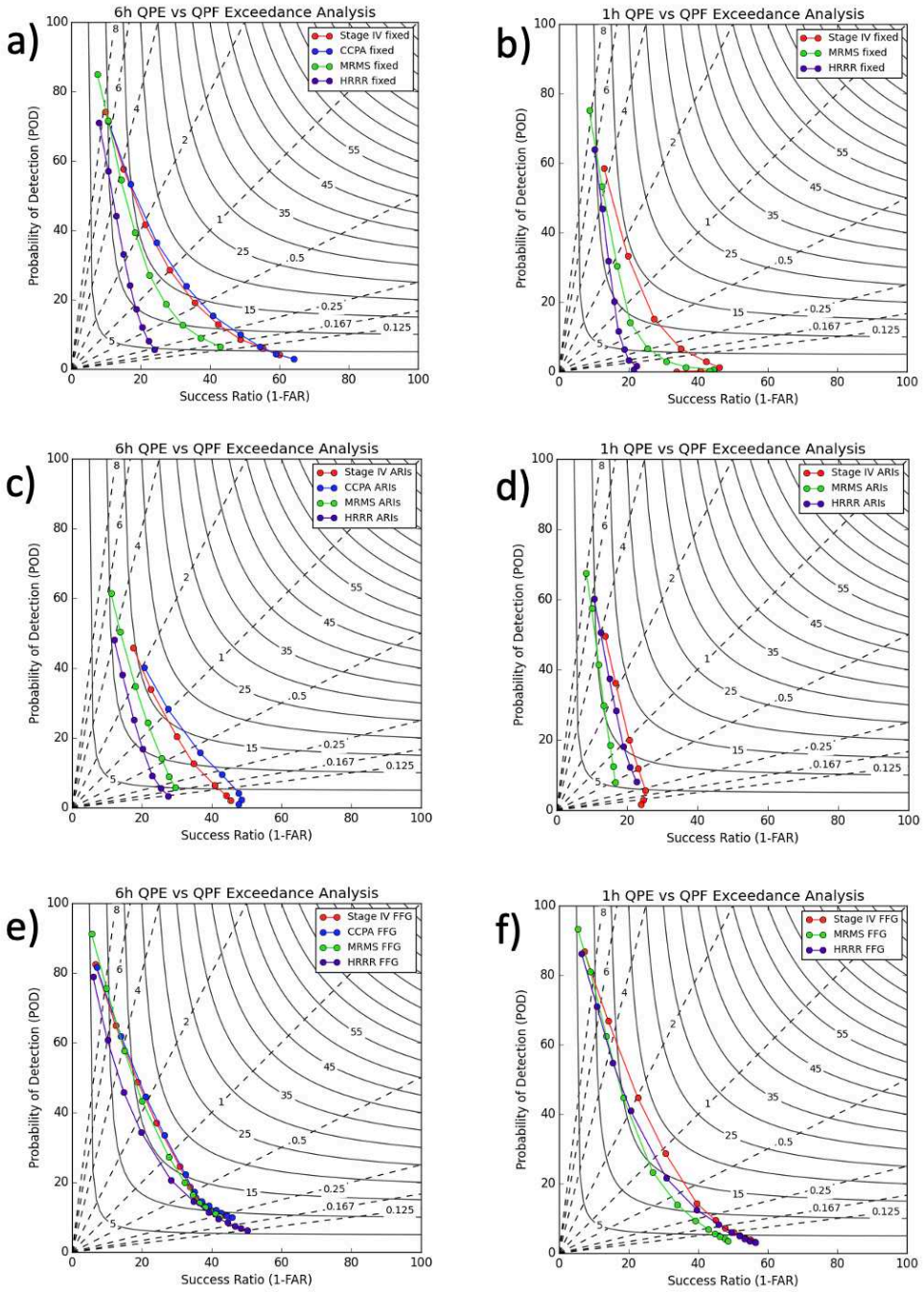

Fig. 9. Performance diagrams evaluating the degree of correspondence between QPE/QPF exceedances of (a-b) fixed thresholds, (c-d) ARIs, and (e-f) FFG, and observed FFRs. The evaluation period is Feb 2015 – Dec 2021 (1199 days included), for (left column) 6-h durations and (right column) 1-h durations. Thresholds shown are, from upper left to lower right of each panel, 25.4, 38.1, 50.8, 63.5, 76.2, 88.9, 101.6, 114.3, and 127 mm $(6\ h)^{-1}$; 1, 2, 5, 10, 25, 50, and 100 y ARIs; and 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5xFFG. Curved lines from upper left to lower right in each panel correspond to 100*critical success index (CSI), while dashed lines correspond to frequency bias.

precipitation amounts. For example, the HRRR QPFs have a higher FAR (lower SR) than CCPA at the heaviest precipitation threshold (127 mm [6 h]$^{-1}$; Fig. 9a, lowest points on the purple and blue curves), indicating that HRRR predicts many more 127 mm (6 h)$^{-1}$ precipitation events than are seen in CCPA. The high FAR in the HRRR also manifests as a lower CSI at these heavier thresholds. The FFG curves behave differently, because those thresholds are generally not static in time.

CCPA (blue curves) exhibits the greatest CSI overall (closest to top right) for all three types of thresholds for the 6-h duration (Fig. 9, left column). Stage IV QPE (red curves) is just slightly lower in CSI, but has an evident shift towards higher frequency bias (more frequent exceedances; Fig. 9, left column). MRMS corresponds to FFRs comparably to Stage IV / CCPA for fixed thresholds, but with a lower SR at the high precipitation thresholds (Fig. 9a), indicating more frequent heavy precipitation events in the MRMS dataset. In terms of FFG exceedances, MRMS corresponds almost as well as Stage IV and CCPA to FFRs (Fig. 9e).

For the 1-h QPE and QPF results (right column of Fig. 9), we see the same relative correspondence of the Stage IV and MRMS QPE for fixed threshold exceedances (Fig. 9b). For the 1-h ARI exceedances (Fig. 9d), we see less decrease in FAR with increasing threshold than was seen for the 6-h duration (Fig. 9c); this stems from relatively more frequent false alarms at the rare ARIs (100y ARI) for the 1-h duration compared to the 6-h duration. HRRR QPF exceedances of ARIs correspond better with FFRs than MRMS exceedances of ARIs at the 1-h duration (Fig. 9d). In general, HRRR QPF exceedances do not correspond to FFRs as well as the QPE datasets, which is not a surprising result given that it is a forecast rather than an observational estimate.

2.3.2. Regional correspondence variations

In this section, we examine regional variations in degree of correspondence between exceedances and FFRs. Figure 10 shows the region definitions used for the regional verification statistics shown in this section; these regions are identical to those used by HS18.

Figure 11 shows performance diagrams for the SW region (shown in Fig. 10). In this region, we see a much more pronounced difference between Stage IV and MRMS for all thresholds and both durations (Fig. 11), with MRMS comparisons exhibiting a much higher frequency bias than seen in the CONUS results (cf. Fig. 9). Comparing frequency bias between Stage IV and MRMS for the 25.4 mm $(1\ h)^{-1}$ threshold (Fig. 11b, uppermost points on the red and green curves), it is seen that MRMS contains ~5 times as many exceedances as Stage IV, and ~5 times as many exceedances as FFRs. HRRR QPF exceedances of this threshold, on the other hand, have a frequency bias near 1 when comparing against FFR occurrences. In general, for the southwestern CONUS, we see that HRRR QPF is competitive with the QPE datasets in terms of correspondence with FFRs. In fact, HRRR 1-h QPF exceeding the 2-year ARI has the highest CSI of any comparison for this region (Fig. 11d). These results are in agreement with previous studies documenting that our ability to model precipitation in sparsely-observed mountainous regions is overtaking the capabilities of our observations (Lundquist et al. 2019). These results can provide context for forecasters interpreting QPE datasets and CAM QPF in the southwestern US.

2.3.3. Summary

In order to summarize our quantitative comparison between QPE/QPF exceedances of various thresholds and FFRs, Fig. 12 shows the best-corresponding thresholds for Stage IV, MRMS, and HRRR QPF. For this evaluation, thresholds are considered optimal when they have

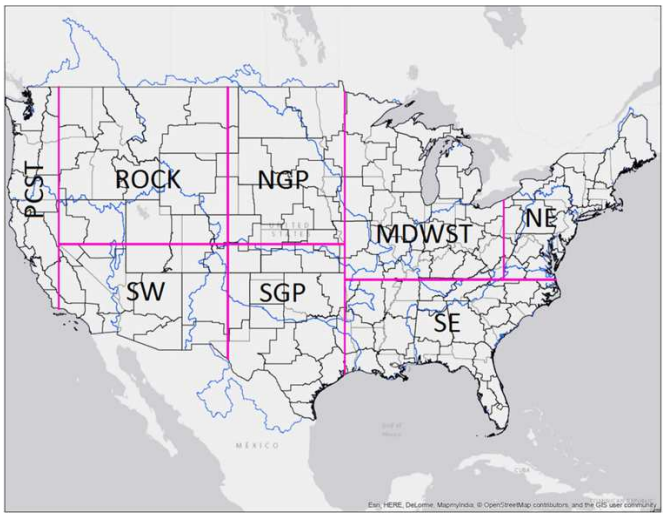

Fig. 10. Map of CONUS showing the eight regions used for correspondence evaluation.

the highest ETS, with a frequency bias falling between 0.5 and 2 (between predicting half as many events as FFRs up to twice as many events). Thresholds are colored green according to ETS, with darker colors associated with higher ETS, following the convention of HS18. For the CONUS as a whole, FFG exceedances emerge with the best correspondence to FFRs for all three datasets for the 1-h duration, and for MRMS and HRRR for the 6-h duration, which is encouraging given the additional information provided by FFG. For Stage IV exceedances, the 2-year ARI threshold has a slightly higher ETS than any FFG threshold.

Regionally, correspondence with FFRs is greatest in the eastern half of the CONUS (Fig. 12), in agreement with Fig 7. Lowest correspondence is seen for the PCST and ROCK regions, largely due to the relative infrequency of FFRs in these regions (cf. Fig. 1). Some interesting patterns emerge regionally in terms of the optimal 6-h thresholds to use for correspondence with FFRs (Fig. 12a). FFG comparisons become inferior to fixed and ARI thresholds in parts of the central and western US for the 6-h duration, with FFGs not corresponding best for any QPE or QPF dataset in the SW or ROCK regions. ARIs emerge as the best thresholds to use for all datasets in the SW region. Fixed 6-h thresholds find utility for several regions, despite their

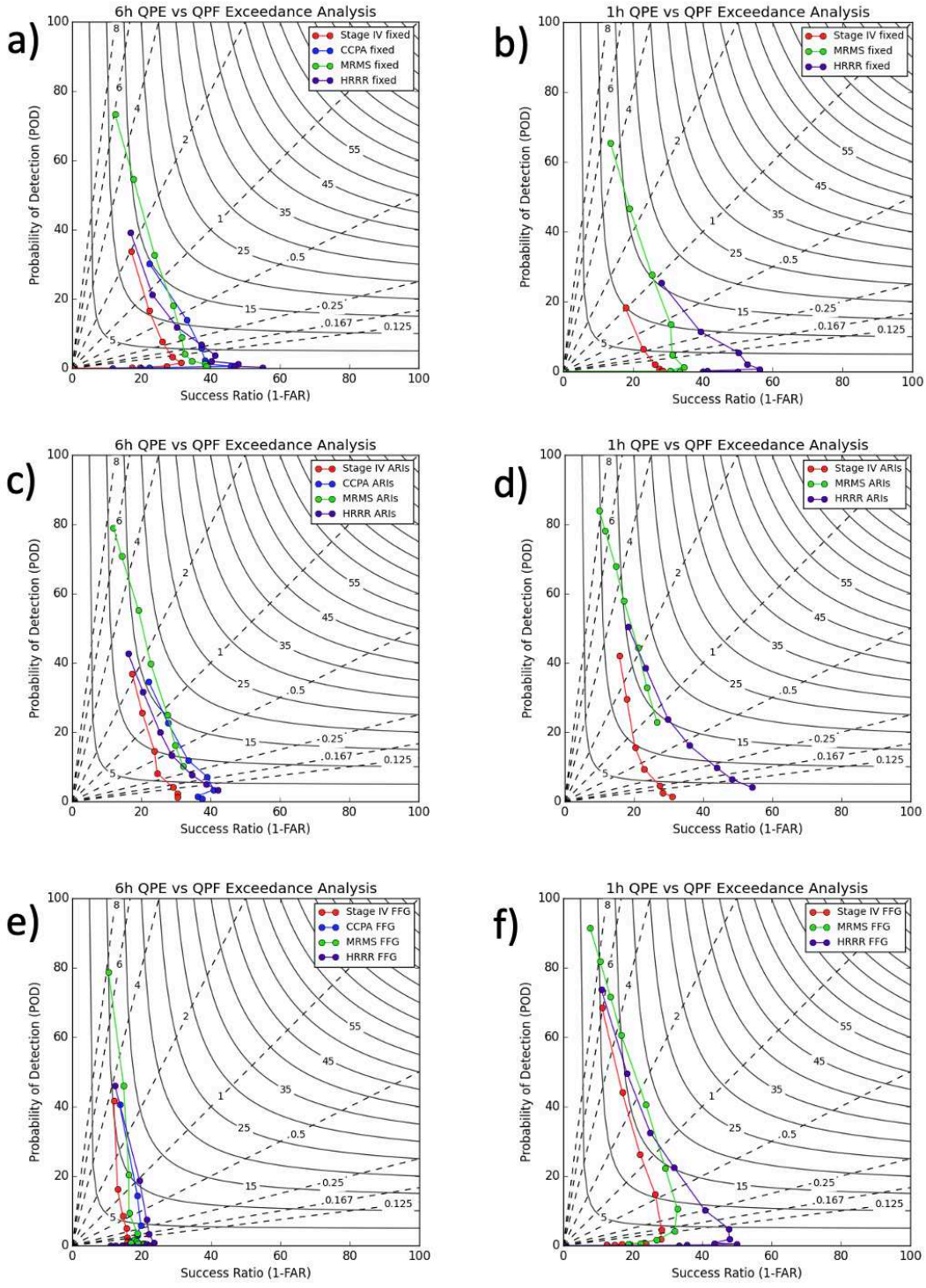

Fig. 11. As in Fig. 9, but for the southwestern CONUS, and showing (a,b) fixed, (c,d) ARI, and (e,f) FFG thresholds for (a,c,e) 6-h and (b,d,f) 1-h durations.

a) 6-h Thresholds

Region	Stage IV		MRMS		HRRR 0-6h	
CONUS	2y	12.43	1.5FFG	11.29	FFG	9.87
PCST	25y	4.91	1.5FFG	2.70	101.6mm	2.11
SW	2y	11.04	10y	13.75	2y	11.66
ROCK	25y	5.52	63.5mm	8.60	50y	6.03
NGP	1.5FFG	12.37	2.5FFG	11.48	2FFG	10.50
SGP	63.5mm	15.75	76.2mm	12.66	FFG	13.43
MDWST	FFG	17.86	FFG	16.50	FFG	13.82
SE	1.5FFG	18.13	2FFG	17.31	1.5FFG	15.45
NE	1y	19.79	FFG	17.38	FFG	14.41

b) 1-h Thresholds

Region	Stage IV		MRMS		HRRR 0-1h	
CONUS	0.75FFG	12.43	1.5FFG	10.32	FFG	10.78
PCST	2FFG	6.50	2FFG	3.03	100y	2.57
SW	2y	10.93	50y	12.83	2y	13.79
ROCK	38.1mm	4.40	50.8mm	7.48	25.4mm	6.06
NGP	FFG	10.93	1.5FFG	9.34	1.5FFG	8.89
SGP	FFG	14.40	1.5FFG	10.91	1.5FFG	11.20
MDWST	0.75FFG	17.69	FFG	15.36	1.5FFG	13.94
SE	FFG	16.92	FFG	14.38	1.5FFG	12.65
NE	2y	17.66	FFG	15.02	FFG	14.87

Fig. 12. As in Fig. 8, but comparing highest ETS thresholds for Stage IV, radar-only MRMS, and HRRR for each region with a frequency bias falling between 0.5 and 2, for (a) 6-h and (b) 1-h duration. The numbers and shading correspond to ETS multiplied by 100.

simple formulation; they have the best correspondence to FFRs for QPE (Stage IV or MRMS) exceedances in SGP, for MRMS exceedances in ROCK, and for HRRR QPF exceedances in PCST. Stage IV 6-h exceedances emerge with the top correspondence in all regions but SW and ROCK (where MRMS exceedances have better correspondence with FFRs).

Results for the 1-h duration are similar, but with some interesting differences (Fig. 12b). FFG comparisons emerge as the best threshold to use more often at this short duration, with fixed thresholds only finding utility in the ROCK region. ARI comparisons are the best threshold to use only for the SW, Stage IV in the NE, and HRRR in the PCST. Again, Stage IV emerges as the top dataset in all regions but SW and ROCK. MRMS exceedances have the best correspondence with FFRs in the ROCK region, and HRRR 0-1-h QPF exceedances correspond best in the SW region (consistent with results shown in Fig. 11).

2.4. Discussion and Conclusions

The correspondence of QPF / QPE datasets exceedances of precipitation thresholds with occurrences of flash floods is a complicated relationship. There are many reasons why we would not expect perfect correspondence, even with a somewhat sophisticated threshold such as FFG. However, the framework introduced by HS18 provides a way of quantitatively evaluating QPE datasets and thresholds for their relative value in flash flood analysis and forecasting, since they are the tools available to operational forecasters. In this article, we have extended the analysis of HS18 to a longer time period, and included, in the same framework, QPF from a state-of-the art CAM.

A key finding from this study, which is consistent with previous work, is that dramatic uncertainties persist in QPE, particularly in sparsely observed regions of the US. The major differences in population of heavy precipitation events between different QPE datasets is

concerning when these datasets are routinely used for many purposes, including model evaluation. As an example, Fig. 5 shows that, even in a relatively well-observed region like central South Carolina, 6-h CCPA contains 20-30 6-h QPEs exceeding 76.2 mm during a seven-year period (Fig. 5d), but MRMS contains 70+ events (Fig. 5g). The more frequent occurrence of heavy precipitation in MRMS also manifests in higher thresholds providing best correspondence with FFRs in the framework of this study. These uncertainties pose major challenges for both the research and operational communities. The MRMS team continues to refine their QPE algorithms using more sophisticated gauge correction and polarimetric information (e.g., Qi et al. 2016; Zhang et al. 2020); these and other innovative approaches are needed to improve the quality of QPE, particularly in the western US.

In agreement with HS18, we find that the skill of correspondences generally is highest in the eastern US, with lower skill in the west. We also find the same recurring deficiencies and biases reported by HS18, including the high bias of Stage IV in the interior western US, and the dependence of 1-h MRMS QPE upon proximity to radars. Consistent with HS18, we find that MRMS generally outperforms Stage IV and CCPA in terms of FFR correspondence in the western US for the 6-h duration, but with a much greater frequency of events. Stage IV exceedances have the highest correspondence with FFRs in the eastern US. We find that, at the 1-h duration, Stage IV exceedances have the best correspondence for almost every region. Exceedances for the 6-h duration have better correspondence with FFRs in all regions except the SW and PCST, where 1-h durations have higher correspondence.

In terms of thresholds, FFG is the best threshold for correspondence with FFRs for most dataset / region combinations and both for 1-h and 6-h durations (Fig. 12). This is true for the CONUS as a whole as well, in contrast to HS18's finding that fixed thresholds provided the best

correspondence for the CONUS. The correspondence of FFG exceedances with FFRs is encouraging, demonstrating the value of the dynamic FFG as a threshold for flash flooding onset. There are, however, some interesting exceptions to this result. In the ROCK region, FFGs do not provide the best correspondence for any dataset. For the 1-h duration, fixed thresholds provide the best correspondence for all datasets, with the specific threshold ranging from 25.4 mm $(1\ h)^{-1}$ for HRRR QPF to 50.8 mm $(1\ h)^{-1}$ for MRMS QPE (also illustrating the high frequency of exceedances in MRMS 1-h QPE). At the 6-h duration, for the ROCK region, relatively high ARI thresholds of 25 y (Stage IV) and 50 y (HRRR) provide the best correspondence. Overall, the ROCK region features the second-lowest correspondence between exceedances and FFRs.

The lowest correspondence is seen for the PCST region. This region is noteworthy for the relatively high thresholds that provide best correspondence with FFRs, including twice FFG for Stage IV and MRMS 1-h QPE, 100y ARI for 1-h HRRR QPF (Fig. 12b), and a fixed threshold of 101.6 mm $(6\ h)^{-1}$ for HRRR 0-6h QPF. The need for very high precipitation thresholds to obtain optimal (although still very poor) correspondence with FFRs stems from the rarity of FFRs in this region (Fig. 1); the requirement for a frequency bias between 0.5 and 1 for the results shown in Fig. 12 necessitates using an extremely trimmed down set of exceedances for any of the datasets shown here.

HRRR forecasts are evaluated here in the same framework as the QPE datasets, and as expected, HRRR QPF exceedances generally have inferior correspondence to FFRs for the CONUS scale for 1-h and 6-h durations. FFG is the best threshold with which to compare HRRR forecasts, both for 6-h and 1-h QPF. However, in certain poorly-observed regions like the SW, HRRR exceedances correspond better with FFRs than any QPE exceedance for the 1-h duration (Fig. 12). This is indicative of the relative skill of the HRRR in predicting short-

duration excessive rainfall events, compared with the relative lack of radar observations in this region. These results argue for the consideration of model QPF when determining a best estimate of QPE in regions of complex terrain and/or sparse observations.

As noted by HS18, it is important to acknowledge the problems associated with the FFR dataset. FFRs have an inherent low bias in rural regions and during the night, and are also subject to reporting differences between WFOs. It is likely that the true number of flash flood events is somewhat higher than that indicated by the FFR dataset, indicating that QPE / threshold comparisons featuring a frequency bias above unity may actually be superior to those with a bias of unity.

Another important issue to note is the limitation of our analysis to non-overlapping hourly 1-h QPEs, and 6-h QPEs between synoptic times; as demonstrated by GV21 and Schumacher and Herman (2021), this has the effect of reducing the number of events. The inclusion of additional MRMS QPE, to include "rolling average" QPEs ending at off-hour times, would be informative, but comparison with other datasets would not be possible.

This study (as well as HS18) has highlighted the regionally varying relationships between QPE / QPF and flash flood events. These variations are somewhat analogous to the varying UH thresholds used in predicting severe weather (Loken et al. 2020), and are an important consideration in the use of any QPE-based dataset for training a machine learning system to predict flash flooding (e.g., Hill and Schumacher 2021; Schumacher et al. 2021). Work also is underway to evaluate probabilistic QPFs from the High-Resolution Ensemble Forecast (HREF; Roberts et al. 2020) system in this framework; these results will be reported in a subsequent manuscript.

Our results highlight that flash flood forecasting is a highly probabilistic problem. Uncertainties are present in both the forcing (QPF) and the response (hydrology, or the threshold for flooding) components of the flash flood prediction problem; state-of-the-art flash flood prediction problems need to approach the forecast from this perspective. The use of a probabilistic FLASH system (Gourley et al. 2017) in combination with ensemble forecasts from the Warn on Forecast System (WoFS; Stensrud et al. 2009, 2013) is one such example, tested recently at the Hydrometeorology Testbed (Martinaitis et al. 2022). Use of convection-allowing ensemble systems, in combination with increasingly advanced hydrologic modeling, will continue to advance the skill of probabilistic flash flood forecast in the coming years.

CHAPTER 3: EXPLORING THE TREATMENT OF PREDICTORS FOR FORECASTING EXCESSIVE RAINFALL WITH RANDOM FORESTS BASED ON A DETERMINISTIC CONVECTION-ALLOWING MODEL

Applying machine learning (ML) approaches to high-impact weather prediction is an active area of research, with numerous candidate systems introduced in recent years (e.g., Herman and Schumacher 2018a; Sobash et al. 2020; Burke et al. 2020; Chapman et al. 2022), and formal evaluation occurring at testbed experiments including NOAA's Hazardous Weather Testbed (HWT; Clark et al. 2021) and Hydrometeorological Testbed (HMT; Trojniak and Correia 2021). Also recently, deterministic and ensemble convection-allowing modeling (CAM) systems have reached a level of maturity where they can, to some extent, directly predict high-impact events (or their storm attribute indicators; Dowell et al. 2022; James et al. 2022; Roberts et al. 2020). In addition to direct prediction of hazardous weather, CAM forecasts have been successfully used as predictors within ML systems in order obtain even more skillful forecasts. Random forests (RFs) show promise for these predictions, both from global ensembles and from CAMs. Successful RF based prediction systems span a wide range of applications, from predicting fog and low visibility for aviation (Herman and Schumacher 2016a), to severe weather forecasting (Loken et al. 2020; Hill et al. 2020), flash flood prediction (Herman and Schumacher 2018a,b; Hill and Schumacher 2021), non-convective windstorm forecasting (Brothers and Hammer 2022), and frontal analysis (Justin et al. 2023).

An initial RF system for prediction of excessive rainfall based on the Global Ensemble Forecast System (GEFS) reforecast (Herman and Schumacher 2018a,b) was transitioned into operations at the Weather Prediction Center (WPC) in 2019 (Schumacher et al. 2021), with day-one forecasts operational since 2020. This system has been demonstrated at the annual Flash

Flood and Intense Rainfall (FFaIR; Barthold et al. 2015) experiment for a number of years, and performs competitively with the operational WPC excessive rainfall outlook (ERO). Since 2018, the same RF framework has been applied to day-one forecasting based on a deterministic CAM, the NSSL-WRF, with the RF configuration largely similar to the GEFS-based system (Hill and Schumacher 2021). Overall, the NSSL-WRF based system does not perform as well as the GEFS based system for the day one period; preliminary testing with a similar system based on the High-Resolution Rapid Refresh (HRRR; Dowell et al. 2022; James et al. 2022) also indicates objectively and subjectively inferior performance to the GEFS-based system (Trojniak and Correia 2021). An outstanding research question is why, to date, deterministic CAM-based RF systems have not performed as well as those based on coarse global ensembles.

One hypothesis for the inferior performance of CAM-based RFs has to do with the assembly of predictors, to which random forests are sensitive (e.g., Sobash et al. 2020). The predictor assembly approach needs to be carefully considered based on characteristics of the input datasets as they relate to the representation of the target phenomenon. Foremost among concerns is the grid spacing (horizontal and vertical), since, for numerical weather prediction, the scale of phenomena able to be represented on the model grid depends on the grid spacing. For example, a global model with 13 km grid spacing will not contain convective-scale features (e.g., Weisman et al. 1997), whereas a 3-km convection-allowing model (CAM) will represent convective scales of motion. At the same time, the higher resolution of CAMs may introduce more redundant information: for example, is the temperature at one 3-km model grid point adding new information that isn't contained at the neighboring point? For a machine learning system such as an RF, it is important to ensure that the signal (for example, model indicators of excessive rainfall in a particular region and day) is adequately captured in the predictors, while

balancing the computational cost of processing higher-resolution output with incorporation of the most relevant inputs.

Herman and Schumacher (2018a) describe extensive sensitivity tests with GEFS-based excessive rainfall prediction models for days two and three, showing regional variability in the results of their tests. In particular, they found that, in most regions, predictions were relatively insensitive to using a 6-h vs. 12-h time step for the predictors. Forecasts were also relatively insensitive to the spatial predictor radius used (although the northeastern US and Pacific Coast regions had improved skill with a broader radius). They also found some sensitivity to the way ensemble information was utilized. However, it is unlikely that these results will hold for an RF using much higher resolution CAM-based predictors. Performing similar experiments for a CAM-based RF would be informative.

Other researchers have explored the formulation of predictors for CAM-based machine learning for severe weather prediction. Loken et al. (2020) describe probabilistic forecasting of day-one severe weather based on an RF using a CAM ensemble, the High-Resolution Ensemble Forecast (HREF) system. They use temporal and spatial aggregation to reduce the dimensionality of the 4-km input dataset, treating environmental and storm attribute fields differently, and also computing ensemble distribution metrics as separate input variables. With their 80-km grid spacing prediction model, they achieve superior performance to operational severe weather outlooks by the Storm Prediction Center (SPC). In a separate effort to use RFs to correct ensemble-based probabilistic precipitation forecasts, Loken et al. (2019) found that using the ensemble mean of each predictor performed just as well as using the predictors from individual ensemble members. More recently, Loken et al. (2022) carried out more extensive experiments exploring the value of retaining individual members from a CAM ensemble vs.

using an ensemble mean value at neighboring gridpoints, finding that use of the ensemble mean at the neighboring gridpoints led to superior performance. They also found that, while ensemble-based storm attribute predictors add little skill for the day two period and beyond, they are quite valuable for the day one period. Clark and Loken (2022) similarly used the Warn-on-Forecast System (WoFS; Lawson et al. 2018) to predict severe weather occurrence using RFs, exploring the contributions of storm attribute and environmental type variables, as well as different ensemble summary predictors.

Sobash et al. (2020) describe a neural network system for severe prediction, based on deterministic 3km WRF runs, in which the mean of the upper air and environmental fields in an 80 km box was used, but the maximum of explicit storm attribute fields. They also include additional spatial and temporal neighborhood means and maxima of explicit and environmental fields, and quantified the impact of their inclusion, finding degraded forecasts when those predictors were not included. They demonstrate that the neural network system, since applied in real-time to the HRRR (R. Sobash, personal communication), is superior to the alternative "surrogate severe" approach described by Sobash et al. (2011).

The goal of this article is to describe a set of controlled sensitivity experiments exploring the treatment of deterministic CAM forecasts as inputs to RFs for excessive rainfall prediction. In particular, we test the impact of using information within a spatial or temporal window, the impact of using finer time step predictor information, and the use of time-lagged ensemble information. This work complements prior studies, which have largely focused on use of CAM ensembles, and on the prediction of the severe convective hazards of tornadoes, hail, and severe convective wind. In exploring how best to condense CAM forecasts for RF predictions, we aim

to shed light on the reasons for inferior performance of CAM-based RFs compared to the GEFS-based RF for excessive rainfall.

The remainder of this article is organized as follows. Section 3.1 outlines the design of the HRRR-based RF system, with which the experiments are carried out. We then describe the sensitivity experiments in section 3.2, results in section 3.3, and discussion and conclusions in section 3.4.

3.1. System Design

RFs (Breiman 2001) are a supervised ML technique consisting of an ensemble of decision trees. Individual decision trees are set up by the algorithm based on their ability to discern between events and non-events. Probabilistic forecasts can be constructed by tallying the resulting prediction from each of the many branches of the RF. The system employed here is described in detail by Herman and Schumacher (2018a,b) and Hill and Schumacher (2021), with differences from their approach noted here. We employ eight independent RF models, one for each of the eight CONUS sub-regions shown in Fig. 10, and the resulting predictions are smoothed at the regional boundaries to avoid sharp gradients or discontinuities. In the following sub-sections, we describe the various components needed for the final trained model: predictands, predictors, and model training, followed by a description of the verification metrics employed in this study.

3.1.1. Predictand assembly

It is critical to define a high-quality predictand (or target vector) for a good RF system. For flash flood prediction, this is a non-trivial problem. Since the product is intended for use at WPC, one option is the current definition of the outlook: probability of precipitation exceeding flash flood guidance (FFG; Sweeney 1992). However, FFG is subject to large differences in

methodology, and thus discontinuities, between different River Forecast Centers (RFCs; Clark et al. 2014; Burke et al. 2023). A simpler approach would be to use flash flood reports (FFRs) as the predictands, but FFRs are subject to substantial regional biases, as described in detail by Herman and Schumacher (2018c). Alternatively, one can follow the approach of Hill and Schumacher (2021), also described by Schumacher et al. (2021), to take advantage of quantitative precipitation estimate (QPE) exceedances of average recurrence intervals (ARIs) to augment FFRs in defining excessive rainfall events. Table 2 describes the predictands constructed for the HRRR-based system described herein, with regionally varying target vectors assigned for the various RF models. While this general approach is somewhat subjective, it provides a configuration for controlled experiments and has been demonstrated to perform reasonably well; Hill and Schumacher (2021) describe sensitivity experiments varying the target vectors for the NSSL-WRF based excessive rainfall RF. The derivation of the ARIs is described by Herman and Schumacher (2018c).

3.1.2. Predictor assembly

Building off Hill and Schumacher (2021), the HRRR-based system uses similar output variables from the HRRR as are used for the NSSL-WRF. We used the same variables (see their

Table 2: Regionally varying target vectors used for the sensitivity experiments described herein. FFR refers to Flash Flood Reports, CCPA refers to Climatology Corrected Precipitation Analysis, and ST4 refers to Stage IV QPE, while the number of years refers to the threshold average recurrence interval (ARI) used.

Region	**Target Vector**
PCST	FFR+CCPA, 2-year
ROCK	FFR+CCPA, 2-year
SW	FFR+CCPA, 2-year
NGP	FFR+CCPA, 1-year
SGP	FFR+CCPA+ST4, 2-year
MDWST	FFR+CCPA, 1-year
SE	FFR+CCPA+ST4, 1-year
NE	FFR+CCPA+ST4, 2-year

Table 1), with the exception of a few variables defined differently (Table 3). In particular, accumulated precipitation in the HRRR is output at 1-h intervals, so the HRRR-based model is trained on run total accumulated precipitation rather than the 3-h accumulation. This is not an ideal configuration because precipitation early in the 24-h period will be included in the predictor array many times, while precipitation in the 09-12 UTC period (33-36 h forecast) will only be used once; sensitivity experiments are underway to quantify the impact of this misconfiguration. The use of updraft helicity is expanded from Hill and Schumacher (2021) through use of 1-h max and 1-h min values for 2-5 km updraft helicity, which has been previously used in RFs for precipitation prediction due to its association with sustained rotating storms (Nielsen and Schumacher 2018, 2020; Smith et al. 2023). We use 0-6 km wind shear values, rather than the mean 0-6 km wind components as in Hill and Schumacher (2021). Finally, we use 700-hPa vertical velocity instead of 0-3 km average vertical velocity. In addition to these meteorological predictors, we use the same static inputs as Herman and Schumacher (2018a) and Hill and Schumacher (2021), which are related to the climatological likelihood of excessive rainfall.

Table 3: Meteorological predictors using in training and forecasting with the RF models.

Predictor	**Description**	**Type**
APCP	Run total accumulated precipitation (kg m^{-2})	STORM
W700	700-hPa vertical velocity (m s^{-1})	STORM
UHMAX	1-h maximum 2-5 km updraft helicity (m^2 s^{-2})	STORM
UHMIN	1-h minimum 2-5 km updraft helicity (m^2 s^{-2})	STORM
CAPE	Surface convective available potential energy (J kg^{-1})	ENV
CIN	Surface convective inhibition (J kg^{-1})	ENV
PWAT	Precipitable water (kg m^{-2})	ENV
MSLP	Mean sea level pressure (Pa)	ENV
T2M	2-m temperature (K)	ENV
Q2M	2-m specific humidity (kg kg^{-1})	ENV
U10	10-m latitudinal horizontal wind speed (m s^{-1})	ENV
V10	10-m longitudinal horizontal wind speed (m s^{-1})	ENV
Z500	500-hPa geopotential height	ENV
USHR6000	0-6-km average latitudinal vector wind difference (m s^{-1})	ENV
VSHR6000	0-6-km average longitudinal vector wind difference (m s^{-1})	ENV

Sensitivity experiments addressing the temporal and spatial aggregation of predictors are described in section 3, but we provide here a brief description of the control experiment configuration, also used by Hill and Schumacher (2021). The model prediction grid is based on an earlier version of the GEFS, with resolution T254L42 (~55 km at 40° latitude; Herman and Schumacher 2018a). Predictors for each gridpoint are taken from the closest 3-km HRRR gridpoint, and also every 10 HRRR gridpoints (30 km) out to 180 km from the prediction point, corresponding to a radius (n) of 6 gridpoints on the GEFS grid; n is hereafter referred to as the predictor radius. The model predictors are then collected in space at 3-h intervals over the 24-h period from 12 UTC to 12 UTC, based on 12- to 36-h forecasts from the 00 UTC HRRR simulation, corresponding to the time period associated with WPC EROs. The number of predictors per training label is then $N = pt(2n + 1)^2$, where t is the number of forecast times (t = 9 here), amounting to 1521 predictors per variable p, and N = 22815 HRRR predictors per training example. The spatial predictor assembly procedure is illustrated in Fig. 13.

For the HRRR-based system, we gave some additional consideration to the masking out of non-land areas in each of the regions shown in Fig. 10. Since our target vectors are based on flash flood reports and QPE exceedances of ARIs (Table 2), it is important to exclude from the training dataset any model prediction for a non-land point, where flash floods can never be observed and ARIs are undefined. This eliminates mis-training, where the RF learns that a certain meteorological pattern is less likely to be associated with flash flooding simply because it occurred over an offshore or Great Lake area. We eliminated all predictor points over oceans or Great Lakes for the experiments described here, amounting to a reduction in number of predictor points by 0-39%. The largest reduction was for the NE region (see Fig. 10), with the land-locked

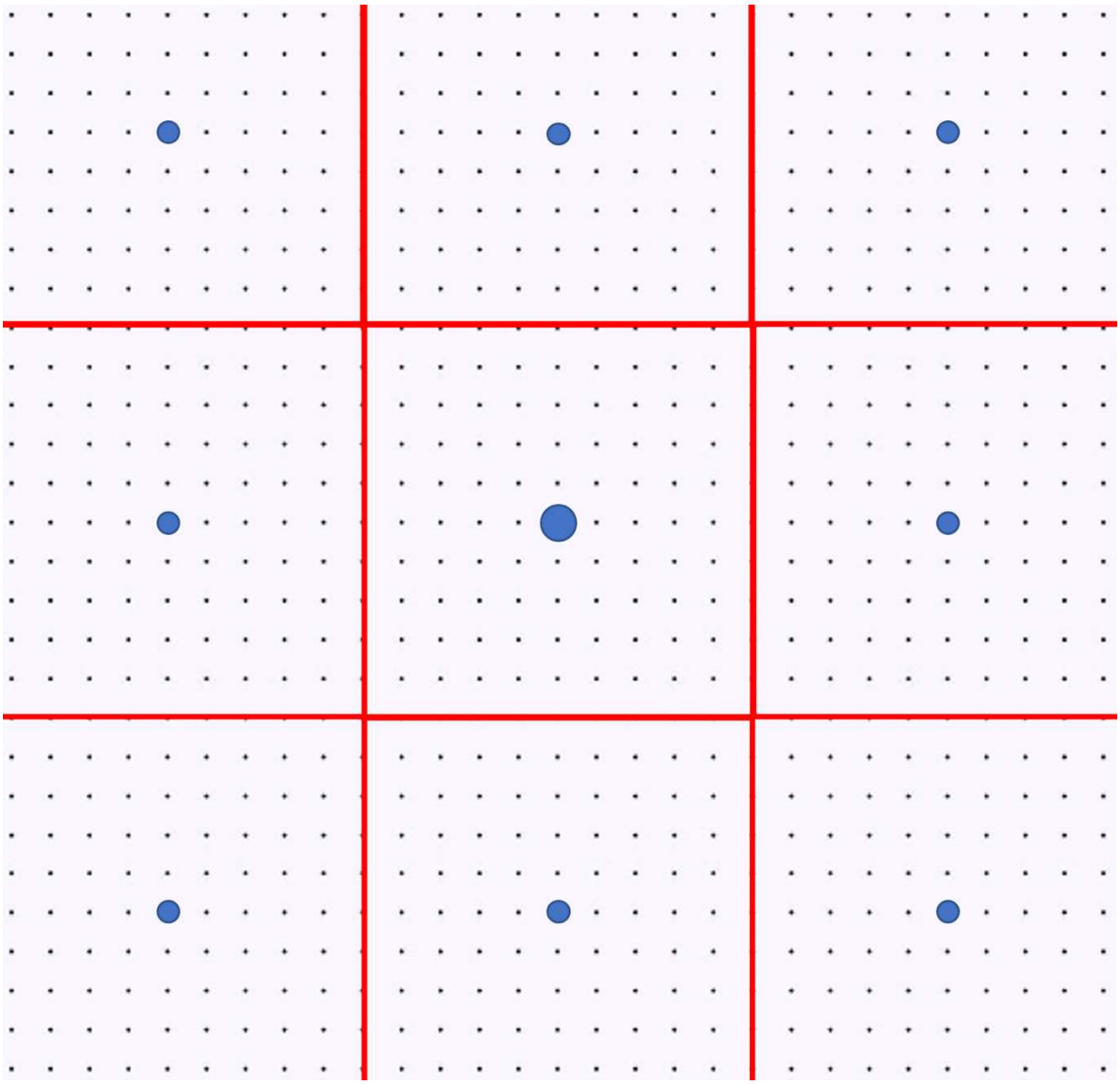

Fig. 13. Illustration of predictor assembly procedure. Tiny black dots represent 3-km HRRR gridpoints. Large blue circles indicate predictor gridpoints (where the forecast is issued). Intermediate blue circles indicate input gridpoints (information from these points, in addition to the predictor gridpoint, is used to make predictions). A predictor radius of only one input gridpoint is shown here, but for the real system, a predictor radius of 6 input gridpoints is used (unless otherwise specified). The red boxes indicate regions over which predictors are spatially averaged (or a spatial max/min is taken) for the spatial aggregation experiments.

NGP and ROCK regions unaffected. Further details on the masking impact are provided in chapter 4.

3.1.3. Training

The training period for the RF was three years (13 Jul 2018 – 31 Aug 2021), with 80 missing days (7% of the 1146-day period, leaving 1066 training days; days were missing due to gaps in the HRRR archive on the NOAA High Performance Storage System). For the time-lagged experiments, there are 128 missing days (leaving 1018 training days). While a three-year training period is relatively short for skillful predictions of a rare event like excessive rainfall, it again provides a testing ground for experiments on predictor assembly. The training includes a period of HRRRv3 (prior to 2 Dec 2020) and HRRRv4 (after 2 Dec 2020); the impact of this mismatch in HRRR version, as well as sensitivity to training length, are described elsewhere. Training was conducted in a manner consistent with Hill and Schumacher (2021), relating predictor variables to occurrence of target vector events (Table 2). The number of decision trees was set to 1000, the maximum number of predictors evaluated at each node was sqrt(N), and entropy was used as the splitting parameter. All models used 120 samples.

3.1.4. Verification approach

In order to evaluate the experimental results, we carry out daily verification of the various forecasts during the independent one-year period 1 May 2022 – 30 Apr 2023. The probabilistic forecasts are evaluated in several ways in a contingency table framework against indications of flash flooding within the Unified Flood Verification Dataset (UFVS; Erickson et al. 2019), which uses NWS storm reports, USGS stream gauge observations of flooding, and also QPE exceedances of flash flood guidance or the 5-year average recurrence interval, with a 40-km radius applied to match the neighborhood used in the definition of the ERO. We use a similar approach to analysis of the results as used by Schumacher et al. (2021), including the fractional coverage of observed events within probability contours, as well as Brier Skill Score (BSS) and

the area under the Relative Operating Characteristic (ROC) curve (AUC). A similar verification approach has been used by developers at the WPC (Erickson et al. 2021). The BSS calculation uses a daily varying climatological forecast constructed from the UFVS during the six-year period 1 May 2017 – 30 Apr 2023. In addition, we demonstrate forecast performance for representative cases. Statistical significance is determined using 100 bootstrap samples of the contingency table, with error bars showing the 2.5th to 97.5th percentile, as shown by Schumacher et al. (2021). Comparison is also made with the WPC ERO issued at 0900 UTC each day. In order to compare the RF predictions with the ERO on a level playing field, the RF predictions are discretized to the same probability contours as are included in the WPC ERO.

3.2. Sensitivity Experiments

In this section, we describe a number of sensitivity experiments intended to explore the use of CAM-based predictors in an RF model for excessive rainfall prediction. Table 4 provides experimental configurations for the sensitivity experiments described here.

3.2.1. Spatial aggregation experiments

Table 4 describes experiments intended to investigate the optimal path for spatial aggregation of predictor information from CAMs. The control experiment (CTRL) configuration is described in section 2. The MEAN and MEAN_MAX experiments explore the impact of spatial aggregation of predictors. Using just the sparse input gridpoints illustrated in Fig. 13 risks missing important information, for example convective storms with torrential rain, from CAM fields in between the input gridpoints. For the MEAN experiment, predictor values are computed as the spatial mean of each variable within the red boxes shown in Fig. 13. For the MEAN_MAX experiment, the spatial aggregation operation varies by input variable; for the environmental predictors (ENV in Table 3), a spatial mean is used, but for the storm attribute

predictors (STORM in Table 3), a spatial max (for APCP, W700, and UHMAX) or min (for UHMIN) is used. All predictors retain the same 3-h time step (with nine input times for the 24-h prediction period). Prior work (e.g., Sobash et al. 2020; Loken et al. 2022) has found beneficial impact from treating environmental and storm attribute fields differently for severe convective weather prediction with RFs; our spatial aggregation experiments are intended to determine if their results apply to a deterministic CAM configuration.

A second set of experiments explores the impact of the predictor radius, *n*. With fine-scale predictors from a deterministic CAM, it is possible that the signals for excessive rainfall that an RF would learn are more local than the 180-km radius used in the default configuration. To explore this question, we carry out experiments reducing *n* from 6 to 4 and 2 (PREDRAD4 and PREDRAD2 in Table 4).

Table 4: Predictor assembly experimental configurations.

Experiment name	**Spatial aggregation**	**Predictor time step**	**Predictor time window**
CTRL	Default (every 10 3km gridpoints)	3 h	None
MEAN	Spatial mean over 10 x 10 3km gridpoints for each input point	3 h	None
MEAN_MAX	As MEAN, but spatial mean only for ENV fields, and max/min for STORM fields	3 h	None
PREDRAD4	As MEAN_MAX, but using predictor radius of 4 (120 km)	3 h	None
PREDRAD2	As MEAN_MAX, but using a predictor radius of 2 (60 km)	3h	None
1H	Default (every 10 3km gridpoints)	1 h	None
1H_1H	Default (every 10 3km gridpoints)	1 h	1 h (temporal mean)
MEAN_MAX_TL	As MEAN_MAX, but averaging across 00, 06, and 12 UTC HRRR runs	3 h	None

3.2.2. Temporal aggregation experiments

We also carry out two temporal aggregation experiments. The 1H experiment uses a 1-h predictor time step, instead of the control run's 3-h time step, to investigate whether additional information can be provided by a finer temporal resolution for the predictors. The 1H_1H experiment uses a 1-h predictor time step, but also uses a 1-h temporal radius, meaning that values from both 1 h before and 1h after each predictor valid time are also used (i.e., three valid hours are averaged together for all predictors). This experiment explores whether additional valuable information would be provided by considering CAM fields across a time range rather than at an instantaneous time.

The final experiment, MEAN_MAX_TL, uses a "time-lagged" ensemble of HRRR initializations to explore whether it is beneficial to consider predictions from multiple HRRR simulations. Specifically, in this experiment we consider information from the 06 and 12 UTC initializations, in addition to the 00 UTC forecast as used in all the other RFs. We take the same spatial predictor aggregation approach of the MEAN_MAX experiment, but average the resulting predictor values across the three initialization times. Forecasters routinely assess run-to-run consistency in hourly HRRR runs in order to estimate the uncertainty of a meteorological situation (e.g., Benjamin et al. 2023); our hypothesis is that having run-to-run consistency quantified in the RF predictors would strengthen the signal for occurrence (or non-occurrence) of excessive rainfall.

3.3. Results

In this section, we summarize the results of our sensitivity experiments. We begin with the overall frequency of issuance of several probability thresholds in the RF systems compared to

the WPC ERO during the evaluation period. We then discuss results of the spatial and temporal aggregation experiments.

3.3.1. Outlook risk issuance frequency

Figure 14 shows maps of the frequency of issuance of the various WPC ERO risk categories during the 356 evaluation days (22 Apr 2022 – 1 May 2023). Marginal risk areas were issued by WPC a maximum of 25% of the time in southern Arizona (not shown), with broad maxima in the interior southeastern US and in Arizona / New Mexico (Fig. 14a). WPC ERO slight risk areas were issued most frequently in Arkansas, and along the lower Ohio River Valley area, with secondary maximum along the Mogollon Rim of Arizona (Fig. 14c). WPC ERO moderate risk areas were quite rare, with a maximum associated with landfalling atmospheric rivers in California (Fig. 14e). High risk outlooks are not shown; during this period of evaluation, there were two in California and two in Florida.

Comparing the issuance frequency of the CTRL RF system (Fig. 14b,d,f), we see that the CTRL system issues much lower probabilities than the WPC ERO overall. The maximum frequency of issuance of the marginal risk is ~30% along the east coast of Florida. The most notable differences in frequency of marginal risk issuance are seen in the interior southeastern US (Mississippi / Alabama), in the northeastern US, and in the Four Corners region associated with the North American Monsoon (Fig. 14b). Slight and moderate risk issuances are very rare in the CTRL system (Fig. 14d,f). Also evident in Fig. 14 are the relatively sharp changes in issuance frequency across region boundaries, particularly between the MDWST and the SE and NE regions (cf. Fig. 10).

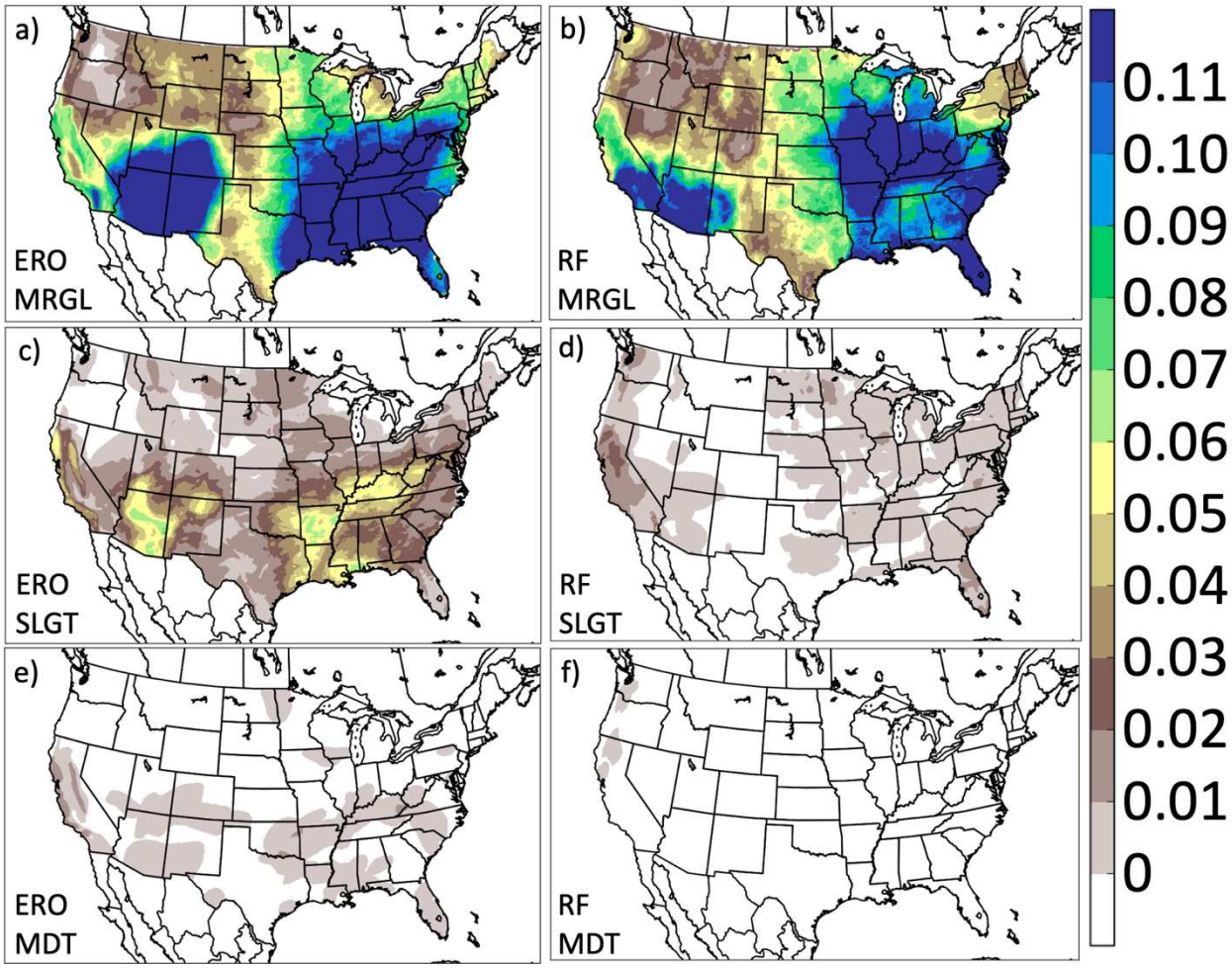

Fig. 14. Frequency of issuance of (a,b) marginal, (c,d) slight, and (e,f) moderate risk areas in (a,c,e) the WPC ERO and (b,d,f) the CTRL RF system during 22 Apr 2022 – 1 May 2023. The three risk categories correspond to 5%, 15%, and 40% chance of exceeding FFG within 25 miles of a point.

3.3.2. Spatial aggregation results

Figure 15a shows BSS for the spatial aggregation experiments over the entire period. On the CONUS scale (rightmost set of bars), we see no benefit from applying a spatial mean to the predictors (MEAN), but statistically significant improvement from applying a spatial max / min for storm attribute fields (MEAN_MAX). All the RF systems are substantially inferior to the WPC ERO in terms of BSS. The inferiority of the RF systems compared to WPC EROs is seen for all the regions, although the degree of inferiority varies. For the ROCK region, the spatial

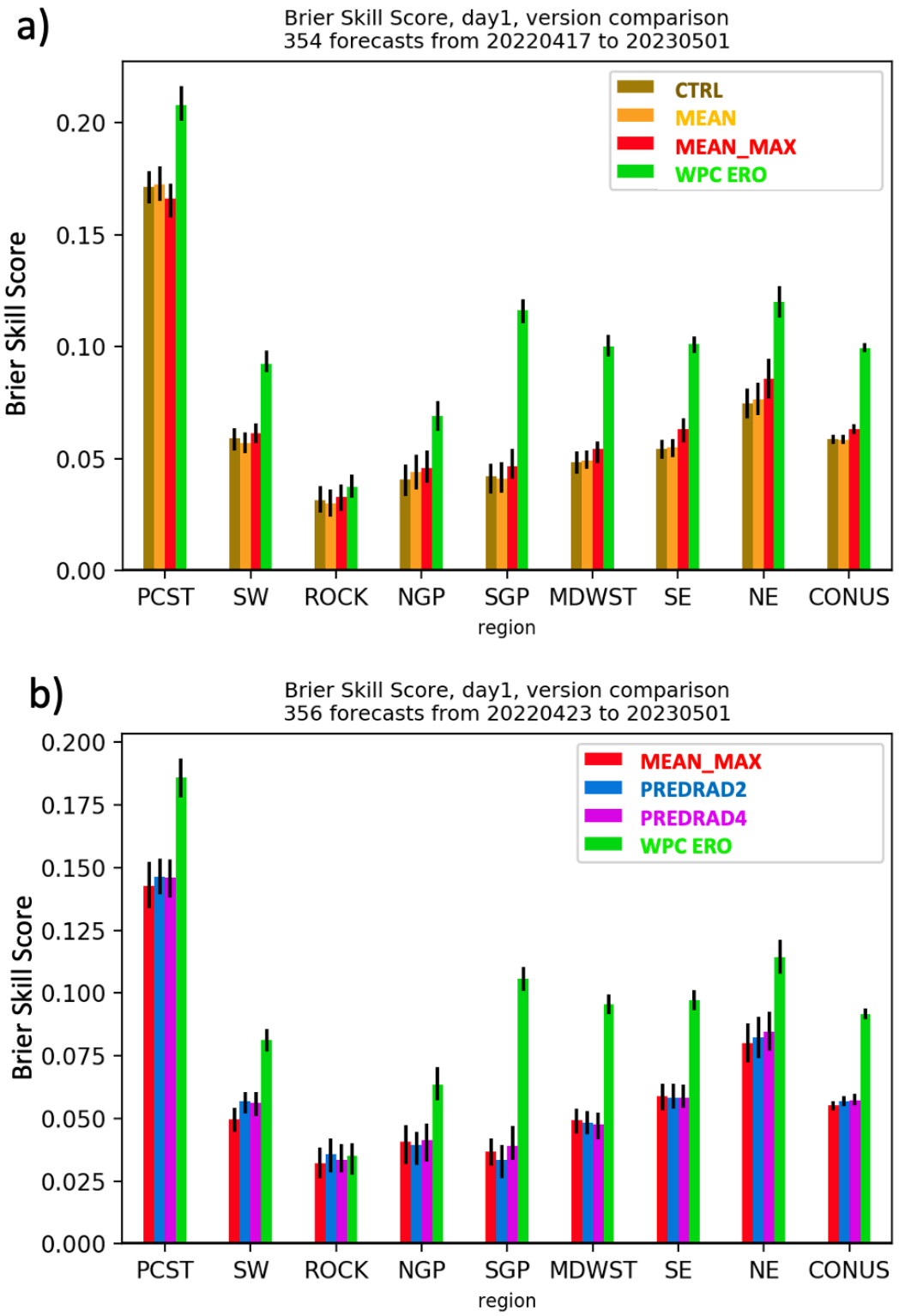

Fig. 15. Brier Skill Score for (a) spatial aggregation experiments and (b) predictor radius experiments. Shown are WPC ERO (green), CTRL (brown), MEAN (orange), MEAN_MAX (red), PREDRAD2 (blue), and PREDRAD4 (purple) by region during the 1-year evaluation period. 95% statistical significance is indicated by the narrow black bars.

spatial aggregation experiments are statistically indistinguishable from the WPC ERO in terms of BSS.

We also see that the benefit from the spatial aggregation varies by region. In particular, the RF forecasts in the eastern US show a clearer benefit from use of spatial max / min of storm attribute fields (MEAN_MAX). In the interior western US, special treatment of storm attribute fields (MEAN_MAX) does not lead to much of a skill improvement. In the PCST, the situation is unique, with marginally statistically significant degradation seen in the MEAN_MAX experiment; however, in this region, MEAN performs comparably to the control run.

For the PREDRAD experiments, we find smaller impacts than in the MEAN and MEAN_MAX experiments (Fig. 15b). On the CONUS scale (rightmost set of bars), both PREDRAD4 and PREDRAD2 do slightly outperform the control run (MEAN_MAX); however, the differences are not statistically significant. The greatest difference is seen in the SW region, where PREDRAD4 and PREDRAD2 both outperform MEAN_MAX; PREDRAD2 achieves the highest BSS in this region. This suggests there may be some benefit to restricting the predictor radius to a smaller distance in the SW.

In maps of the difference in BSS between the control and spatial aggregation experiments (Fig. 16), some of the same patterns emerge as are evident in Fig. 15. The BSS difference between MEAN and CTRL is small and somewhat mixed throughout the CONUS (Fig. 16a). In terms of the MEAN_MAX experiment (Fig. 16b), we see a clearer improvement over CTRL in the eastern US, in agreement with Fig. 15. In the rest of the CONUS, differences with CTRL are generally larger than seen for the MEAN experiment, although there are also regions of degradation in the UT/CO/WY area westward into NV/ID and northern CA. These areas have

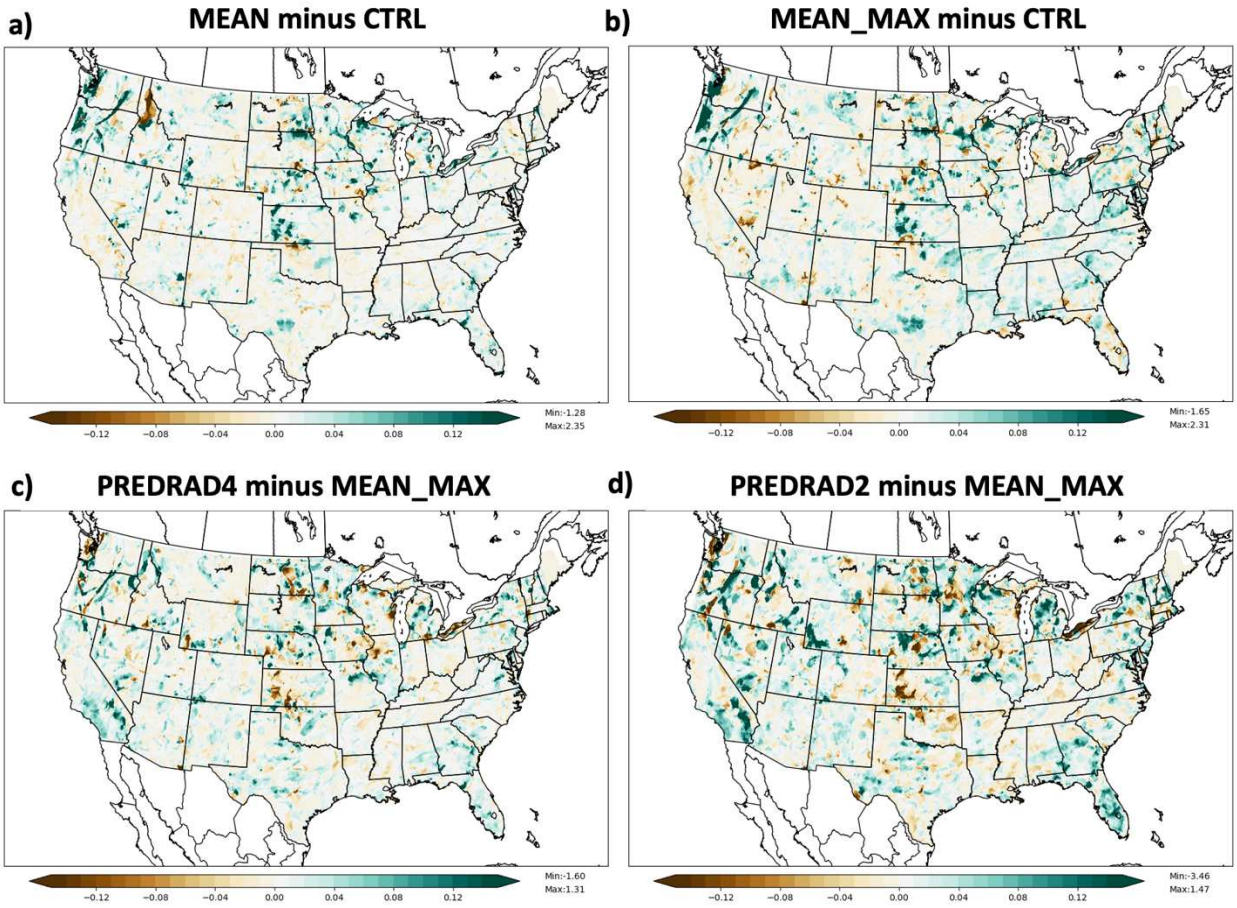

Fig. 16. Maps of BSS difference between (a) MEAN and CTRL, (b) MEAN_MAX and CTRL, (c) PREDRAD4 and MEAN_MAX, and (d) PREDRAD2 and MEAN_MAX during the 1-year evaluation period.

very few excessive rain events in the UFVS during the 2022-23 evaluation period (not shown), so BSS in these regions is likely unduly affected by a single forecast.

For the PREDRAD experiments (Fig. 16c,d), the BSS differences are generally less coherent. PREDRAD2 and PREDRAD4 appear to have systematically higher BSS in parts of the southwestern US, and also in Florida, but systematically lower BSS in some other regions. This agrees with the bulk statistics shown in Fig. 15b.

The area under the relative operating characteristic (ROC) curve, which measures forecast resolution, or the ability of the forecasts to distinguish excessive rainfall events from

nonevents, is shown in Fig. 17. Over the CONUS as a whole (Fig. 17, rightmost set of bars), the WPC ERO has superior resolution to any of the RF systems. However, we see that MEAN_MAX improves upon the resolution of the CTRL and MEAN experiments. Regionally, we see that the resolution improvement in MEAN_MAX holds true in every region but PCST. In fact, MEAN_MAX has comparable resolution to the WPC ERO in the MDWST, NGP, ROCK, and PCST regions.

For the PREDRAD experiments (Fig. 17b), there is very little impact of the predictor radius value upon the resulting forecast resolution. Consistent with Fig. 15, we see the largest impact of a reduced predictor radius in the SW region, with PREDRAD2 having the largest area under the ROC curve.

The reliability of the forecasts is assessed using a reliability diagram; for this evaluation, we retained a finer set of probability categories than included in the WPC ERO in order obtain a picture of reliability across a wide range of forecast probabilities (Fig. 18). Reliable forecasts will lie along the diagonal line, indicating forecasts of a given probability verify with the same probability threshold. As shown in Fig. 18, the RF systems all underpredict the probability of excessive rainfall. For example, 20% probability forecasts from CTRL verify ~35% of the time (brown curve in Fig. 18a). The spatial aggregation experiments do not show substantially improve the reliability of the RF forecasts (Fig. 18a); the results become noisy at higher probability thresholds due to relatively infrequent forecasts of extreme events.

For the PREDRAD experiments (Fig. 18b), there does not appear to be any additional improvement to reliability from restricting the predictor radius; MEAN_MAX, PREDRAD4, and

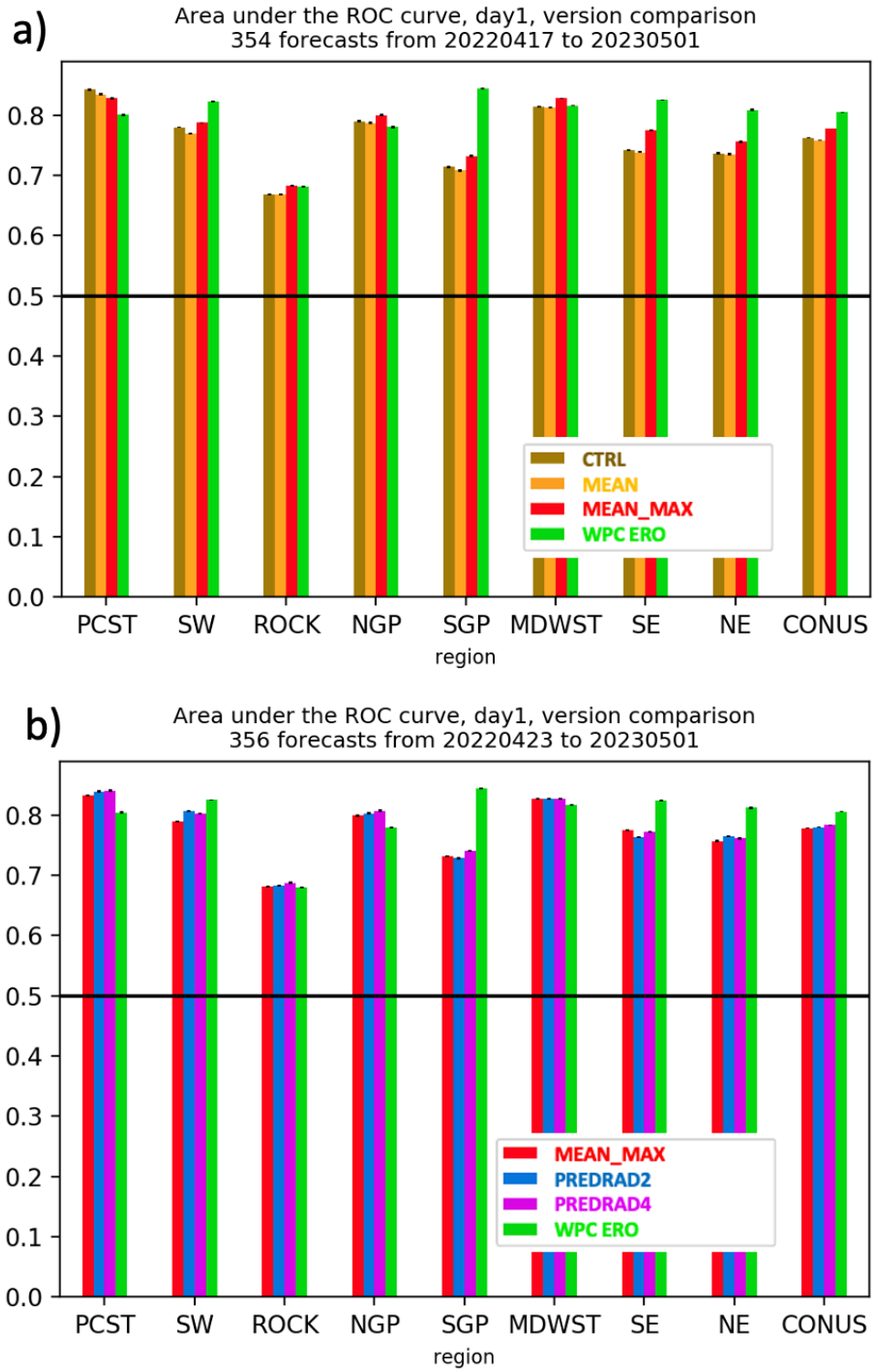

Fig. 17. Area under the relative operator characteristic (ROC) curve by region during the 1-year evaluation period, showing WPC ERO (green), and (a) CTRL (brown), MEAN (orange), and MEAN_MAX (red), and (b) MEAN_MAX (red), (c) PREDRAD2 (blue), and (d) PREDRAD4 (purple). 95% statistical significance is indicated by the narrow black bars.

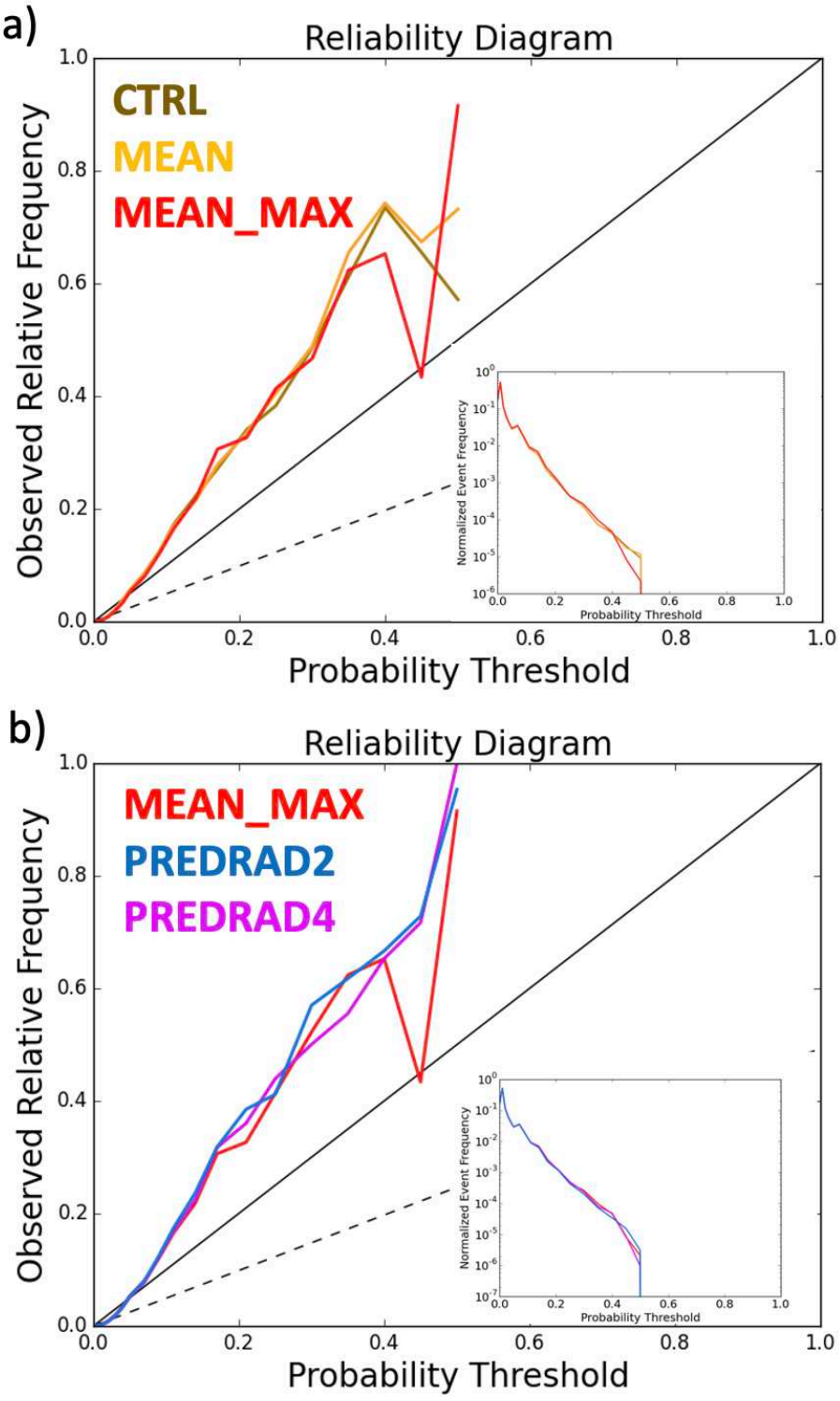

Fig. 18. Reliability diagram of RF forecasts, showing (a) CTRL (brown), MEAN (orange), and MEAN_MAX (red), and (b) MEAN_MAX (red), PREDRAD2 (blue), and PREDRAD4 (purple) experiments, during the 1-year evaluation period. Inset graphs describe the normalized frequencies of forecast at each probability threshold. Solid black diagonal lines and dashed black lines indicate perfect reliability and no skill, respectively.

PREDRAD2 are all similarly unreliable.

In order to determine if the impact of spatial aggregation is more prominent at certain times of year, we examine a monthly time series of BSS for the CONUS (Fig. 19). The BSSs for the three spatial aggregation experiments are nearly statistically indistinguishable during the majority of the year, but slightly higher BSS is seen from the MEAN_MAX experiment compared to CTRL and MEAN in several months, most notably during the warm season (April to September). This is likely due to the relatively smaller scale of warm season convective precipitation events relative to cool season synoptic precipitation events.

As shown in Fig. 19b, there does not appear to be a systematic difference in BSS between the PREDRAD experiments and the MEAN_MAX experiment, and what differences there are fail to reach statistical significance. This is consistent with the results of Fig. 15b, suggesting there is not a substantial benefit to restricting the predictor radius for the RF.

3.3.3. Temporal aggregation results

Turning our attention to the temporal aggregation experiments, Fig. 20a shows BSS for the temporal aggregation experiments. Once again, we see that the WPC ERO outperforms all of the RF experiments. We do not see a statistically significant benefit to using a 1-h time step rather than a 3-h time step on the CONUS scale (compare the CTRL and 1H experiments in the rightmost set of bars in Fig. 20a). Looking at different regions, there are not statistically significant differences between the experiments except in the PCST region, where a 1-h time step (1H), as well as using a 1-h temporal window (1H_1H) leads to statistically significant degradation compared to the control run (CTRL).

On the other hand, using predictors averaged across three separate HRRR initialization times (MEAN_MAX_TL), leads to a significant increase in BSS on the CONUS scale and in

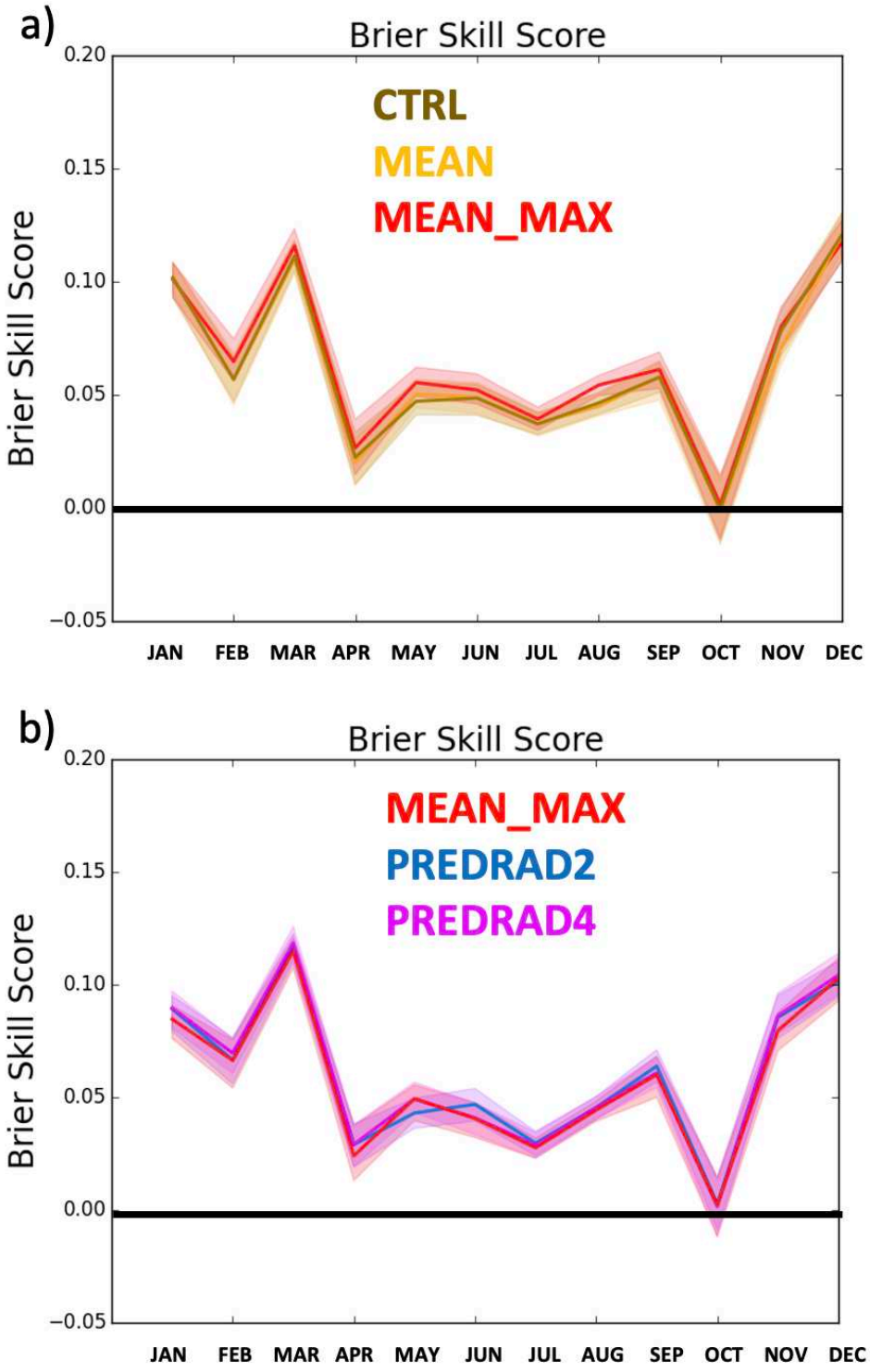

Fig. 19. Monthly time series of BSS among the spatial aggregation experiments during the 1-year evaluation period. Thick lines indicate the BSS, and shading indicates 95% confidence interval.

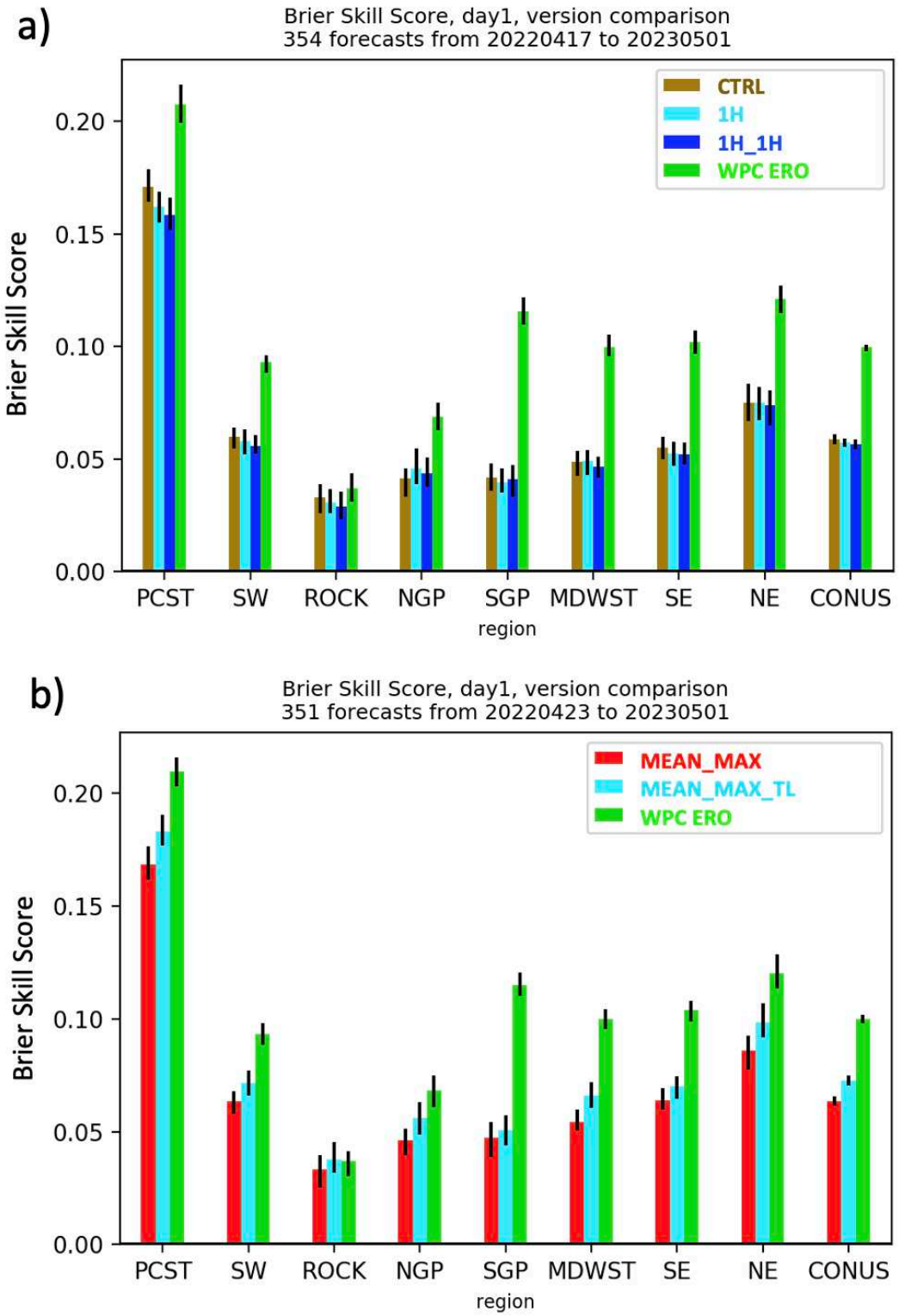

Fig. 20. Brier Skill Score for WPC ERO (green), (a) temporal aggregation experiments: CTRL (brown), 1H (cyan), and 1H_1H (dark blue), and (b) time-lagged input experiment: MEAN_MAX (red) and MEAN_MAX_TL (cyan), by region during the 1-year evaluation period. 95% statistical significance is indicated by the narrow black bars.

most regions (Fig. 20b). This means that the use of different simulations for the same time period provides the RF with a clearer indication of the true risk of excessive rainfall. This result agrees with our expectations, and implies that the use of ensemble information is one factor in why deterministic CAM-based RFs have not performed as well as global ensemble-based RFs.

Spatially varying impacts of the shorter time step are apparent, with larger BSS differences (although mixed in direction) seen in the northern US (Fig. 21a). Figure 21b shows a difference map between MEAN_MAX and MEAN_MAX_TL. The impacts of using time-lagged predictor information are more positive overall, with benefits seen all around the CONUS, in agreement with Fig. 21b.

In terms of the area under the ROC curve, there are relatively few differences between the CTRL, 1H, and 1H_1H experiments in any region (Fig. 22a). However, use of time-lagged predictor information leads to a modest increase in forecast resolution for the CONUS and for most regions (Fig. 22b).

As with the spatial aggregation experiments, all systems underpredict probabilities of excessive rainfall, but we see modestly improved reliability with the 1H and 1H_1H experiments (Fig. 23a). For the time-lagged input experiment (Fig. 23b), we see slightly improved reliability (less underprediction) at probabilities less than about 20%, but degraded reliability (increased underprediction) at higher probabilities. This suggests that, at least for higher probability excessive rainfall scenarios, the inclusion of time-lagged predictor decreases confidence and leads to a less reliable forecast, despite leading to improved BSS (Fig. 20a). This makes sense, as it is likely that there will be substantial run-to-run variability in the HRRR forecast in some scenarios, which could decrease confidence. It is important to note that the sample size is quite small at forecasted probabilities of greater than 20% (see inset to Fig. 23b).

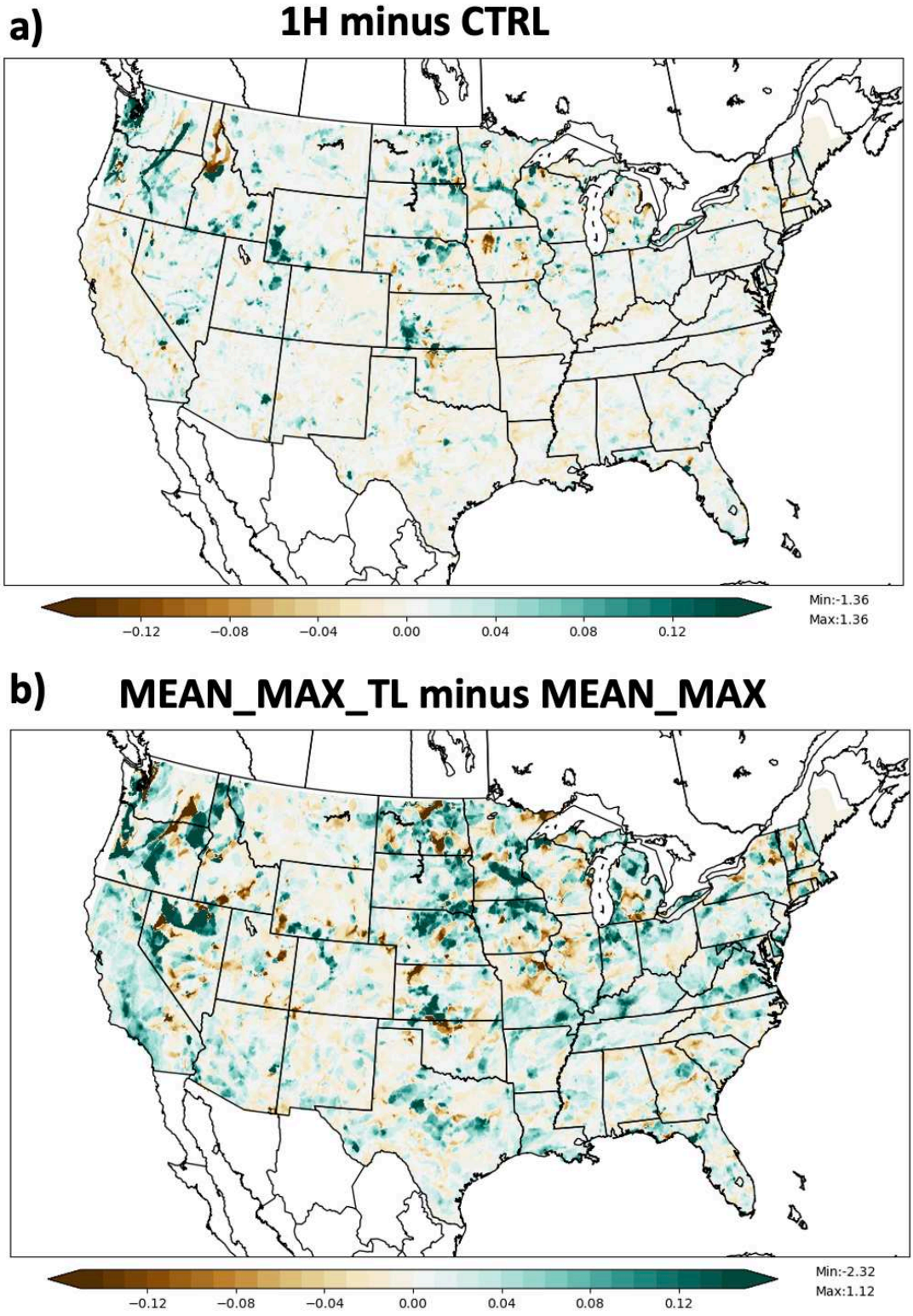

Fig. 21. Map of BSS difference between (a) 1H and CTRL, and (b) MEAN_MAX_TL and MEAN_MAX during the 1-year evaluation period.

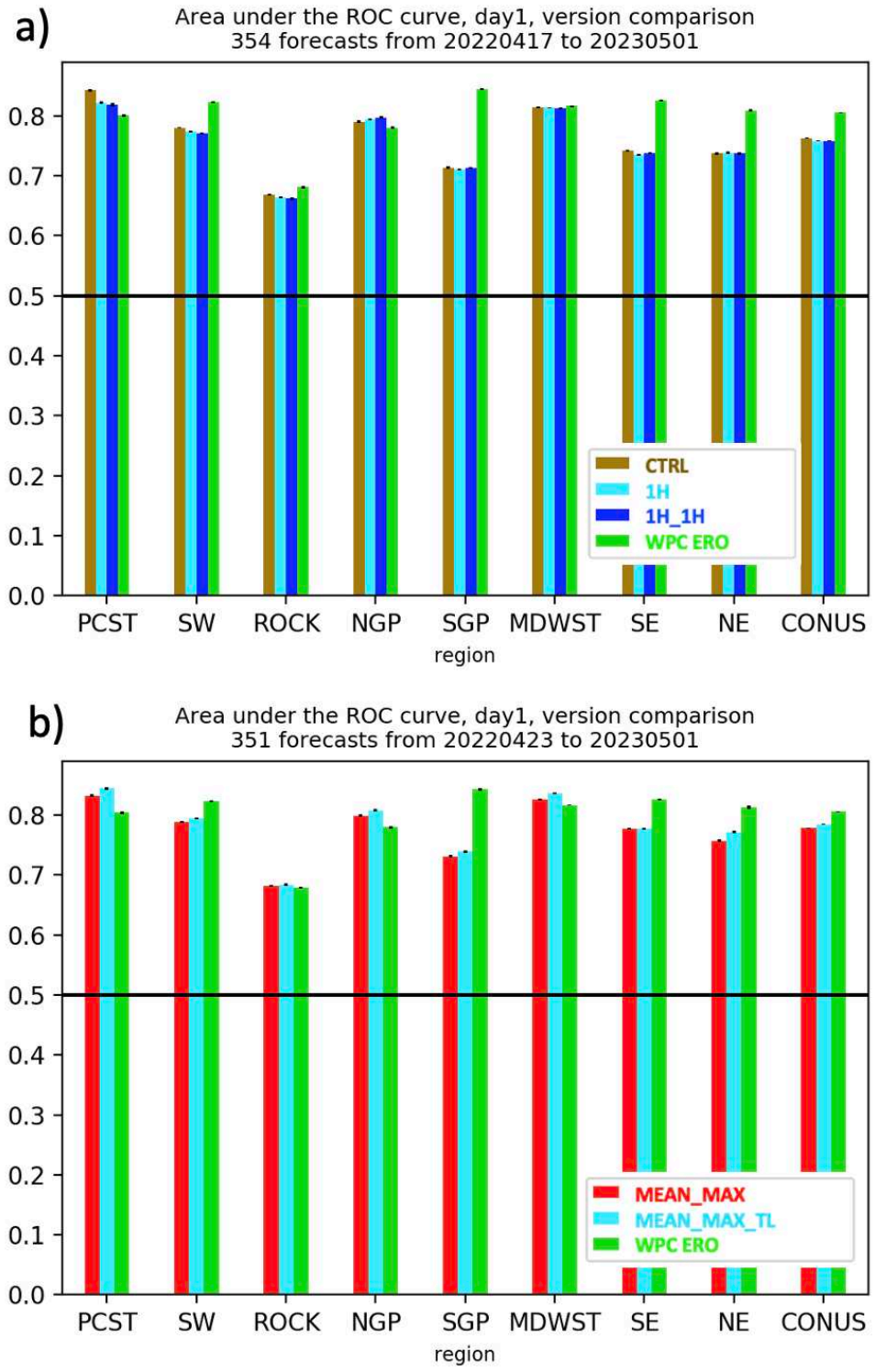

Fig. 22. Area under the relative operator characteristic (ROC) curve by region during the 1-year evaluation period, showing WPC ERO (green), (a) temporal aggregation experiments: CTRL (brown), 1H (cyan), and 1H_1H (dark blue), and (b) time lagging experiments: MEAN_MAX (red) and MEAN_MAX_TL (cyan). 95% statistical significance is indicated by the narrow black bars.

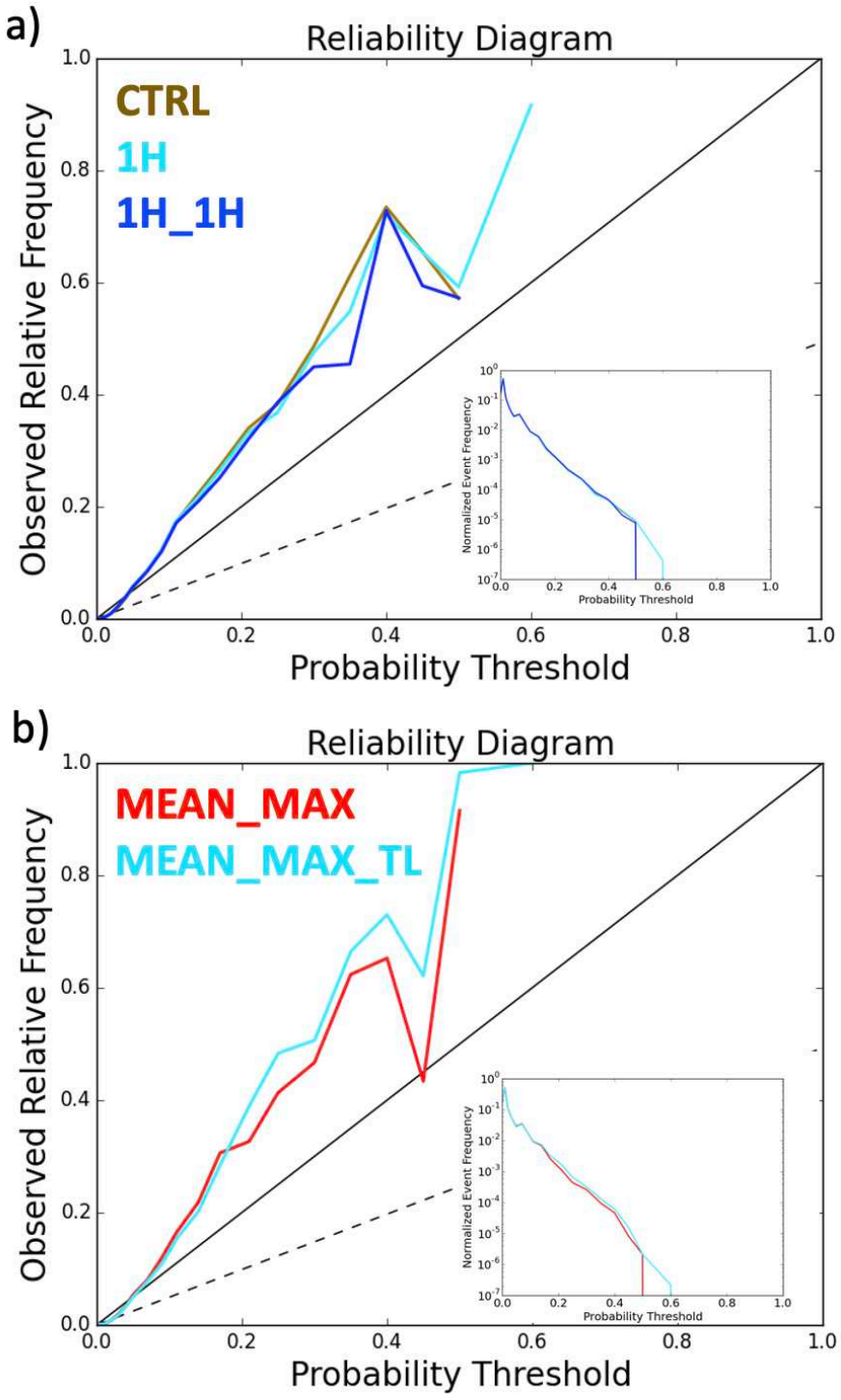

Fig. 23. Reliability diagram of RF forecasts, showing (a) temporal aggregation experiments: (brown) CTRL, (cyan) 1H, and (dark blue) 1H_1H, and (b) time-lagged input experiment: (red) MEAN_MAX and (cyan) MEAN_MAX_TL, during the 1-year evaluation period. Inset graphs describes the normalized frequencies of forecast at each probability threshold. Solid black diagonal lines and dashed black lines indicate perfect reliability and no skill, respectively.

In contrast to the results from the spatial aggregation experiments, the increase in BSS from the MEAN_MAX_TL configuration with respect to the MEAN_MAX configuration is seen fairly consistently across the annual cycle (Fig. 24). There are a few months in which the BSSs are statistically indistinguishable (for example, June – July and September – October). However, it appears that the use of time-lagged predictor information from the HRRR provides benefits both in the warm season and the cool season. Benefits do appear slightly greater during the cool season (October – April).

3.3.4. Feature contributions

In this section, in order to shed light on the mechanisms for the benefit stemming from spatial predictor aggregation and use of a shorter predictor time step, we examine feature contribution metrics calculated with the tree interpreter python package (Saabas 2016). Loken et

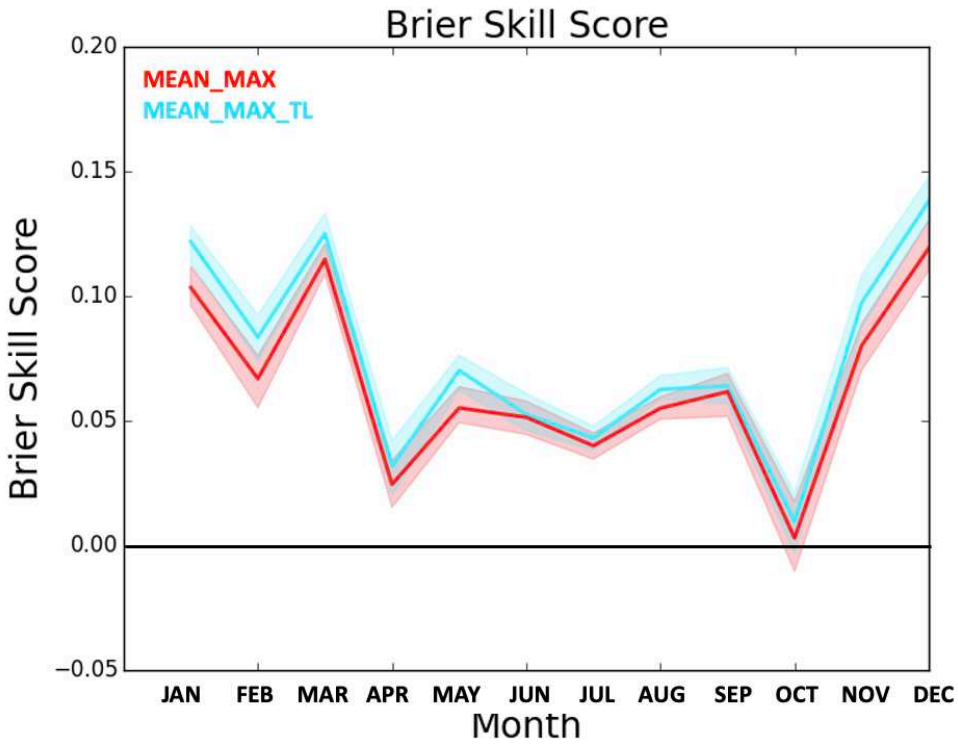

Fig. 24. Monthly time series of BSS between the MEAN_MAX and MEAN_MAX_TL experiments during the 1-year evaluation period. Thick lines indicate the BSS, and shading indicates 95% confidence interval.

al. (2022) provide a brief review of the various approaches available for interpreting ML / RF model results, including the strengths and weaknesses of the tree interpreter approach. Tree interpreter is a local interpretability method which measures the mean contribution of each predictor over all nodes in each tree within the RF. The resulting contributions reflect the influence of each predictor upon the resulting forecast, for a specific forecast case.

Figure 25 compares feature contributions for the CTRL, MEAN_MAX, and 1H experiments for the entire CONUS and the one-year evaluation period. In general, APCP and PWAT emerge as the two predictors with the largest positive contribution for both experiments (Fig. 25a). Relatively large positive contributions are also seen from Q2M. APCP has a much larger mean positive contribution in the MEAN_MAX experiment than in the CTRL experiment. This indicates that the use of spatial aggregation allows the RF to use information from the HRRR APCP field more effectively to improve forecasts (cf. Fig. 15a). Differences in mean positive contribution are also seen for other predictors, including UHMAX and UHMIN (with ~3 times the positive contribution in MEAN_MAX compared to CTRL). These fields, in addition to APCP and W700, are storm attribute fields (cf. Table 3); the RF is able to more effectively use information from APCP, UHMIN, and UHMAX when a spatial maximum or minimum is used. The W700 predictor does not contribute much more in MEAN_MAX than in CTRL, perhaps because this field is complicated by the presence of non-storm values related to synoptic systems or gravity waves.

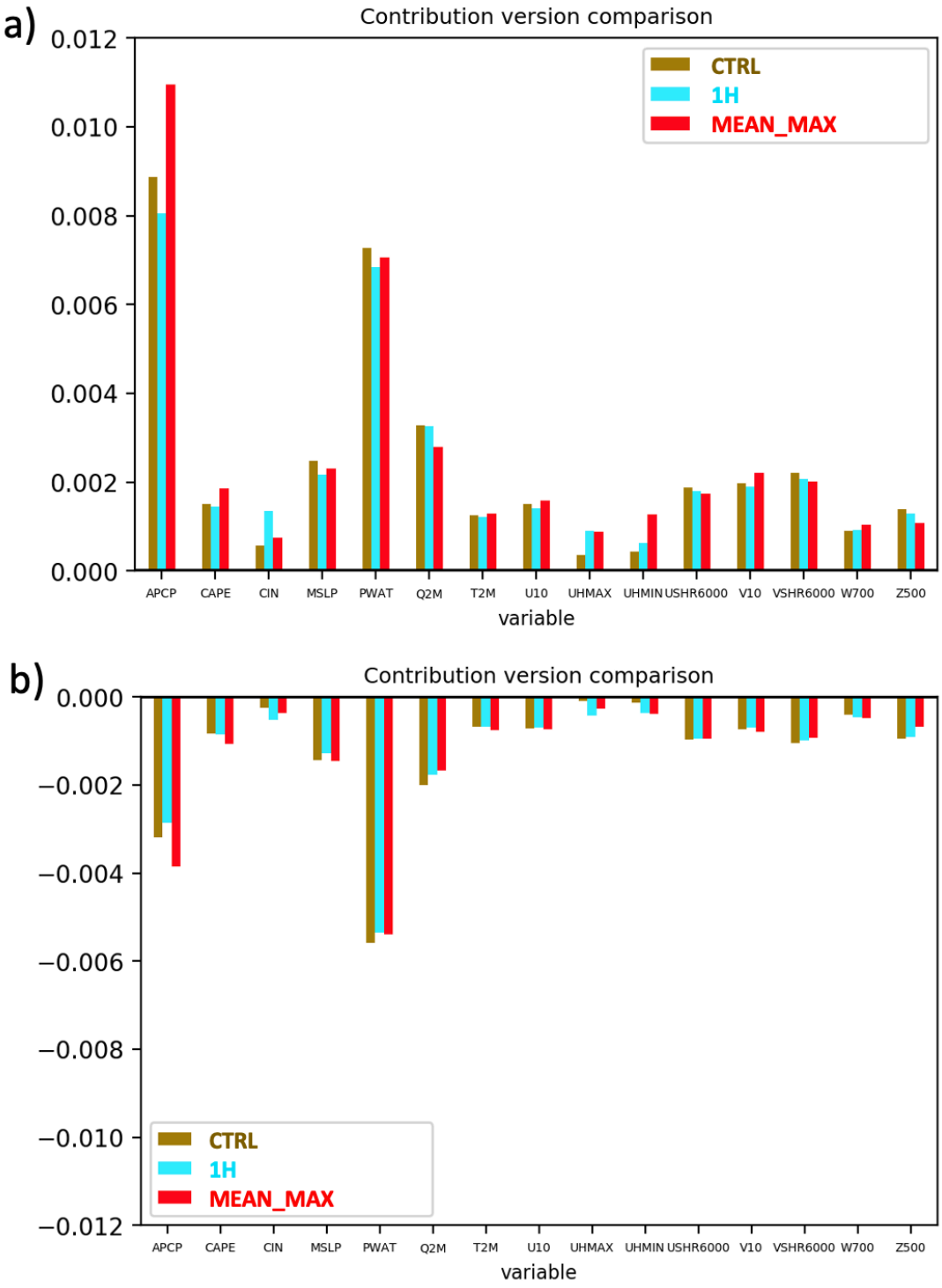

Fig. 25. Mean positive (a) and (b) predictor contributions for (brown) CTRL, (red) MEAN_MAX, and (cyan) 1H experiments, for every other day during the one-year evaluation period (178 days), as determined with the tree interpreter package. Positive and negative contributions are summed across all predictor time steps for each forecast, and then averaged across cases.

For the 1H experiment, differences in contributions from the control run are relatively more modest. Feature contributions are generally decreased with an hourly time step relative to the CTRL experiment, with the exception of a few fields, including CIN, UHMAX, and UHMIN. It makes sense that UHMAX and UHMIN would provide more information on an hourly time scale, since those fields are 1-h maximum and minimum fields from the HRRR.

To summarize the tree interpreter results, use of spatial aggregation allows the RF to more effectively use information from storm attribute fields like APCP, UHMAX, and UHMIN to increase the predicted probability of excessive rainfall where appropriate. Using a 1-h predictor time step does not dramatically impact RF forecast skill, although it does all the RF to glean more information from 1-h maximum or minimum fields from the HRRR (such as the UHMAX and UHMIN predictors).

3.3.5. Sample forecasts

In this section, we show some example forecasts comparing the different sensitivity experiments in order to illustrate how the differences in forecast skill, as well as the differences in predictor contributions, appear for a single forecast. Figure 26 shows predictions for the 12 UTC 26 Jul – 12 UTC 27 Jul 2022 period. This was a day of excessive rainfall in the central Appalachians of eastern Kentucky and southern West Virginia, and also an active day in the southwestern US. Widespread excessive rainfall occurred in Arizona, New Mexico, and Colorado, and also in a swath from southern Illinois to southern West Virginia. The control RF predicted excessive rainfall probabilities of 10-20% in parts of the southwestern US, and an east-west swath of 10-15% probabilities along the threat axis from southern Illinois to central Virginia (Fig. 26a). The MEAN_MAX experiment, in this case, predicted slightly higher probabilities of excessive rainfall in the southwestern US (compare the extent of 10-15%

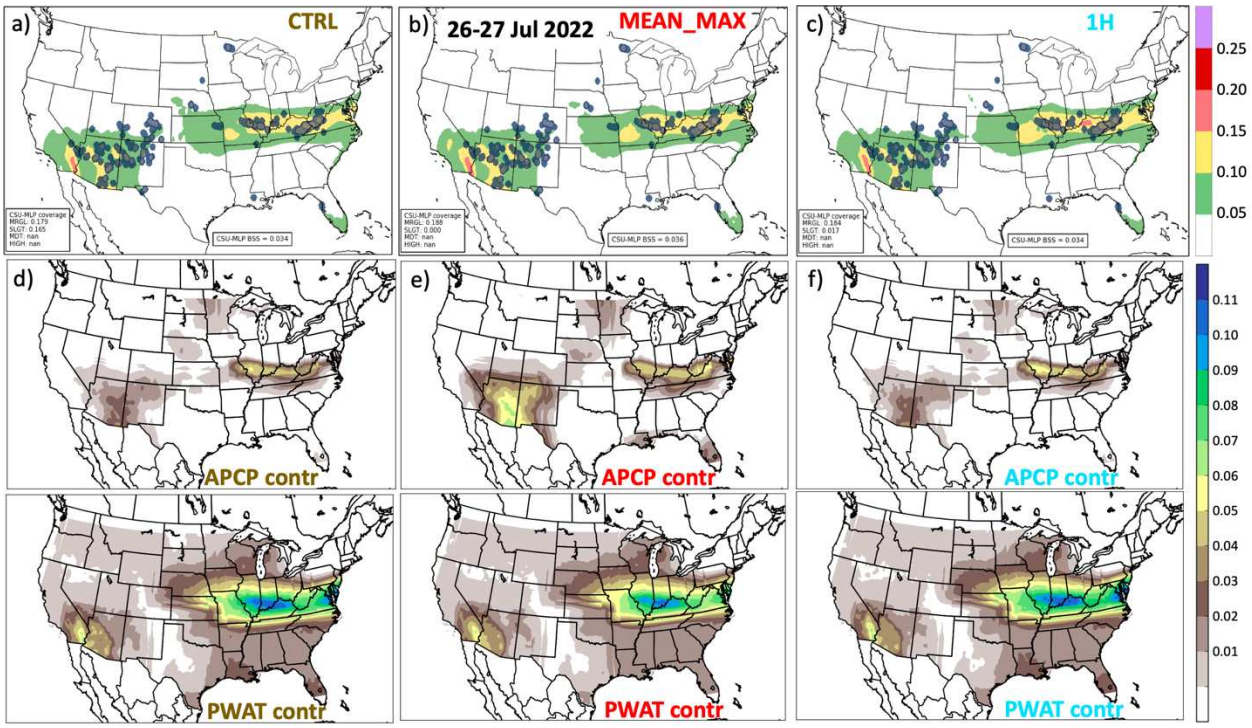

Fig. 26. Sample forecasts and a subset of predictor contributions for the 12 UTC 26 Jul – 12 UTC 27 Jul 2022 period, based on the 00 UTC 26 Jul 2022 HRRR run. Shown are (top row) the probabilistic forecasts of excessive rainfall (note the color scale which is different from the operational WPC outlooks), (middle row) contributions summed across all time steps from the APCP predictor, and (bottom row) contributions summed across all time steps from the PWAT predictor. Shown are (a,d,g) CTRL, (b,e,h) MEAN_MAX, and (c,f,i) 1H experiments.

probability in Arizona between Fig. 26a,b), but forecasted similar probabilities as the CTRL run in the Midwest (Fig. 26b). The 1H experiment forecast was similar to the CTRL forecast, with slightly higher probabilities in the east (Fig. 26c). Examining the feature contributions for this case, we see the strikingly larger positive contribution from the APCP predictor in the MEAN_MAX experiment compared to the CTRL experiment (Fig. 26d,e) in the SW, associated with small-scale convective heavy precipitation in this case (not shown). In contrast, in the Midwest, where the heavy precipitation was larger in scale, we see more similar contributions from the APCP predictor in the MEAN_MAX and CTRL experiments (Fig. 26d,e). APCP contributions in the 1H experiment appear similar to the CTRL experiment (Fig. 26f). For the

environmental PWAT predictor, broad but small positive contributions are seen in the SW region, with larger contributions in the Midwest, for all three experiments (Fig. 26g,h,i).

Another case, 12 UTC 5 Sep – 12 UTC 6 Sep 2022, is shown in Fig. 27. This case featured small-scale bands of excessive rainfall producing convection in the northeastern US (not shown). In this case, excessive rainfall was observed broadly across the eastern US, but with a swath of more widespread occurrences from north-central Pennsylvania eastward through New York, Connecticut, Rhode Island, and eastern Massachusetts. For this event, the CTRL experiment predicted a broad 20-25% risk of excessive rainfall from Pennsylvania to Massachusetts, with a small region of 25-30% risk in Pennsylvania and southern New York (Fig. 27a). In contrast, the MEAN_MAX experiment predicted much higher probabilities of excessive rainfall, greater than 25% across the entire area (Fig. 27b).

Examining the feature contributions, it is clear that the APCP predictor makes a much larger positive contribution in the MEAN_MAX experiment compared to the CTRL experiment (Fig. 27d,e). This larger contribution stems from the use of a spatial maximum for the APCP predictor field, which provides large benefits to the forecast when the HRRR prediction is for very small-scale precipitation features. The greater contribution from the MEAN_MAX experiment extends into the southeastern US, where scattered excessive rainfall was also observed (Fig. 27e).

Summarizing the results of these case studies, there is a clear indication that the benefit of using a spatial maximum for the APCP predictor field depends upon the spatial scale of the expected precipitation event. For large-scale events, the RF does not gain much information from using a spatial maximum of the APCP predictor field, since the field at sparse input gridpoints already provides a good representation of the expected precipitation. However, for

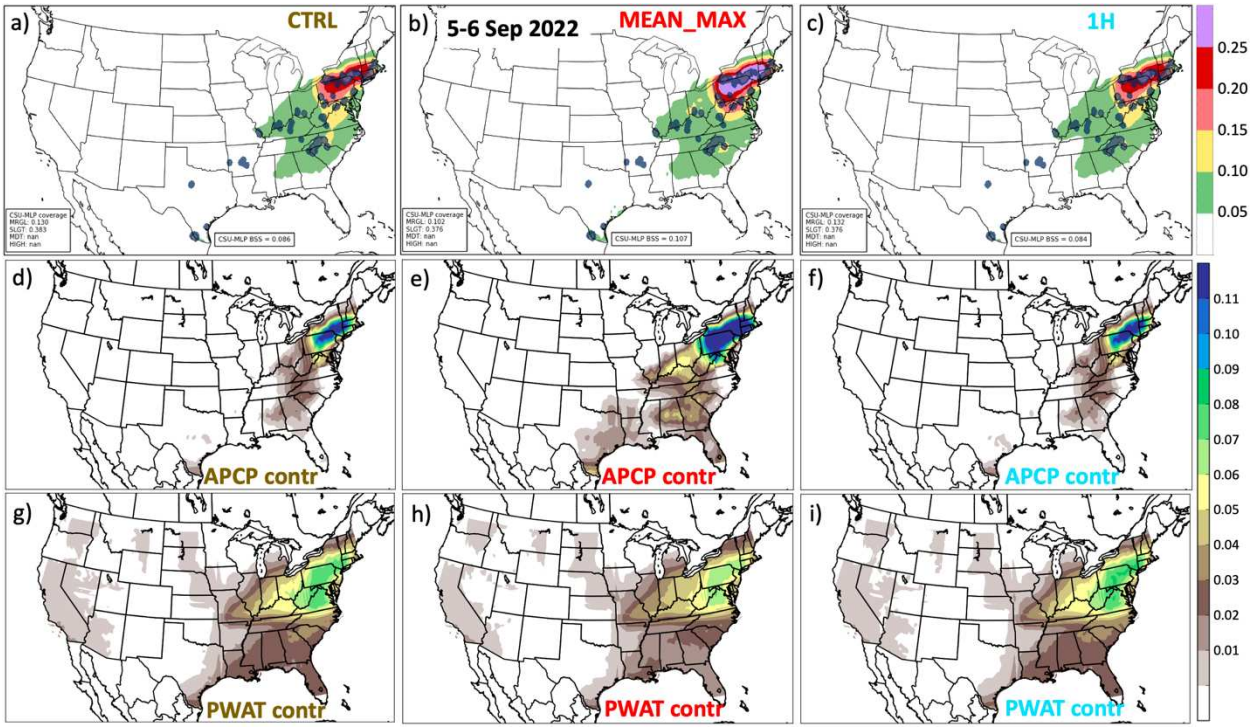

Fig. 27. As in Fig. 26, but for the 12 UTC 5 Sep – 12 UTC 6 Sep2022 period, based on the 00 UTC 5 Sep 2022 HRRR run.

small-scale precipitation events, there is a great benefit to taking a spatial maximum of the APCP predictor field, to ensure that the RF is able to have a good representation of the causative heavy precipitation event.

3.4. Discussion and Conclusions

Predictor assembly is an important question for all RFs, which have been gaining popularity as a tool for high-impact weather prediction. This study sheds light on the question of predictor assembly for a deterministic CAM, as a complement to the study of Loken et al. (2022) for predicting severe weather based on a CAM ensemble. Our study also sheds light on possible reasons for the inferior performance of RF-based prediction tools for excessive rainfall in the southwestern US.

To summarize our results, we find that use of spatial aggregation (use of a spatial mean, with even better performance stemming from use of a spatial maximum or minimum for storm attribute predictors) leads to a statistically significantly improvement to RF forecasts of excessive rainfall, with greater improvement seen for precipitation events which feature small-scale precipitation maxima. We also find that use of a 1-h time step rather than a 3-h time step, and use of a 1-h temporal window for predictor information, does not lead to any substantial forecast improvement.

Regarding the benefit of spatial aggregation, it is interesting that the benefit is seen most strongly for events with small-scale precipitation maxima. This suggests that the CTRL experiment, which uses only sparse input information from the predictors, does not have a coherent signal with which to predict small-scale excessive rainfall events, and it is important to aggregation predictor information over a spatial radius for these types of events. Benefit of the spatial aggregation approaches is seen in most regions of the country, with the exception of the PCST, indicating that some fraction of excessive rainfall events around the rest of the US are characterized by small spatial scales that can only be accounted for using these spatial aggregation approaches.

The lack of benefit stemming from a shorter time step (as shown in the 1H experiment) indicates that the signals for excessive rainfall are generally adequately captured with a 3-h time step. However, we do find slightly larger feature contributions from the 1-h maximum and 1-h minimum updraft helicity predictors from the HRRR (UHMAX and UHMIN) when using a 1-h predictor.

Another set of experiments exploring the radius of predictor information being used by the RF revealed minimal sensitivity to this parameter, although with slightly improved

predictions in the SW region when using a shorter predictor radius. Additionally, use of predictor information averaged across three HRRR initializations, instead of the single 0000 UTC HRRR simulation, led to statistically significant benefit in all regions of the US. The benefit stemming from use of time-lagged ensemble information suggests that the availability of ensemble information is an important factor in the difference in forecast skill between the global ensemble-based RFs for excessive rainfall and those based on deterministic CAMs.

During the Flash Flood and Intense Rainfall (FFaIR) experiment 2021, several real-time versions of RFs for excessive rainfall were formally evaluated by participants, including the GEFS-based system described by Schumacher et al. (2021), the NSSL-WRF based system described by Hill and Schumacher (2021), and an earlier version of the HRRR-based system described here. It was noted in the experiment report that the HRRR-based RF performed worst of all the RFs, particularly in the southwestern US, and it was hypothesized that the short training period (only ~2 years at the time, and not including an active monsoon season) was the cause of the low probabilities and generally poor performance. This study suggests that another factor in the performance of the HRRR-based RF in the southwestern US is the use of spatial predictor aggregation.

Our experiments focus on the operational HRRR model, a deterministic CAM, but future work should explore the use of CAM ensembles for excessive rainfall prediction, as has been done for severe weather prediction (e.g., Loken et al. 2022). The operational CAM ensemble is currently the High-Resolution Ensemble Forecast (HREF), but development is underway on a formal 3-km ensemble prediction system based on the Unified Forecast System (UFS).

Some features of the tree interpreter-based contributions suggest opportunities for RF improvement. It is evident in Figs. 26 and 27 that there are sharp gradients in the predictor

contributions near the region boundaries (cf. Fig. 10). Also evident in these figures are relatively high probabilities of excessive rainfall along the coast of southern California where no excessive rainfall was observed. Erroneous high probabilities along the coast of southern California were noted on many days during the 2023 FFaIR experiment; further examination reveals that the RF is recognizing meteorological patterns in this area that are associated with excessive rainfall when they occur elsewhere in the SW domain, but are very unlikely to lead to excessive rainfall in the coastal mountains of southern California. Taken together, these results suggest that the regional definitions for the RF could be improved. Alternatively, one could envision a system with no regional breakdown, in which the final forecast results from the sum of predictions from several different RFs designed to capture different types of excessive rainfall events.

Our results suggest that careful consideration of the spatial scales of excessive rainfall events and their environmental indicators can provide some guidance on how to best assemble predictors for an RF-based system. The spatial aggregation experiments shown here led to improvements in almost all regions of the US, but most strongly in the eastern US, while use of a shorter predictor radius led to the greatest improvements in the SW US. Excessive rainfall events in the SW are a known challenge for WPC, with this region experiencing the most frequent "missed" damaging flash flood events (Williamson et al. 2023). For this reason, our results are important for future RF systems to take into account, and may hold promise for developing improved early warning capabilities regarding potential flash flooding during the North American monsoon.

CHAPTER 4: PROGRESS ON RANDOM FORESTS FOR PREDICTION OF EXCESSIVE RAINFALL BASED ON AN OPERATIONAL CONVECTION-ALLOWING MODEL

Flash flooding is one of the leading weather-related causes of death and property damage in the United States (Ashley and Ashley 2008), with billions of dollars in damage and dozens of lives lost in the annual mean (e.g., Ashley and Ashley 2008; NWS 2017a). Population growth in flash flood prone regions (Downton et al. 2005; Pielke and Downton 2002), as well as projected increases in flash flooding associated with climate change (Prein et al. 2017), underscore the growing importance of timely and accurate flash flood forecasts. Characterized by short time scales and watershed-dependent hydrologic responses, flash floods require a fundamentally different forecasting paradigm than longer-term river flooding.

Flash flood prediction is a very challenging problem, with multiple sources of uncertainty. QPFs from numerical weather prediction (NWP) models are prone to errors related both to initial condition and model errors, particularly in the rapidly-evolving and sensitive environments of deep, moist convection. Even with perfect QPFs, the hydrologic response to a given amount of precipitation is highly spatially variable, and exhibits changes in time due to land surface processes. These factors suggest that flash flood prediction is an inherently probabilistic forecasting challenge. An additional complication is the fact that there is no universally accepted definition of a flash flood. A wide variety of different forecasting approaches have been developed to address flash flood threats both from an atmospheric perspective (i.e., quantitative precipitation forecasts QPF) and a hydrologic perspective (e.g., Sharif et al. 2006; Javier et al. 2007; Chen et al. 2013; Broxton et al. 2014). Heavy rainfall prediction development has included early ingredients-based techniques (e.g., Maddox et al. 1978, 1979; Doswell et al. 1996), analog techniques (e.g., Marty et al. 2012), increasingly

sophisticated data assimilation and NWP systems (e.g, Yussouf et al. 2016; Yussouf and Knopfmeier 2019), nowcasting systems (e.g., Sun et al. 2020; Radhakrishnan and Chandrasekar 2020), as well as statistical and machine learning approaches (Herman and Schumacher 2018a,b).

Operational forecasters at the NOAA Weather Prediction Center (WPC) have responsibility to issue Excessive Rainfall Outlooks (EROs) for 24-h periods from 12 UTC to 12 UTC each day, out to five days lead time (Burke et al. 2023). These outlooks (issued in some form for 45 years, but revised to include three risk categories in 2004; Erickson et al. 2021) are defined as the probability of exceeding Flash Flood Guidance (FFG) within 25 miles of a point, and are intended to reflect the expected probability of flash flooding. The category thresholds were recently revised to better reflect the true probabilities (Erickson et al. 2021). WPC forecasters have access to a wide variety of observational datasets for monitoring occurrence of heavy rainfall and flooding, as well as operational and experimental deterministic and ensemble NWP systems.

Machine learning (e.g., McGovern et al. 2019) has shown significant promise in application to various problems in the atmospheric sciences, and has recently been applied to the flash flood prediction problem. Among machine learning approaches, random forests (RFs) have been widely used for forecasting in the last few years, with strengths including their inherent probabilistic predictions for occurrence of a well-defined event, their ability to find signals in massive datasets, and their accounting for biases in inputs. Applications have included severe convective hazards (Hill and Schumacher 2021, Loken et al. 2020, Gagne et al. 2017), low visibility and cloud ceilings (Herman and Schumacher 2016a), aviation turbulence (Williams 2014) and occurrence of high winds along sensitive interstate highway sections (Brothers and

Hammer 2022). Herman and Schumacher (2018a,b), and Schumacher et al. (2021) describe a novel RF excessive rainfall prediction system based on the global ensemble forecast system (GEFS) reforecast dataset. The GEFS system has relatively coarse horizontal grid spacing, as well as limitations related to its convective parameterization and other aspects of its data assimilation and physics suite. While development of the RF excessive rainfall forecasting algorithm for days two and three was originally motivated by the lack of convective-allowing model (CAM) guidance for that time window, success with the GEFS-based RF raises the question of whether a similar approach could be adopted to utilize information in current state-of-the-art convection-allowing models (CAMs) within the day one time window. CAMs, in particular the operational High-Resolution Rapid Refresh (HRRR; Dowell et al. 2022, James et al. 2022), exhibit skill for predicting the timing and location of heavy rainfall (e.g., Szoke et al. 2015, 2018; Herman and Schumacher 2016b).

Hill and Schumacher (2021) describe work towards a RF excessive rainfall prediction system based on the NSSL-WRF model, using a seven-year training dataset. They find major dependence on the construction of the target vector in different regions of the CONUS, largely due to biases in quantitative precipitation estimate (QPE) datasets during their training period. In addition, they find improved forecasts with spatially aggregated predictors, possibly due to reduced convective-scale noise in the meteorological input data. One of the main motivations for use of the NSSL-WRF model was the long training period with a static model; however, it remains to be quantified how short of a training period can produce reasonably skillful forecasts. Additionally, it is unclear how detrimental would be a change in model configuration during the training period.

While NSSL-WRF has the strength of a long period of record, it has been static for many years with no active development, and has no data assimilation. The HRRR, in contrast, features storm-scale ensemble data assimilation and continued physics development until recently (Dowell et al. 2022). The HRRR has recently been successfully used in a neural network prediction system for occurrence of severe convective hazards (Sobash et al. 2020), and represents an interesting candidate for application of a RF system for predicting excessive rainfall. In this article, we build upon the work of Hill and Schumacher (2021) to explore excessive rainfall prediction based on RFs using deterministic CAMs. In particular, we want to answer the following research questions: Does the relatively short period of record for the HRRR seriously undermine the RF approach? How significant are HRRR version changes for interrupting the training dataset? Can we get any improvement from using more recent initialization times (since the HRRR runs more than once per day)? And finally, has any progress been made on improving HRRR-based RFs over three years of development and evaluation from the WPC Flash Flood and Intense Rainfall Experiment (FFaIR)?

The remainder of this article is organized as follows. Section 4.1 outlines the design of the HRRR CSU-MLP system. We then describe a number of sensitivity experiments in section 4.2, and results of realtime evaluation of forecasts in section 4.3. We provide a discussion and conclusions in section 4.4.

4.1. System design

The system employed here is described in some detail in chapter 3; here we review the RF configuration. We develop eight independent machine learning models, one for each of the eight CONUS sub-regions (shown in Fig. 10). In the following sub-sections, we describe the

various components needed for the final trained model: predictands, predictors, and model training.

4.1.1. Predictand assembly

It is critical to define a high-quality predictand (also known as labels or the target vector) for a good RF system. For flash flood prediction, this is a non-trivial problem. Since the product is intended for use at WPC, one option is the definition of the outlook: probability of precipitation exceeding FFG (Sweeney 1992). However, FFG is subject to large differences in methodology, and thus discontinuities, between different River Forecast Centers (RFCs; Clark et al. 2014). A simpler approach would be to use flash flood reports (FFRs) as the predictands, but FFRs are subject to substantial regional biases, as described in detail by Herman and Schumacher (2018c). Alternatively, one can follow the approach of Hill and Schumacher (2021), also described by Schumacher et al. (2021), to take advantage of QPE exceedances of average recurrence intervals (ARIs) to augment FFRs in defining excessive rainfall events. Table 5 describes the predictands constructed for the HRRR-based system described herein. "HRRR2021" indicates the initial version of the HRRR-based RF system, following the predictand assembly used for the "NSSL1" model evaluated during FFaIR 2020 (Trojniak et al.

Table 5: Regionally varying target vectors used for "HRRR2021", the preliminary version of the HRRR RF system, evaluated during FFaIR 2021, and "HRRR2022" and "HRRR2023", the subsequent version evaluated at FFaIR 2022 and 2023.

Region	**HRRR2021**	**HRRR2022-2023**
PCST	FFR+CCPA, 2-year	FFR+CCPA, 2-year
ROCK	FFR+CCPA, 2-year	FFR+CCPA, 2-year
SW	FFR+CCPA, 2-year	FFR+CCPA, 2-year
NGP	FFR+CCPA, 1-year	FFR+CCPA, 1-year
SGP	FFR+CCPA+ST4, 2-year	FFR+CCPA, 1-year
MDWST	FFR+CCPA, 1-year	FFR+CCPA, 1-year
SE	FFR+CCPA+ST4, 1-year	FFR+CCPA+ST4, 1-year
NE	FFR+CCPA+ST4, 2-year	FFR+CCPA, 2-year

2020), while "HRRR2022" indicates the optimized model evaluated at FFaIR 2022. Since the impact of regionally varying target vectors was previously assessed by Hill and Schumacher (2021), we do not report on controlled tests examining these settings; however, the configuration for HRRR2022 - HRRR2023 follows the best-performing regional definitions found by Hill and Schumacher (2021).

A more straightforward approach to defining the target vector used for the flash flood RF system is enabled by the availability of a Unified Flood Verification System (UFVS, Erickson et al. 2019), developed by the WPC. The UFVS consists of a combination of flash flood observations and proxies from a number of data sources, including LSRs, USGS stream gauges, and Stage IV exceedances of FFG and ARIs. The UFVS is unlikely to miss any potential flash flood events, and thus could serve as a replacement for the somewhat arbitrary and subjective selection of regional target vectors based on Stage IV and CCPA exceedances of ARIs. It also has the advantage of including both static thresholds (from the ARIs) and information about antecedent and hydrologic conditions (from FFG).

4.1.2. Predictor assembly

Building off Hill and Schumacher (2021), the HRRR-based system uses similar output variables from the HRRR as are used for the NSSL-WRF. We used the same variables (see their Table 1), with the exception of a few variables defined differently (Table 3). In particular, accumulated precipitation in the HRRR is output at 1-h intervals, so the HRRR-based model is trained on run total accumulated precipitation; as noted in section 3.1.2, this is not an ideal configuration, and sensitivity tests are underway to quantify the impact of this choice. The use of updraft helicity is expanded from Hill and Schumacher (2021) through use of 1-h max and 1-h min values for 2-5 km updraft helicity (UH). 2-5 km UH has been previously used in RFs for

precipitation prediction due to its association with sustained rotating storms (Nielsen and Schumacher 2018, 2020: Smith et al. 2023). We use 0-6 km wind shear values, rather than the mean 0-6 km wind components as in Hill and Schumacher (2021). Finally, we use 700-hPa vertical velocity instead of 0-3 km average vertical velocity. In addition to these meteorological predictors, we use the same static inputs as Herman and Schumacher (2018a) and Hill and Schumacher (2021), which are related to the climatological likelihood of excessive rainfall.

As described by Hill and Schumacher (2021), the fine scale model predictors (3 km in the case of the HRRR) are collected in space at 3-h intervals over the 24-h period from 12 UTC to 12 UTC, based on 12- to 36-h forecasts from the 00 UTC HRRR simulation, corresponding to the periods associated with WPC EROs. For HRRR2021 (versions described in Table 6), predictors are spaced by 30 km (10 grid points), and up to 180 km from the forecast point, corresponding to a radius (n) of 6 nearby HRRR grid points. The number of predictors per training label is then $N = pt(2n + 1)^2$, where t is the number of forecast times ($t = 9$ here), amounting to 1521 predictors per variable p, and N = 22815 predictors per training example. For the HRRR2023 version, we use a 1-h time interval instead of a 3-h interval based on sensitivity experiments in chapter 3; in this case t = 25, and N = 63375.

Previous studies (e.g., Loken et al. 2022; Hill et al. 2021) have found benefit to spatial averaging of predictors within an RF system. Thus, we use different aggregation approaches for the different versions of the HRRR-based RF, shown in Table 6. The HRRR2022 version uses the OPT_AVG formulation of spatial averaging described by Hill and Schumacher (2021). The HRRR2023 version includes every 3 km HRRR gridpoint within a 60 km radius to calculate the predictor values at each prediction point, using a spatial mean for environmental fields and a spatial maximum or minimum for storm attribute fields, based on the experiments in chapter 3.

Table 6: Versions of the HRRR RF system demonstrated at FFaIR during 2021-23, with their configurations.

Version	**HRRR2021**	**HRRR2022**	**HRRR2023**
Training period	13 Jul 2018 – 31 Jul 2020 (HRRRv3)	13 Jul 2018 – 31 Oct 2021 (HRRRv3-HRRRv4)	1 Apr 2020 – 31 Mar 2023 (HRRRv4)
Training period length	750 days	1207 days	1095 days
Masking	Offshore points	Offshore and non-CONUS points	Offshore and non-CONUS points
Predictor assembly	HRRR2021 approach (see text)	OPT_AVG approach of Hill and Schumacher (2021)	MEAN_MAX approach of chapter 3
Predictor time step	3 h	3 h	1 h
Target vectors	HRRR2021 (see Table 5)	HRRR2022 (see Table 5)	HRRR2022 (see Table 5)

For the HRRR-based system, we gave some additional consideration to the masking out of non-land areas in each of the regions shown in Fig. 10. Since our target vectors are based partially on flash flood reports (Table 5), it is important to exclude from the training dataset any model prediction for a non-land point, where flash floods can never be observed. This eliminates mis-training, where the RF system learns that a certain meteorological pattern is less likely to be associated with flash flooding simply because it occurred over an offshore or Great Lake area. Table 7 describes the degree of masking for each region; in our initial HRRR2021 configuration, the NE region had 39% of its original points removed (although this one region did have some gridpoints masked out in the original configuration). Other regions had between 3% (SGP) and 20% (SE) of their points removed, while the continental regions ROCK and NGP regions were unchanged. Some additional masking of Canadian and Mexican land areas, as well as a small portion of the Pacific Ocean for the PCST region, was carried out for the HRRR2022 configuration (see rightmost column in Table 7).

Table 7: Number of spatial gridpoints included in each region in the original configuration, HRRR2021 version, and HRRR2022 version. Shown in parentheses is the fractional size of the resulting region, after masking out offshore and non-CONUS points, compared with the original configuration. Versions / regions where a change was made are highlighted in bold. Fig. 28 shows results comparing the original configuration with the HRRR2021 version.

Region	**Original Configuration**	**HRRR2021**	**HRRR2022-2023**
PCST	288 or 284	**235 (82%)**	**232 (81%)**
ROCK	704	704 (100%)	704 (100%)
SW	512	**481 (94%)**	**459 (90%)**
NGP	484	484 (100%)	484 (100%)
SGP	528	**514 (97%)**	**477 (90%)**
MDWST	630	**562 (89%)**	**540 (86%)**
NE	396 or 276	**242 (61%)**	242 (61%)
SE	576	**459 (80%)**	459 (80%)

Figure 28 shows the impact of masking upon forecast verification, for a relatively short verification period from 1 Aug to 15 Oct 2020. Note that the number of events occurring during this period is highly spatially variable. However, it is evident that some of the regions with masking (e.g., PCST and NE) exhibit significantly improved forecasts in terms of Brier Skill Score (BSS). Examining Relative Operating Characteristic (ROC) curves (not shown) reveals that these improvements are due mostly to large gains in probability of detection with the additional masking. This indicates that excluding from the training misleading gridpoints where observed events can never occur tends to increase the chance that the RF can predict an event.

4.1.3. Training

The training periods for the various configurations are described in Table 7. The three training periods included 54, 85, and 77 days, respectively, in which the 00 UTC HRRR run was not available; these days are excluded from the training. Training was conducted in a manner consistent with Hill and Schumacher (2021), relating predictor variables to occurrence of target vector events (Table 5). The number of decision trees was set to 1000, the maximum number of

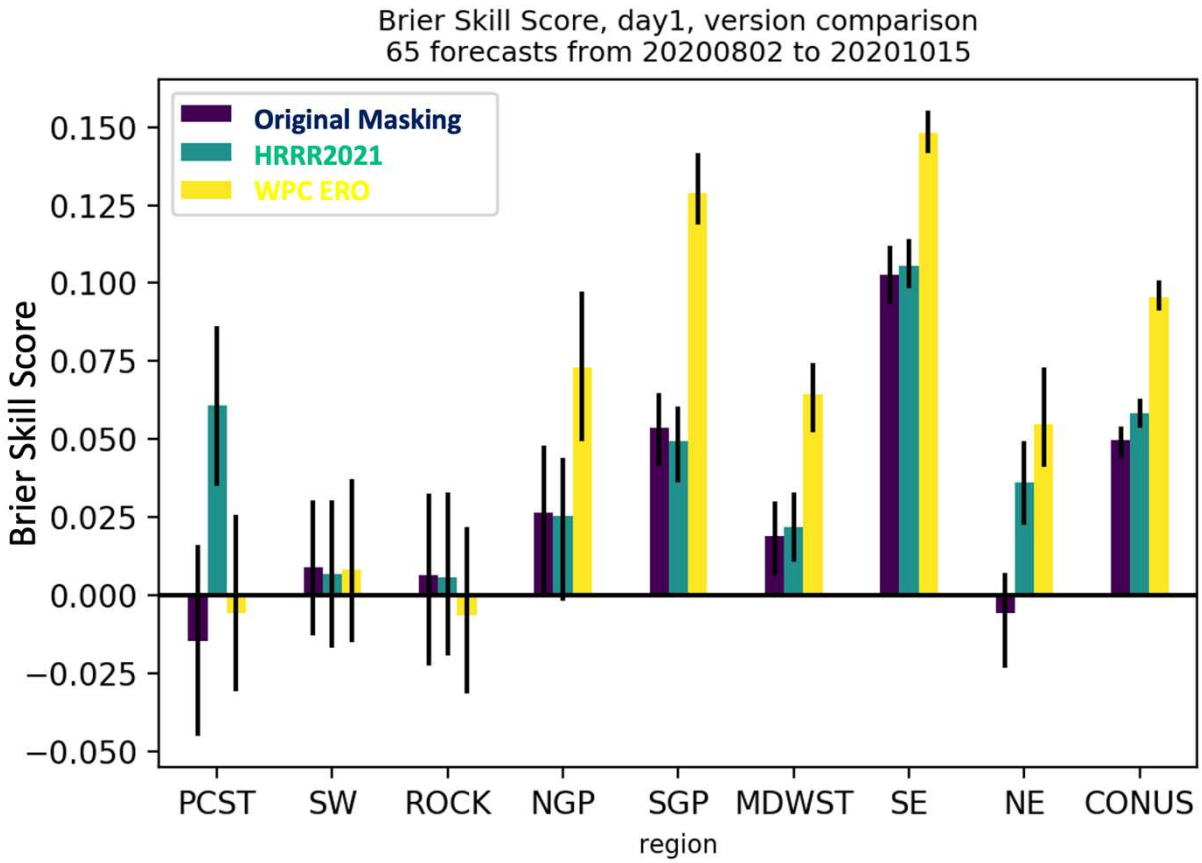

Fig. 28. Regional Brier Skill Score of HRRR-based RFs, showing (purple) the original configuration without offshore areas masked out from the training data, (green) HRRR2021 version which includes masking out offshore areas, and (yellow) WPC ERO, during 2 Aug – 15 Oct 2020.

predictors evaluated at each node was sqrt(N), and entropy was used as the splitting parameter. The HRRR2021 version used the same number of samples across all regions: 120, while the HRRR2022-23 versions used a regionally varying number following the testing of Hill and Schumacher (2021; see Table 8).

The initial version of the HRRR-based system (HRRR2021), with its 2-year training dataset, was noted to provide poor forecasts over the SW region and some adjoining regions (Trojniak and Correia 2021); it was hypothesized that this was due to the relative dearth of monsoon-related flash flood activity in the North American monsoon (NAM) region during the 2018 – 2020 training period. These results, combined with the relatively active 2021 NAM

Table 8: RF regional model configuration: minimum number of samples required to split an internal node.

Region	HRRR2021	HRRR2022-2023
PCST	120	16
ROCK	120	30
SW	120	30
NGP	120	120
SGP	120	120
MDWST	120	120
SE	120	120
NE	120	120

season, motivated the development of the HRRR2022 version, with an extension of the training period through Oct 2021.

The latest version of the HRRR-based system (HRRR2023) used a three-year training period based exclusively on HRRRv4; while HRRRv4 was implemented operationally on 2 Dec 2020, it was run experimentally at NOAA GSL for more than six months prior to the operational implementation. Access to these experimental HRRR data allows the training period to extend back to early 2020.

4.1.4. Forecast verification

As described in chapter 3, we verify the probabilistic RF forecasts against indications of flooding from the UFVS. We evaluate forecasts of independent sets of days (shown in Table 9a,b) in terms of Brier Skill Score (BSS), which uses a daily varying climatological baseline forecast constructed from the UFVS during the six-year period 1 May 2017 – 30 Apr 2023. Statistical significance is determined using 100 bootstrap samples of the contingency table, with error bars showing the 2.5th to 97.5th percentile, as shown by Schumacher et al. (2021). Comparison is also made with the WPC ERO issued at 09 UTC each day. In order to compare RF predictions with the ERO on a level playing field, the RF predictions are discretized to the

same probability contours as are included in the WPC ERO. Forecast resolution is also evaluated in terms of the area under the Relative Operating Characteristic (ROC) curve (AUC).

4.2. Sensitivity experiments

In this section, we describe a number of sensitivity experiments intended to explore the potential utility of a CAM-based RF model for excessive rainfall prediction, and also to determine the reasons for the inferior performance of deterministic CAM-based systems as compared with global ensemble-based system. Table 9 shows the experiments discussed in this section.

4.2.1. Impact of operational model upgrades

A long-standing question in the ML community has been related to the impact of changes (of varying significance) in the configuration of operational models used for training ML systems. Operational upgrades of numerical weather prediction models are quite frequent, potentially interrupting training periods and undermining the capability of the ML to learn bias

Table 9a: Experimental configurations, part 1. All experiments use the HRRR2021 target vector configuration (see Table 1).

Configuration	**Unmasked**	**HRRRv3 / HRRRv4 mismatch**	**Extended training period**	**HRRR+FFG**
Training period	13 Jul 2018 – 31 Jul 2020	13 Jul 2018 – 31 Jul 2020	13 Jul 2018 – 31 Aug 2021	13 Jul 2018 – 31 Aug 2021
HRRR version for training	V3	V3	V3-V4	V3-V4
HRRR realtime version	V3	V4	V4	V4
Masking	Original	2021	2021	2021
HRRR init for training	00Z	00Z	00Z	00Z
HRRR init for forecast	00Z	00Z	00Z	00Z
Evaluation period	2 Aug – 15 Oct 2020	2 Aug – 3 Dec 2020	17 Apr 2022 – 1 May 2023	5 Jul 2022 – 31 Aug 2023

Table 9b: Experimental configurations, part 2.

Configuration	**Realtime 06Z, HRRR2021**	**Realtime 12Z, HRRR2021**	**Realtime 06Z, HRRR2022**	**Realtime 12Z, HRRR2022**
Training period	13 Jul 2018 – 31 Jul 2020	13 Jul 2018 – 31 Jul 2020	13 Jul 2018 – 31 Oct 2021	13 Jul 2018 – 31 Oct 2021
HRRR version for training	V3	V3	V3-V4	V3-V4
HRRR realtime version	V3-V4	V3-V4	V4	V4
Masking	2021	2021	2021	2021
HRRR init for training	00Z	00Z	00Z	00Z
HRRR init for forecast	06Z	12Z	06Z	12Z
Evaluation period	01 Aug 2020 – 23 Apr 2022	01 Aug 2020 – 23 Apr 2022	23 Apr 2022 – 31 Aug 2023	23 Apr 2022 – 31 Aug 2023

characteristics and other performance features. In this section, we describe a sensitivity experiment quantifying the impact of a version mismatch in the HRRR-based RF system (see Table 9a).

In this experiment, we used the HRRR2021 version of our RF model, and compared results from applying it to the daily 00 UTC operational HRRR (consistent with the model version used in the 2018-2020 training) with results from applying it to the daily 00 UTC experimental HRRRv4, for the time period during which both were available (Aug – Dec 2020, prior to the 2 Dec 2020 operational implementation of HRRRv4). Figure 29 shows the results from this four-month test. Note that the experimental HRRRv4 was subject to much more frequent outages due to its running on a NOAA research and development machine (~25% of days are missing). It is evident that the version mismatch in the experimental HRRRv4 experiment does lead to some degradation in forecast skill on the CONUS scale (and in most

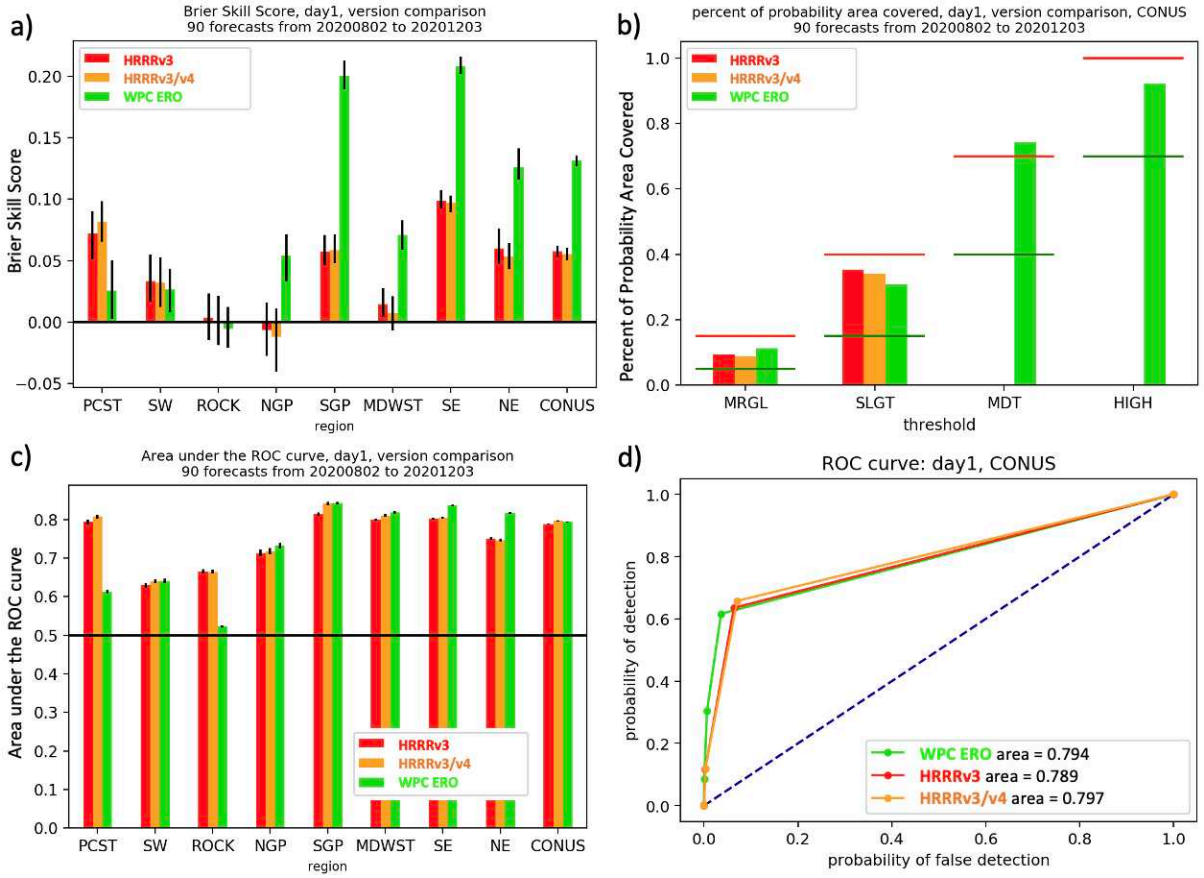

Fig. 29. Impact of daily HRRR version passed into the HRRR2021 model, for 2 Aug – 3 Dec 2020. Shown are (a) regional BSS, (b) fractional coverage of observations for each probability category: marginal (5-15%), slight (15-40%), moderate (40-70%), and high (>70%), (c) regional area under the ROC curve, and (d) CONUS ROC curves. Experiments shown are (red) HRRR2021 trained and applied using HRRRv3, (orange) HRRR2021 trained with HRRRv3 but applied to HRRRv4, and (green) WPC ERO.

sub-regions). However, we see that the differences are generally not statistically significant for this short intercomparison period. Note that HRRRv4 featured major changes to the DA methodology (Dowell et al. 2022), but those changes mostly affected shorter lead times (James et al. 2022); at the lead times used in the RF, the bias characteristics of HRRRv4 were not dramatically different from HRRRv3. For this reason, these results may not be generally applicable to other model version upgrades. However, these results do suggest that model upgrade discontinuities are not always a major problem for RF training.

4.2.2. Impact of initialization time mismatches

Here we describe the results of another experiment in which we explore the impact of an initialization time mismatch between the trained RF model and the daily HRRR forecast to which the model is applied. We carried out this experiment with both the HRRR2021 and HRRR2022 versions of the model, which were trained on the 0000 UTC initialization of the HRRR, and explored applying these trained models to the 06 and 12 UTC HRRR initializations. Once again, the mismatch in initialization time may undermine the capability of the RF to learn the performance characteristics of the HRRR; however, the assimilation of more recent observations ought to improve the HRRR forecasts for the later initialization times. We aim to quantify the net impact of these effects.

Overall, we see increasing forecast degradation with increasing difference in initialization time between the trained model and the HRRR forecast to which the model is applied (rightmost column of Fig. 30a). This makes sense, as there is some temporal coherence (or run-to-run-consistency) to the hourly HRRR forecasts; forecasters often use this information to gauge uncertainty (e.g., Benjamin et al. 2023). The difference in BSS between the 00 UTC and 06 UTC initializations is relatively small (Fig. 30a,c). However, there are substantial regional differences in the degree of degradation stemming from a 12-h initialization time mismatch (Fig. 30a,d). The largest degradation during this ~21-month period is seen in the MDWST and especially the NGP region, and to some extent in the SE (Fig. 30a,d). In the SW region, we actually see an improvement in BSS when there is a 12-h initialization time mismatch (Fig. 30a,d). On the other hand, in almost all regions, the area under the ROC curve is highest when there is a 12-h initialization time mismatch (Fig. 30b).

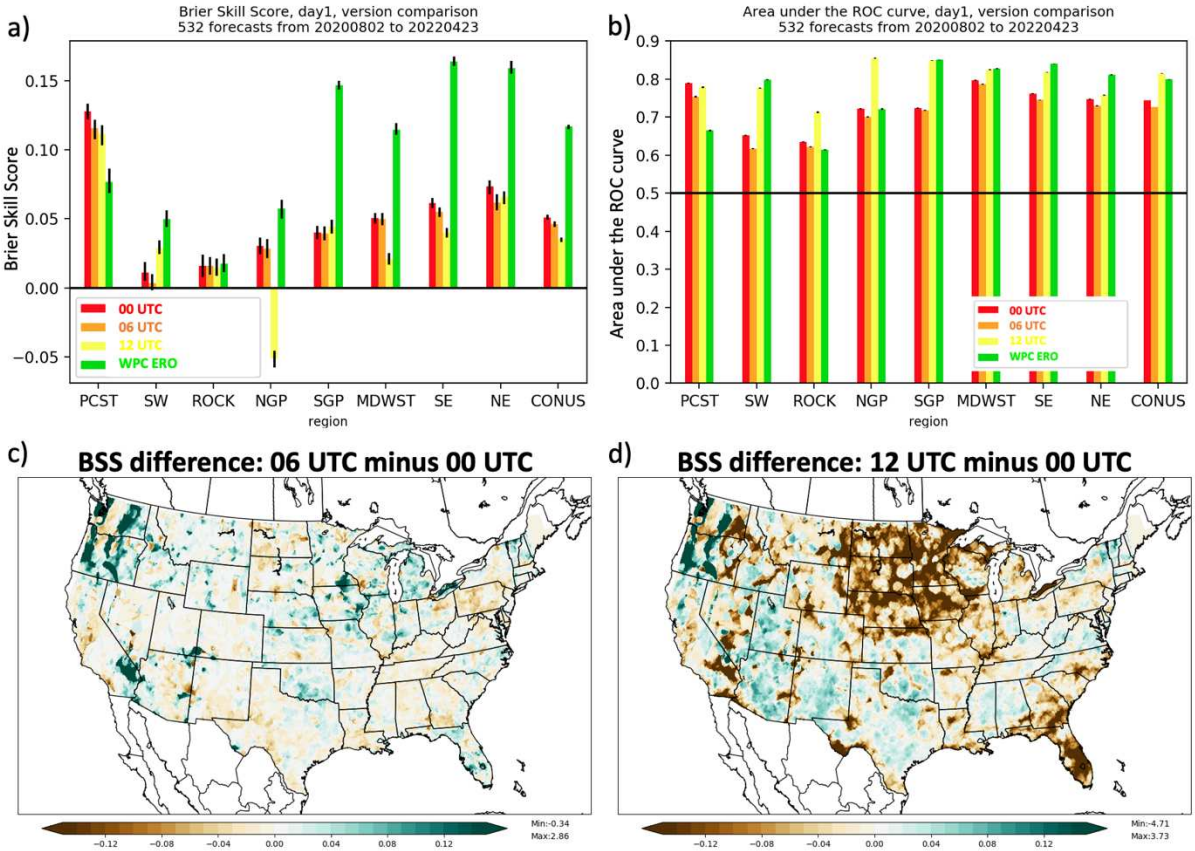

Figure 30: Experimental results comparing application of the HRRR2021 RF model to (red) 00 UTC, (orange) 06 UTC, and (yellow) 12 UTC daily HRRR simulations during 2 Aug 2020 – 23 Apr 2023. Shown are (a) regional BSS, (b) regional area under the ROC curve, (c) map of the BSS difference between the 06 UTC version and the 00 UTC version, and (d) map of the BSS difference between the 12 UTC version and the 00 UTC version.

The increase in area under the ROC curve with a 12-h initialization time mismatch stems mostly from the increase in frequency of marginal risk issuance (5% probability contour; not shown), with the 12 UTC experiment exhibiting a higher probability of detection (POD) but also a higher probability of false detection (POFD) with these probabilities (not shown). This indicates that the 12 UTC HRRR initialization has more signal for excessive rainfall events than the 00 UTC initialization. However, as is evident from Fig. 30a, the increased probabilities from

the 12 UTC experiment lead to a degradation in BSS in most regions. The only exceptions are the SW region and to some extent the SGP region, where the higher probabilities lead to a large increase in POD but not much increase in POFD. Apparently, in the SW and SGP regions for this ~21-month period, the increased signal for excessive rainfall from the 12 UTC HRRR initialization outweighed the degradation stemming from the 12-h initialization time mismatch.

Figure 31 shows the same experiment as Fig. 30, but with the more recent HRRR2022 RF version; note that there is no overlap between the evaluation periods shown in Figs. 30 and 31. With the HRRR2022 RF, and for this more recent time period, the degradation in BSS in the north-central US with an initialization time mismatch is much more muted (Figs. 31a,c,d). However, we still see the highest area under the ROC curve in most regions with a 12-h initialization time mismatch (Fig. 31b), and we still see indications of improved BSS in the SW and SGP regions when there is a 12-h initialization time mismatch (Fig. 31a,d).

Summarizing the initialization time experiments, applying the RF which was trained on the 00 UTC HRRR initialization to other HRRR initialization times does in general lead to a degradation in BSS. Larger impacts are generally seen when applying the system to the 12 UTC initialization compared to the 06 UTC initialization, presumably because the differences in HRRR forecasts are much greater. Applying the RF to the 12 UTC initialization leads to higher probabilities of excessive rainfall in almost every region, particularly using the HRRR2021 RF. This leads to an interesting north-south contrast in the impact of the initialization time mismatch. In the north-central US (the NGP and MDWST regions), the increase in probabilities leads to a large increase in POFD, which dramatically degrades the BSS for the HRRR2021 experiment. In contrast, in the south-central US (SW and SGP regions), the increase in probabilities leads to a

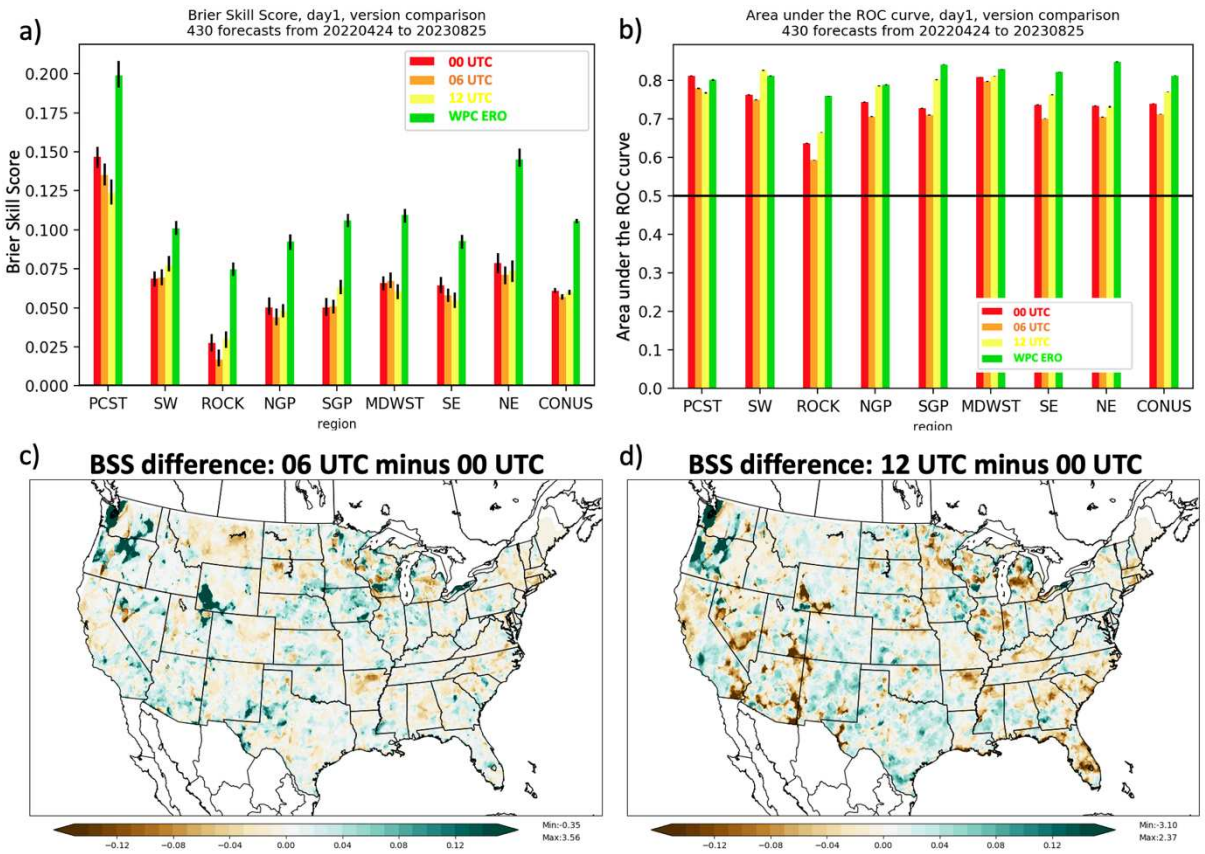

Figure 31: As in Fig. 30, but using the HRRR2022 RF model, and for the time period 24 Apr 2022 – 25 Aug 2023.

substantial increase in POD with minimal change to POFD, resulting in improved BSS with both the HRRR2021 and HRRR2022 experiments.

Because of the different experimental periods shown in Figs. 30 and 31, it is impossible to separate the effect of different seasons and different meteorological regimes during the two periods from the effect of the differences between the two trained RFs, HRRR2021 and HRRR2022. However, we can examine the frequency of UFVS events during the two periods; Figure 32 shows maps of the frequency of UFVS events during the two experimental periods shown in Figs. 30 and 31. One difference between the two periods is the relative lack of excessive rainfall events in the north-central portion of the US during 2 Aug 2020 – 23 Apr 2022

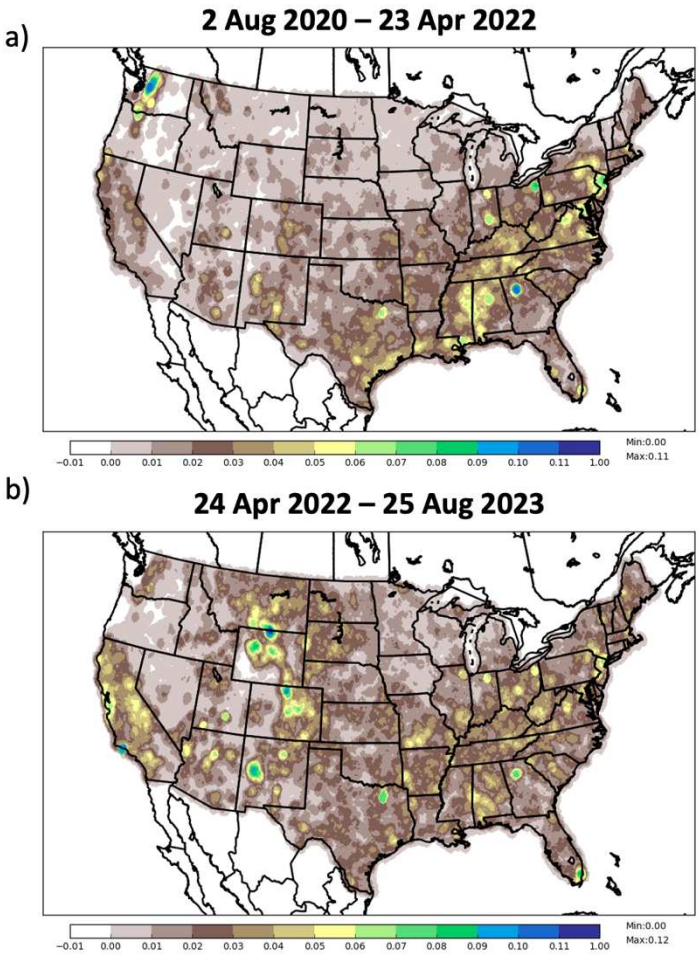

Figure 32: Relative frequency of UFVS events during (a) 2 Aug 2020 – 23 Apr 2022, and (b) 24 Apr 2022 – 25 Aug 2023.

(Fig. 32a), compared to the 24 Apr 2022 – 25 Aug 2023 experimental period (Fig. 32b). This suggests that the dramatic impacts of initialization time mismatch in the northern US shown in Fig. 30 may be based on only a few excessive rainfall events from this period. In contrast, the frequency of excessive rainfall events in the southern US (SGP and SW regions) appears relatively similar between the two experimental periods (Fig. 32), suggesting that the impact of initialization time mismatch seen in this region may be more robust.

4.2.3. Impact of training period length

In this section, we explore the impact of extension of the training period from two years to three years. In general, it is expected that a longer training period would be beneficial for

forecast skill, because it would capture a wider range of possible scenarios in the RF training. It is challenging to obtain a long training period with an operational CAM, because operational models are updated relatively frequently (every ~2 years for HRRR; Dowell et al. 2022). For this reason, our training period length experiment has the additional complication of mixing HRRR versions in the RF training period. It is a high priority for future work to evaluate the impact of extending the training period with a static model version, for example with the HRRRv4 which has now been in operations for nearly three years.

With that caveat in mind, we examine the impact of extending the training period for our RF from two to three years, which was one of the key recommendations from FFaIR 2021 (Trojniak and Correia 2021). Figure 33 shows the results of the experiment. We do see statistically significant BSS improvements to the RF predictions coming from a training period extension in this framework. The benefit is quite regionally variable, with the greatest BSS improvement seen when using a longer training period in the PCST and SW regions; BSS is virtually unchanged with a longer training period in central portions of the CONUS (Fig. 33a). In terms of forecast reliability (Fig. 33b), the extended training period does improve the underforecasting of the HRRR2021 version for the SLGT and MDT risk categories.

Figure 33d shows a map of the difference in BSS between the two RFs, and similar patterns emerge as are indicated in Fig. 33a: substantial BSS improvements are seen in the western CONUS, with more mixed results in the central and eastern US.

Due to the operational HRRR upgrade on 2 Dec 2020, it is possible that the training length extension benefit would be even greater if a static version of the underlying CAM was used for this experiment; future work should examine this question in a more controlled experiment.

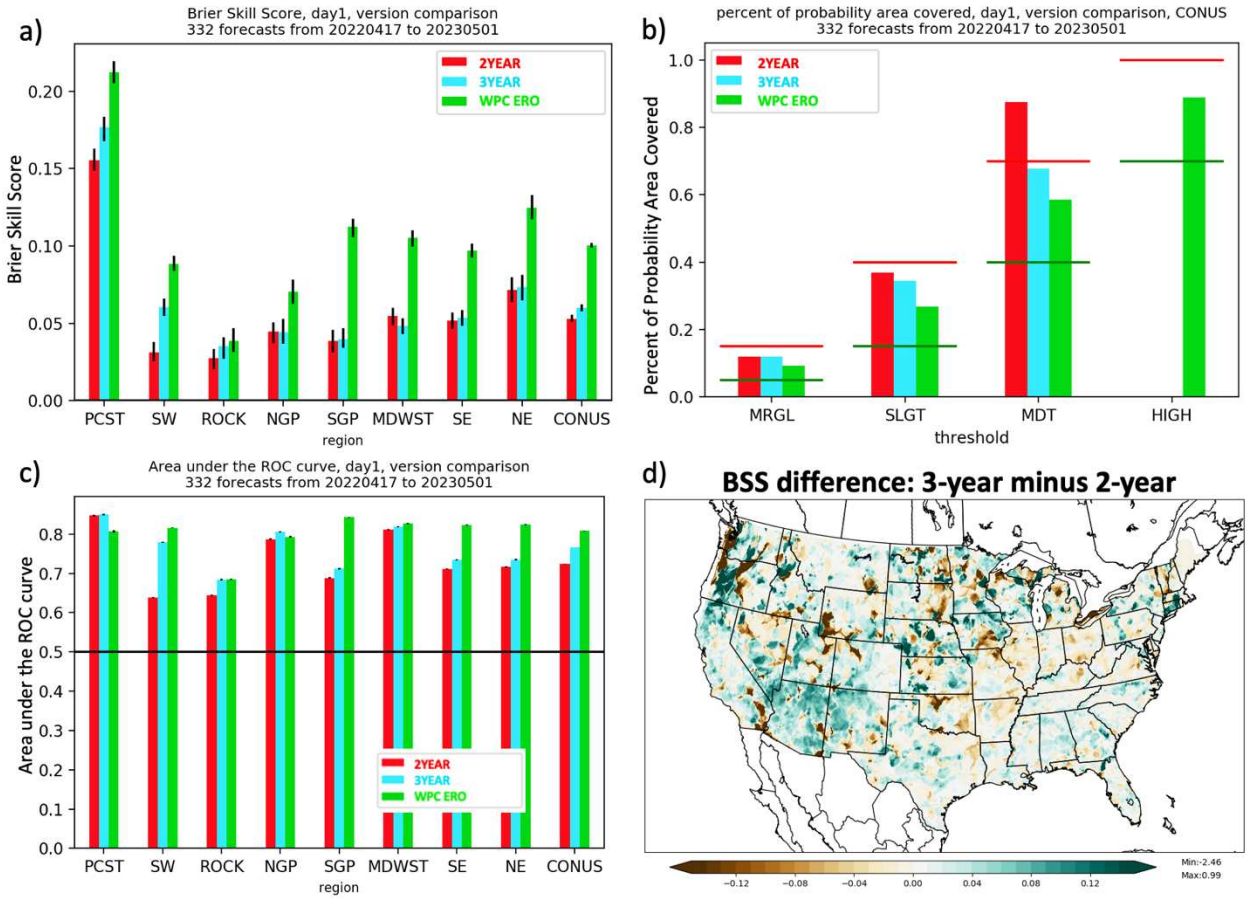

Figure 33: Experimental results comparing performance of the HRRR2021 RF model (red) to an identical model with a training period extended to three years but also mixing HRRRv3 and HRRRv4 (cyan) during 17 Apr 2022 – 1 May 2023. Shown are (a) regional BSS, (b) fractional coverage of observations for each probability category: marginal (5-15%), slight (15-40%), moderate (40-70%), and high (>70%), (c) regional area under the ROC curve, and (d) map of the BSS difference between the 3-year-training period version minus HRRR2021.

4.2.4. Impact of an additional predictor on soil conditions

One potential weakness of RFs trained exclusively on model forecasts is the lack of predictor information regarding the soil state or antecedent conditions. The effects of heavy rainfall are known to be highly variable depending on the underlying land surface type as well as soil wetness. To explore this question, we trained an additional RF which has a 16th predictor:

gridded 6-h flash flood guidance (FFG). FFG is produced by the RFCs, and is intended to reflect the amount of precipitation in a given duration that would be expected to lead to bank-full conditions. Including FFG as an additional predictor allows the RF to learn about values of FFG that are more likely to lead to flash flooding. Note that FFG exceedances are one component of the UFVS used for forecast verification, but there are other components including stream gauge observations and local storm reports of flooding. We used the daily 00 UTC 6-h gridded FFG.

As shown in Fig. 34a, including FFG leads to a statistically significant improvement in BSS on the CONUS scale (rightmost set of bars). Regionally, BSS improvement from including FFG is seen in every region except the SE and NGP. In terms of forecast reliability, including FFG greatly reduces the underprediction bias in the HRRR2022 RF (Fig. 34b), which is seen most dramatically for the slight and moderate risk categories (15% and 40% probability contours). Including FFG also leads to an improvement in forecast resolution in all regions (Fig. 34c). Regionally, the most notable BSS improvements are seen in northern California, the area from southeastern Wyoming to central New Mexico, and in Texas. Some of these BSS differences may be due to regional variability in FFG methodology. It is clear that, overall, including an additional predictor related to soil saturation and/or antecedent conditions is beneficial for the performance of the RF.

4.3. FFaIR evaluation

Since 2021, the HRRR-based RF for excessive rainfall has been evaluated both objectively and subjectively by participants at the Flash Flood and Intense Rainfall Experiment (FFaIR; Barthold et al. 2015). As shown in Table 6, each subsequent year's HRRR-based RF has had a different configuration aimed at improving the forecast skill. In this section, we

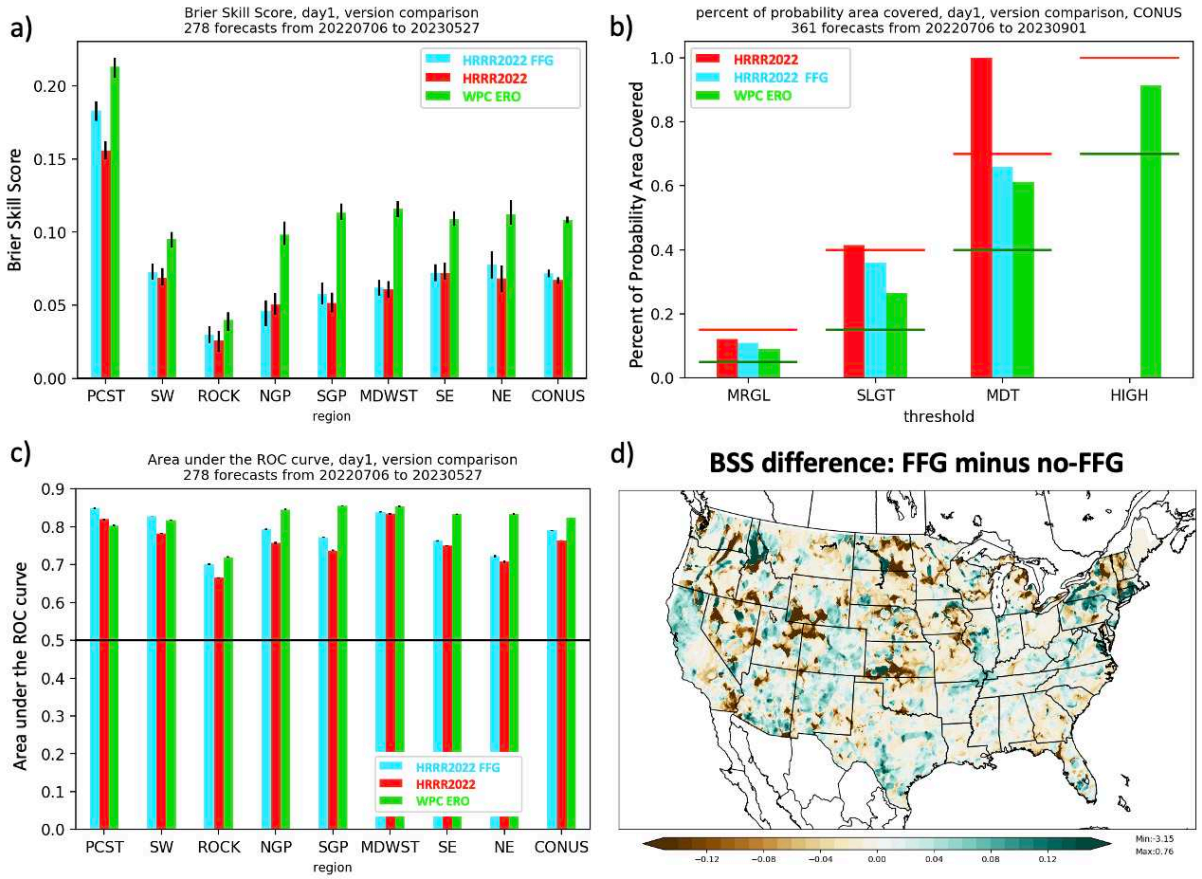

Figure 34: Experimental results comparing performance of the HRRR2022 RF model (red) to an identical model using FFG as an additional predictor, for 6 Jul – 27 May 2023. Shown are (a) regional BSS, (b) fractional coverage of observations for each probability category: marginal (5-15%), slight (15-40%), moderate (40-70%), and high (>70%), (c) regional area under the ROC curve, and (d) map of the BSS difference between the RF including FFG and the traditional RF (HRRR2022).

present quantitative verification of the three HRRR-based RF versions evaluated at FFaIR, in addition to subjective feedback from FFaIR participants.

FFaIR has been held annually during the summer months since 2013, with a focus on evaluating the utility of high-resolution models and ensembles for flash flood forecasting, exploring new tools and approaches for combining meteorological and hydrologic information, and exploring improvements to WPC's operational forecasts (Barthold et al. 2015). The experiment is generally configured with a fixed set of participants each week, including representatives from academia, research laboratories, and operational forecasting offices. The

experiments were held in person at the Hydrometeorology Testbed during 2013-2019, and virtually during 2020-2022 due to the COVID-19 pandemic. The 2023 FFaIR featured a hybrid in person / virtual format. As with all real-time experiments, there is a great deal of variability in the meteorological regime from year to year; for FFaIR, this corresponds to substantial regional shifts in the locations of the country experiencing excessive rainfall from year to year. It is important to keep this in mind when synthesizing the FFaIR results.

4.3.1. FFaIR 2021

In 2021, FFaIR was held from 21 Jun – 23 Jul 2021 (Trojniak and Correia 2021). Due to the COVID-19 pandemic, the experiment was held virtually, with two dedicated facilitators from HMT. The experiment had 58 participants spanning the four weeks of the experiment (evaluation occurring Monday – Friday). While a major focus of the experiment was an evaluation of early prototypes of the Rapid Refresh Forecast System (RRFS) based on the Finite Volume Cubed-sphere (FV3) dynamical core, a part of the evaluation focused on machine-learning based excessive rainfall outlooks provided by Colorado State University, including the first version of the HRRR-based system described herein (HRRR2021).

As noted by Trojniak and Correia (2021), heavy rainfall events over the CONUS during FFaIR 2021 were characterized by forcing associated with frontal passages and the return of the NAM, the latter of which was largely absent during FFaIR 2020. Major events from the first two weeks of FFaIR included heavy rainfall across northern Missouri associated with a stalled frontal boundary on 24-25 Jun, flash flooding in southern Utah on 28-29 Jun, and heavy rainfall in Colorado and Oklahoma on 1-2 Jul (Trojniak and Correia 2021). During the latter half of FFaIR 2021, there was extreme rainfall in Pennsylvania / New Jersey on 12-13 Jul leading to issuance of a Flash Flood Emergency, heavy rain in New York on 14 and 19 Jul, and widespread

monsoon activity in the southwestern US including training convection in southern Utah during 15-16 Jul which led to a train derailment, flash flooding and debris flows near Flagstaff, Arizona, during 13-14 Jul, and a number of burn scar debris flows in Colorado during 20-29 Jul (Trojniak and Correia 2021).

The HRRR2021 RF did not receive favorable subjective ratings during FFaIR 2021. Figure 35 shows the subjective scores during (top) the first two weeks of FFaIR 2021 and (bottom) the final two weeks of FFaIR 2021. The HRRR2021 received the lowest scores of any of the CSU RFs during both periods. However, during the second half of the experiment, when the NAM was active, HRRR2021 ratings dropped even as ratings for the other RFs generally increased. This indicates that the low ratings were due in part to poor performance (too low probabilities) in the southwestern CONUS. The FFaIR report (Trojniak and Correia 2021) hypothesized that the poor performance of the HRRR2021 in this region was due to the lack of an active monsoon in the training period, and put forth the recommendation that the HRRR-based RF be retrained for a longer period which should include the active 2021 NAM season. This recommendation motivated the extension of the training period for the subsequent version of the HRRR-based RF (Table 6).

Quantitative verification of the HRRR2021 (Fig. 36) agrees with the subjective ratings of Trojniak and Correia (2021). Figure 36 shows a quantitative evaluation of the HRRR2021 RF against the other CSU ML-based EROs, as well as the 09 UTC WPC EROs, during the nearly three years of overlapping forecasts. Note that evaluation over this period involves a much broader spectrum of seasonal and meteorological environments than just the FFaIR 2021 period. Also, note that the NSSL-WRF RF shown in Fig. 36 corresponds to the “NSSL2” model shown in Fig. 35. On the CONUS scale, the HRRR2021 has a lower BSS than the other two RFs (Fig.

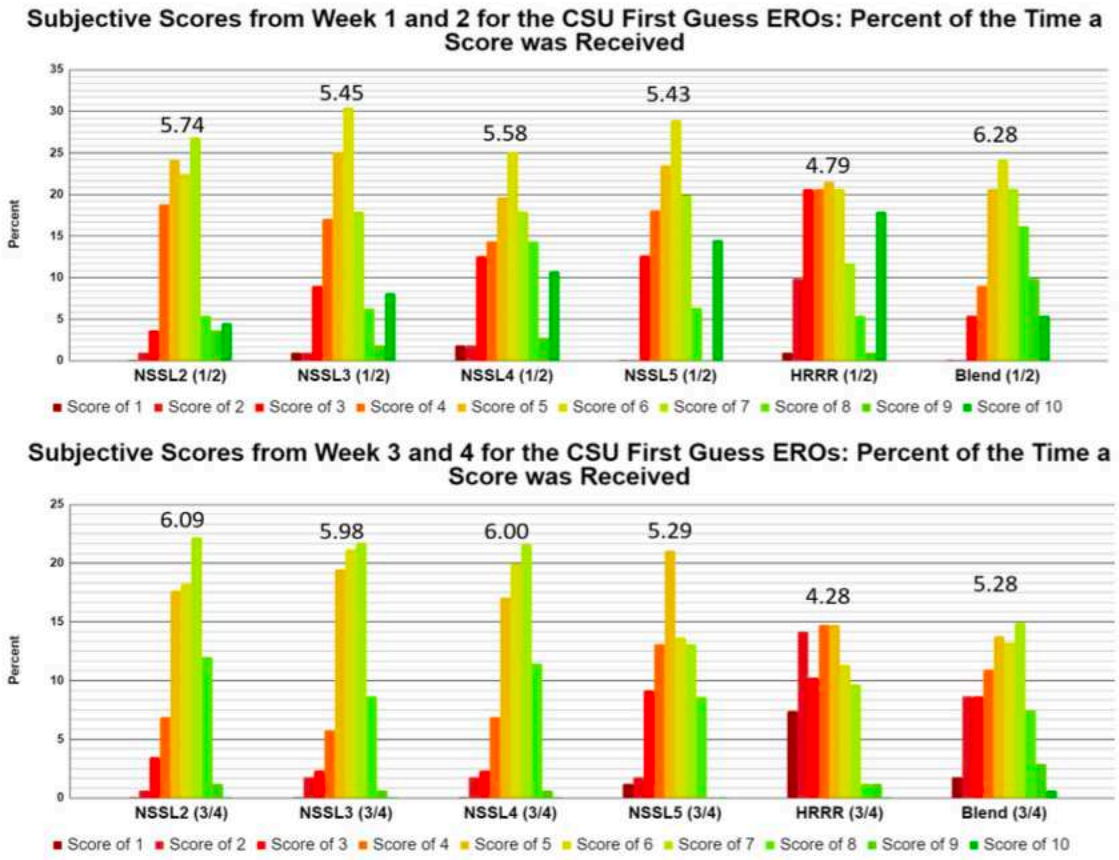

Fig. 35. Summary of subjective scores of the various RFs evaluated by FFaIR 2021 participants during (top) 21 Jun – 2 Jul 2021 and (bottom) 12-23 Jul 2021. The HRRR2021 RF is labeled as "HRRR". Figure is from Trojniak and Correia (2021).

36a). It performed particularly poorly with respect to the other RFs in the SW, ROCK, and SGP regions (Fig. 36a). In terms of reliability, Fig. 36b shows that HRRR2021 underpredicts the probability of excessive rainfall. Figure 36d shows that HRRR2021 has a lower POD than the other RFs, resulting in a smaller area under the ROC curve (Fig. 36c).

In summary, the HRRR2021 RF significantly underpredicted the risk of excessive rainfall, undermining the ability of the predictions to compete with other candidate RFs and the operational WPC ERO. As the first candidate RF for excessive rainfall based on a deterministic CAM, it is not surprising that further development was needed in order to address forecast deficiencies. Subsequent HRRR-based RF versions are described in the following sections.

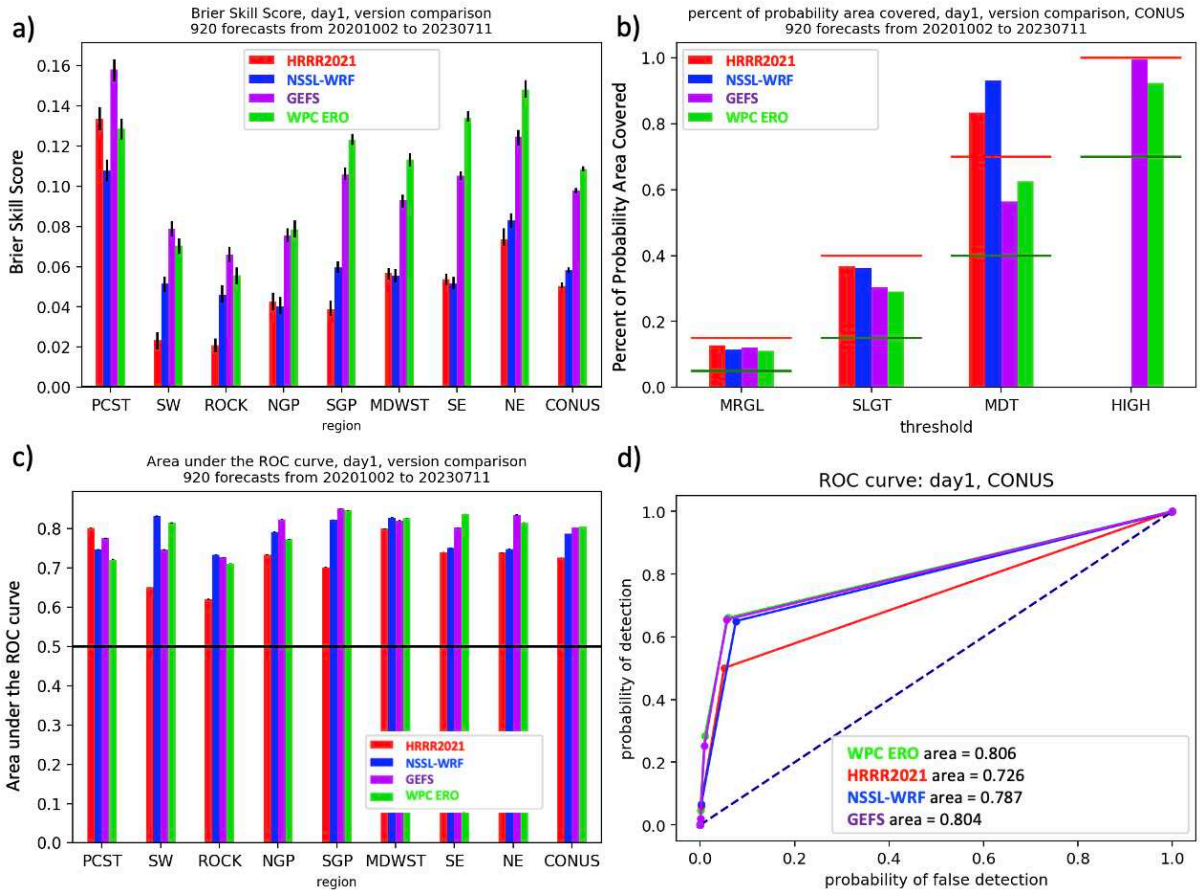

Fig. 36. Comparison of HRRR2021 version with WPC ERO and other CSU ML-based EROs during 2 Oct 2020 – 11 Jul 2023. Shown are (a) regional BSS, (b) fractional coverage of observations for each probability category: marginal (5-15%), slight (15-40%), moderate (40-70%), and high (>70%), (c) regional area under the ROC curve, and (d) ROC curve, showing HRRR2021 RF (red), NSSL-WRF RF (blue), GEFS RF (purple), and WPC ERO (green).

4.3.2. FFaIR 2022

In 2022, FFaIR was virtually held from 21 Jun – 22 Jul 2022 (Trojniak and Correia 2022), and featured a continued focus on evaluation of prototypes of the RRFS. The FFaIR 2022 period was characterized by a relative dearth of large-scale excessive rainfall events. In particular, there were relatively few mesoscale convective systems (MCSs), which are normally a major contributor to excessive rainfall over central and eastern portions of the CONUS. The NAM was relatively active during 2022, although many of the excessive rainfall events associated with the NAM occurred after the formal FFaIR period. Significant flash flooding

events during FFaIR 2022 included training convective elements in southern West Virginia and western Virginia during 12-13 Jul, and training cells in southern Kentucky / northern Tennessee during 21-22 Jul.

The HRRR2022 version was designed to address some of the concerns raised in FFaIR 2021 (Table 6). The training period was extended from two years to three years, to include the active 2021 NAM season. HRRR2022 also incorporated the OPT_AVG spatial averaging approach of Hill and Schumacher (2021).

Once again, during FFaIR 2022, the HRRR-based RF, HRRR2022, received poor subjective ratings. Figure 37 shows subjective ratings from the FFaIR 2022 participants. Participants noted that HRRR2022 "struggled to forecast an excessive rainfall risk, but when it did, the marginal risk areas were too large and noisy." It was also noted that the offshore risk areas were distracting and made the forecasts difficult to look at.

Figure 38 shows objective validation of the HRRR2022 RF for a 2022-2023 evaluation period. Comparing Fig. 36 with Fig. 38, it is evident that the HRRR2022 RF exhibits improved forecasts compared to the HRRR2021 RF. On the CONUS scale, HRRR2022 performs comparably to the NSSL-WRF in terms of BSS (Fig. 38a). The HRRR2022 continues to underforecast excessive rainfall probabilities compared to the other systems (Fig. 38b), and has inferior forecast sharpness compared to the NSSL-WRF RF (Fig. 38c,d). Notably, HRRR2022 exhibits similar BSS to the NSSL-WRF in the SW region (Fig. 38a), reflecting the improvements incorporated in the HRRR2022 version.

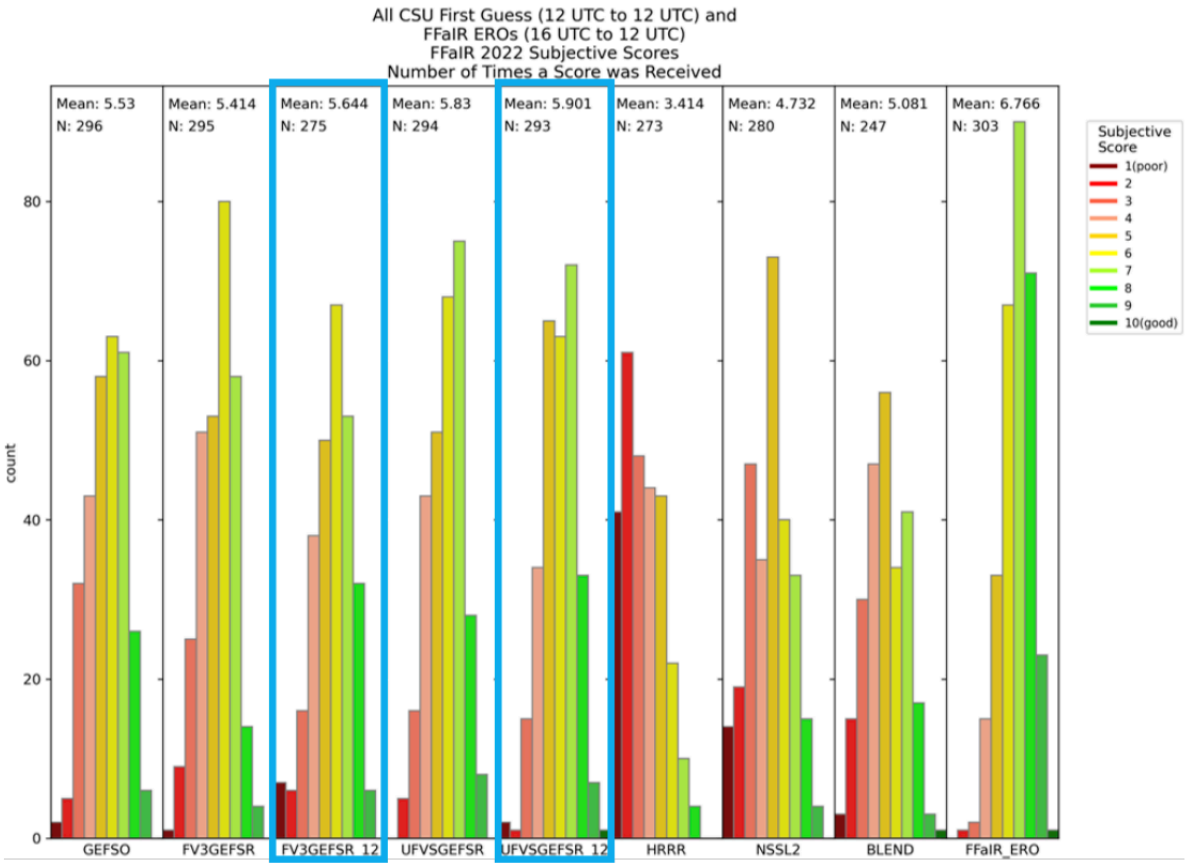

Fig. 37. Summary of subjective scores of the various RFs evaluated by FFaIR 2022 participants during 21 Jun – 23 Jul 2022. The HRRR2022 RF is labeled as "HRRR". The blue boxes indicate RFs that are initialized from 1200 UTC model data. Figure is from Trojniak and Correia (2022).

4.3.3. FFaIR 2023

The 2023 incarnation of FFaIR was somewhat modified from prior years. For the first time, the experiment was held in a hybrid format, and was extended to a longer time period, extending from 5 Jun – 11 Aug 2023 (Trojniak and Correia 2023), with three weeks off during 3-7, 17-21, and 24-29 Jul. Because the experiment only ended recently, final objective and subjective verification results are not yet available, so we present here our own objective evaluation of the HRRR2023 version of the HRRR-based RF.

The HRRR2023 RF was designed to address some noted problems with the HRRR2022

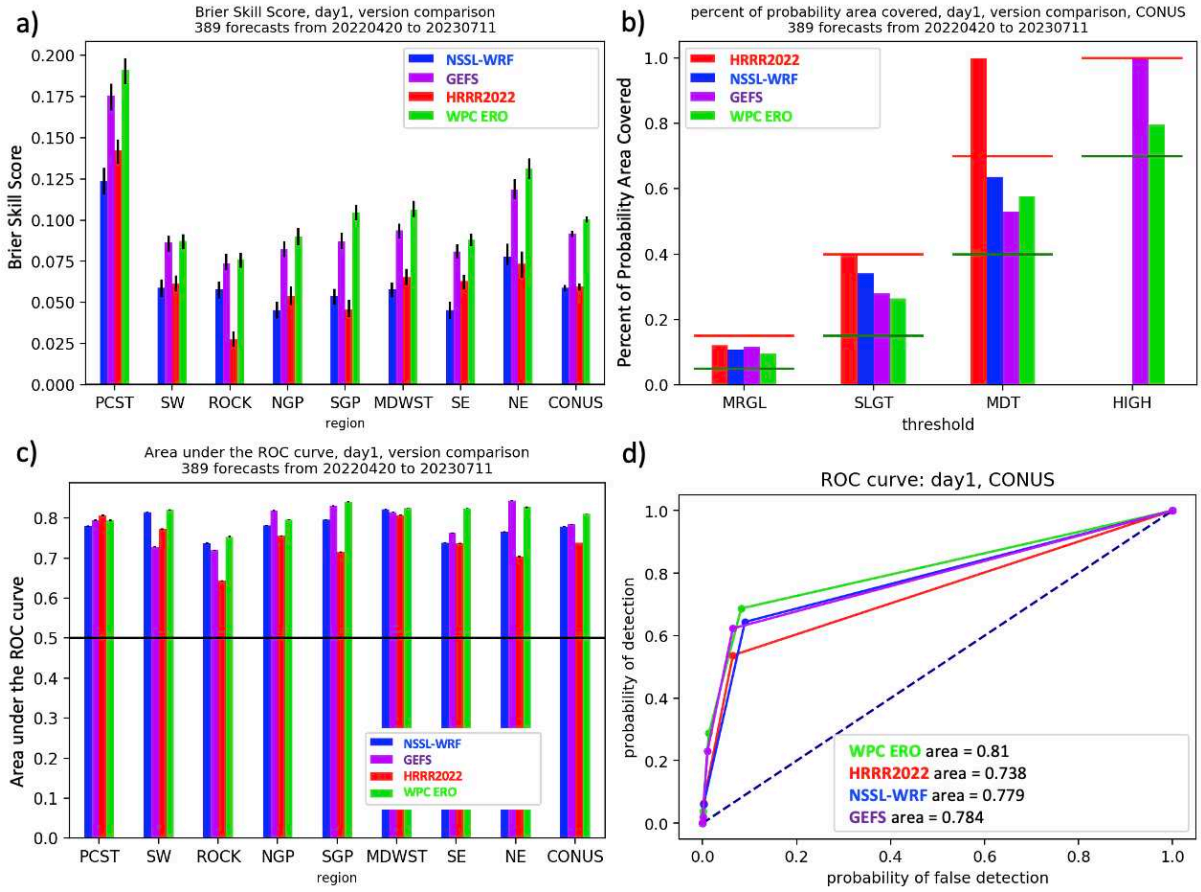

Fig. 38. As in Fig. 36, but showing HRRR2022 RF, and for the period 20 Apr 2022 – 11 Jul 2023.

RF. In particular, the training period was revised to extend from 1 Apr 2020 – 31 Mar 2023, and to include only HRRRv4. In addition, based on the results of sensitivity experiments described in chapter 3, the predictor assembly procedure was modified to use a spatial maximum or minimum of all nearby 3-km gridpoints for storm attribute predictor fields, and a spatial mean for environmental predictor fields. The predictor time step was also shortened from three hours to one hour. Finally, offshore probabilities were masked out, following the recommendation from FFaIR 2022.

Figure 39 shows quantitative evaluation of the HRRR2023 version for the period in which it has been running. On the CONUS scale, the HRRR2023 version exhibits a slightly

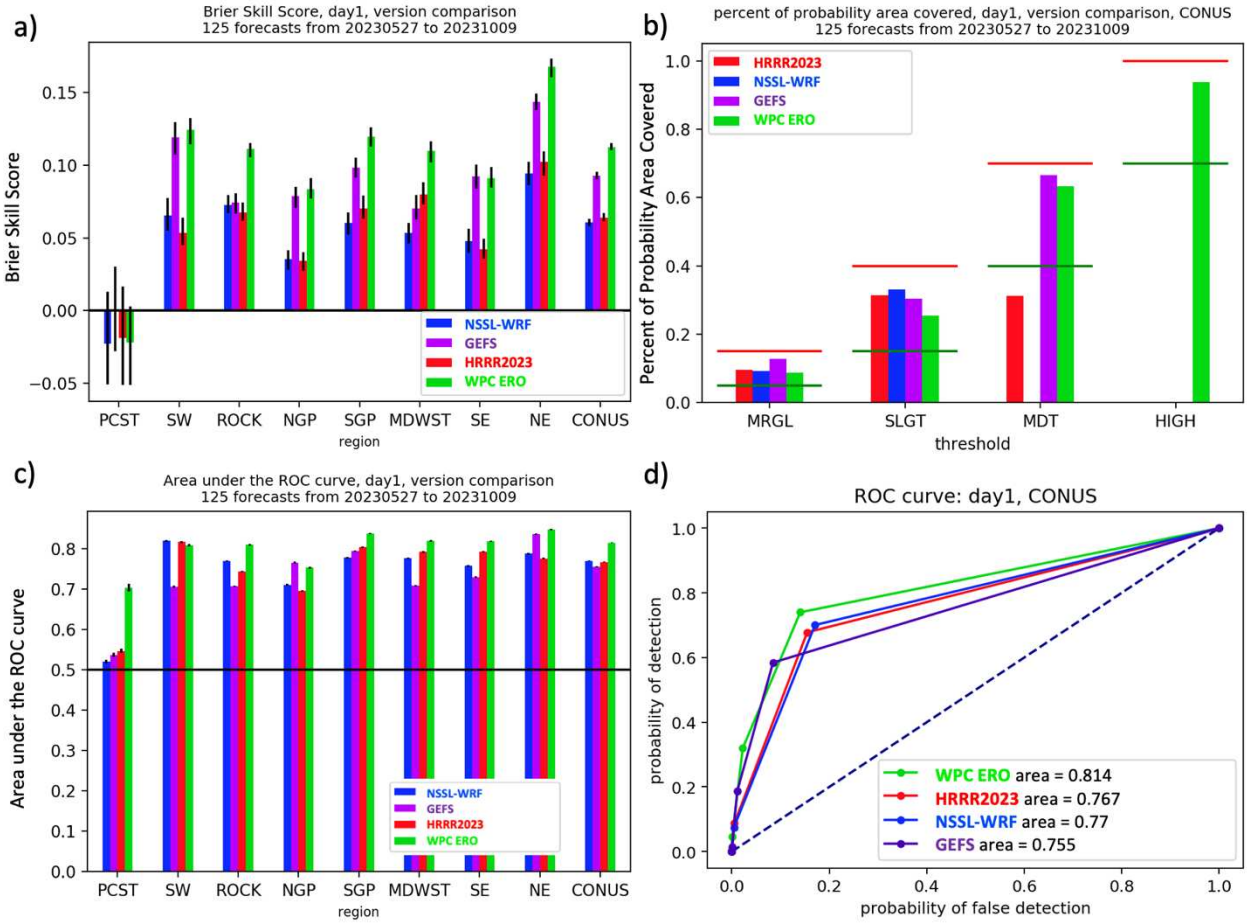

Fig. 39. As in Fig. 36, but showing HRRR2023 RF, and for the period 27 May – 9 Oct 2023.

higher BSS than the NSSL-WRF RF, which is an improvement over the HRRR2022 version (Fig. 39a). In terms of reliability (Fig. 39b), the HRRR2023 RF has a more optimal fractional coverage of UFVS events in the marginal and slight risk categories compared to the HRRR2022 RF (cf. Fig. 38b). The HRRR2023 RF also, for the first time, exhibits a comparable degree of resolution to the NSSL-WRF RF (Fig. 39c,d). Note that the relatively short period of evaluation means that excessive rainfall events were hardly observed in the PCST region, so it is difficult to draw any conclusions about the HRRR2023 RF performance in this region.

To summarize the various versions of the HRRR-based RF, Fig. 40 shows the three versions compared for the ~four-month period for which all three RFs are available. Focusing

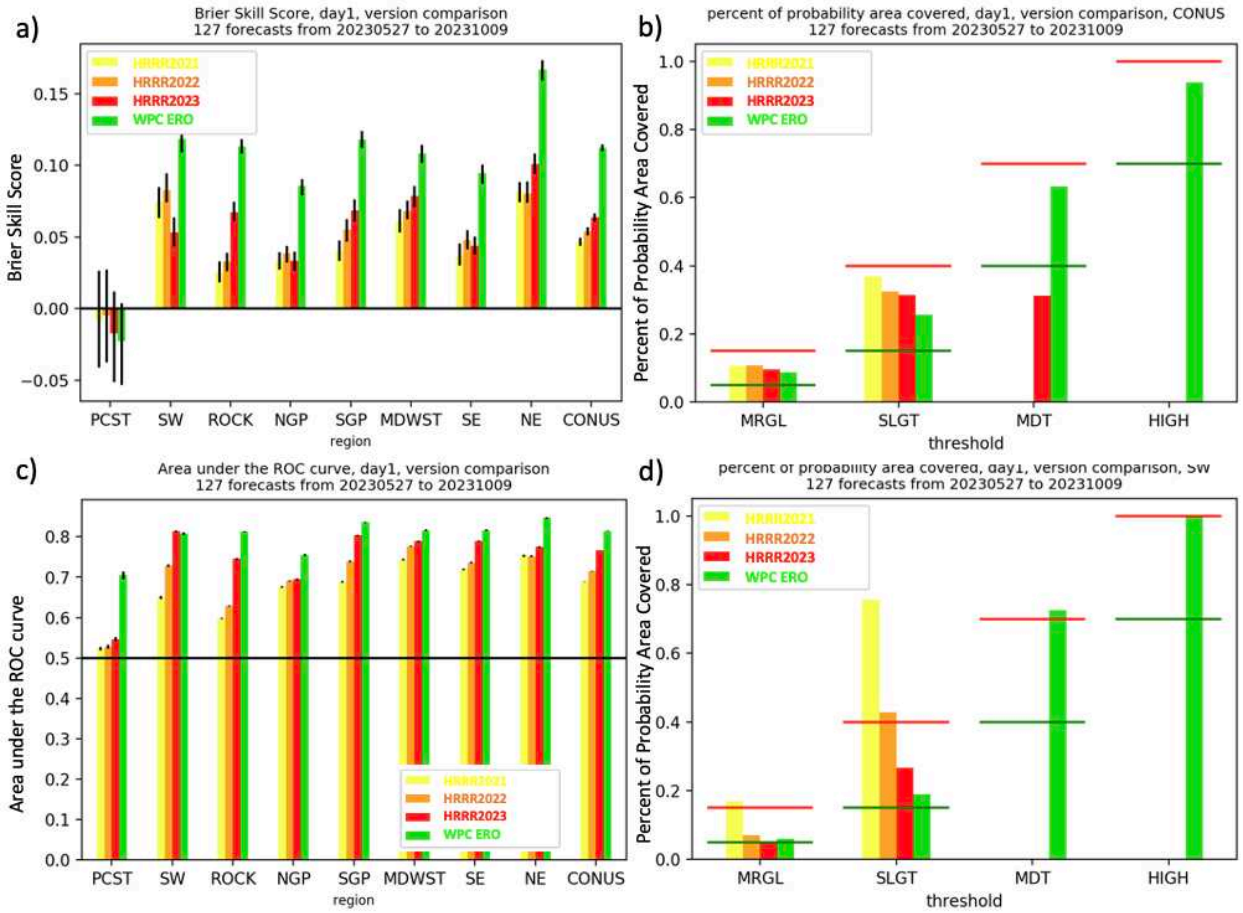

Fig. 40. Comparison of three versions of the HRRR-based RF with WPC ERO during 27 May – 9 Oct 2023. Shown are (a) regional BSS, (b) fractional coverage of observations for each probability category: marginal (5-15%), slight (15-40%), moderate (40-70%), and high (>70%) for the CONUS, (c) regional area under the ROC curve, and (d) fractional coverage of observations for each probability category: marginal (5-15%), slight (15-40%), moderate (40-70%), and high (>70%) for the SW region.

on the CONUS-wide BSS (Fig. 40a, rightmost set of bars), we see the steady increase in skill of each year's HRRR-based RF, with each year statistically significantly better than the last. This year-over-year improvement is also seen in some of the individual regions, including the NE, MDWST, SGP, and ROCK regions. In terms of reliability, we also see a year-over-year improvement in the SW (Fig. 40d), manifested most notably as a decrease in underprediction bias for the slight risk category (second set of bars). In terms of resolution, the regional area

under the ROC curve also exhibits a year-over-year improvement (Fig. 40c), both regionally and on the CONUS scale, with the HRRR2023 RF having the highest resolution in all regions.

Interestingly, the largest exception to the BSS improvements outlined above is in the SW region. In this region, and for this short time period, the HRRR2023 RF exhibits a lower BSS than the HRRR2022 RF (second set of bars in Fig. 40a). Figure 40d shows the reliability of the different HRRR-based RFs in the SW region for this period, in terms of the fractional coverage of events in each of the WPC risk categories. As is evident in Fig. 40d, the HRRR-based RF in this region has exhibited a shift from underprediction of excessive rainfall risk with the HRRR2021 RF, to a more reliable prediction of risk with the HRRR2023 RF; however, for the marginal risk category (5-15% probability), the HRRR2023 RF verifies near the lower probability bound (5%; nearly an overprediction). As documented by Gallo et al. (2016), the area under the ROC curve tends to reward overpredictions, leading to a net increase in area under the ROC curve in this region; the HRRR2023 RF actually exhibits superior AUC to the WPC ERO (Fig. 40c, second set of bars). But for BSS (Fig. 40a), the overprediction is penalized such that the HRRR2023 RF has lower BSS than the HRRR2022 RF in the SW region.

Given results described in chapter 3 exploring the impact of the spatial predictor assembly approach used in the HRRR2023 RF, as well as the shorter predictor time step, the question arises whether these results are representative of the RF performance overall. The NAM season of 2023, encapsulated by this time period, was quite unusual, with a dearth of overall classic NAM excessive rainfall events during most of the season, but a major event with the landfall of Hurricane Hilary in California on 20 Aug 2023. Figure 41 shows a time series of BSS for the SW region during the 27 May – 15 Sep 2023 period. BSS is calculated from forecasts aggregated over week-long periods, going from Friday to Thursday of each week. The

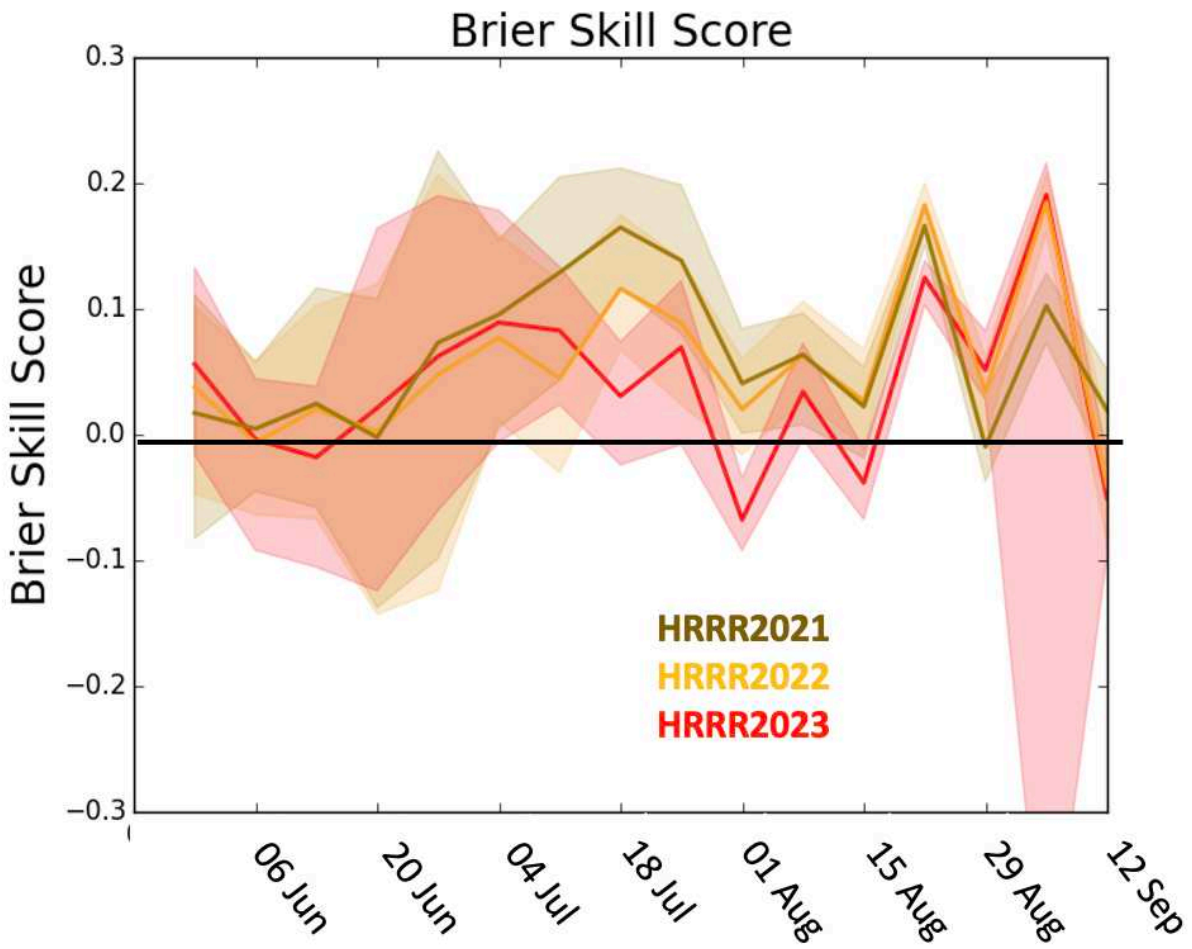

Fig. 41. Weekly time series of Brier Skill Score in the SW region during summer 2023, comparing the HRRR2021, HRRR2022, and HRRR2023 RFs. Solid lines are the BSS, with the shading indicating the 95% confidence interval.

breadth of the confidence intervals is inversely proportional to the number of UFVS events during each week-long period; it is evident that the first half of the evaluation period featured relatively few excessive rainfall events in the SW region. Starting in mid to late July, the BSS of the HRRR2023 RF drops to near zero for several weeks, and is notably lower than the BSS of either the HRRR2021 or HRRR2022 RFs. Excessive rainfall related to Hurricane Hilary took place during 19-21 Aug, and is captured in the second-to-last weekly point in the time series; all the RFs had relatively higher scores during this week, but the HRRR2023 RF still had the lowest BSS. The final week in the time series, 26 Aug – 1 Sep, featured a NAM surge and fairly

widespread excessive rainfall in the SW region; for this week, the HRRR2023 RF had the highest BSS.

Figure 42 shows some example forecasts from the three RFs. The forecasts appear broadly consistent with the results shown in Fig. 41: each subsequent year's HRRR-based RF version predicts higher probabilities of excessive rainfall in the SW region. For the Hurricane Hilary event on 20-21 Aug 2023 (Fig. 42a-c), the HRRR2021 RF predicts excessive rainfall probabilities that appear too low (10-25% in most of the regions that ended up experiencing widespread excessive rainfall). These probabilities are increased in the HRRR2022 RF, and increased further in the HRRR2023 RF, but with the introduction of some false alarm area, particularly in the eastern portion of the risk area (Arizona and southwestern Utah). For this case, based on the BSS time series shown in Fig. 41, it appears that the penalty from the HRRR2023 RF false alarm outweighed the benefit of higher probabilities where excessive rainfall was observed.

The case of 24-25 Aug 2023 is shown in Figs. 42d-f; in this case, the HRRR2021 RF forecasted probabilities of less than 5% throughout the SW region. The HRRR2022 RF had a small area of 5-10% probability in the Four Corners region. But the HRRR2023 RF did a much better job highlighting the true risk area, with a region of 15-20% probabilities near the Four Corners and a 10-15% risk area extending northeastward into western Colorado. A final case, 1-2 Sep 2023, is shown in Fig. 42g-i; major flash flooding occurred in Las Vegas, Nevada, on this day. Again, the HRRR2023 RF has the highest probabilities (>25% in southern Arizona and far southeastern California), which appeared appropriate on this day. Also, interestingly, the HRRR2023 RF had a much more successful forecast in Florida on this day.

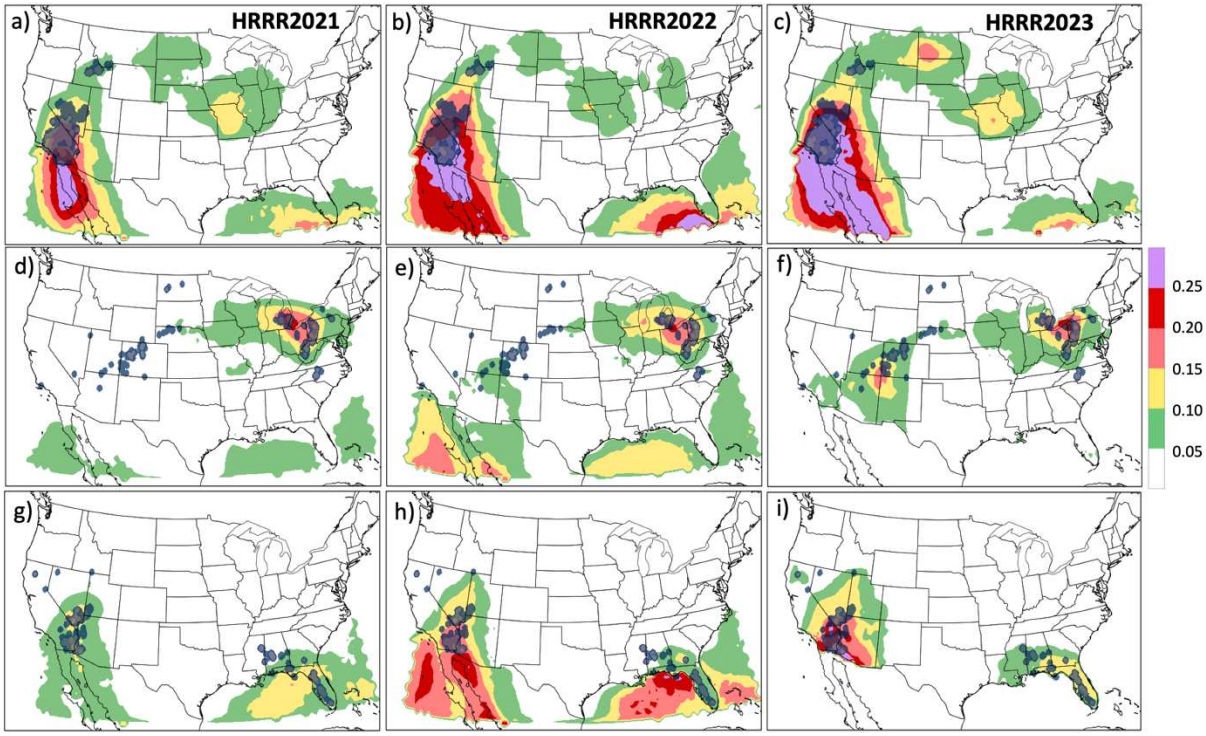

Fig. 42. RF forecasts of probability of excessive rainfall for (a-c) 12 UTC 20 Aug – 12 UTC 21 Aug 2023, (d-f) 12 UTC 24 Aug – 12 UTC 25 Aug 2023, and (g-i) 12 UTC 1 Sep – 12 UTC 2 Sep 2023. Shown are (left column) HRRR2021, (middle column) HRRR2022, and (right column) HRRR2023 RFs. Note that the color scheme is different from the operational WPC ERO risk category convention. UFVS events occurring during each 24-h period are overlaid on the forecasts.

To summarize, these preliminary results on the HRRR2023 RF suggest that the new system may have an overprediction bias in the SW; however, the unusual nature of the NAM season during 2023, in addition to the relatively short period of evaluation, suggest that additional verification is needed to obtain more robust results.

4.4. Discussion and Conclusions

In this article, we have described the first RF for excessive rainfall prediction based on an operational CAM, the HRRR. Training RFs using operational models leads to some unique challenges, which we explore using sensitivity experiments. Without the availability of a "re-forecast" dataset, such as that used for the GEFS-based system (Herman and Schumacher

2018a), time periods with a frozen model configuration are relatively short in duration for training an RF to predict rare events. We explore the impacts of this reality through testing the effects of an extended training period which mixes HRRR versions, and through testing the impact of applying an RF trained on HRRRv3 to the operational HRRRv4. Overall, our results suggest that there is a forecast degradation associated with a mismatch between the model version used in training versus that used for the daily application. We also find that extending the training period from two to three years does lead to statistically significant forecast benefit, with more benefit seen in the SW and PCST regions.

Applying a HRRR-based RF trained on the 00 UTC simulations to HRRR simulations initialized at different times does lead to a forecast degradation, although the degree of degradation varies by region of the CONUS. Interestingly, we find slightly improved forecasts in the southern CONUS when applying the 00 UTC RF to the 12 UTC daily HRRR simulations. Future work should clearly explore the value of re-training the HRRR-based RF using the 06 and 12 UTC simulations.

A separate experiment exploring the impact of including an additional predictor based on the gridded FFG for each day reveals statistically significant improvement to BSS when FFG is included. Including FFG as an additional predictor allows the RF to forecast higher probabilities of excessive rainfall more frequently, which leads to a bulk improvement in BSS. This experiment can be thought of as a first step towards integrating hydrologic predictors into RF-based excessive rainfall predictions.

We provide an evaluation of three versions of the HRRR-based RF demonstrated at the annual FFaIR experiment during 2021-23. Overall, each year's HRRR-based RF has improved upon the previous year's system, with quantitative forecast improvements seen in terms of BSS,

forecast reliability, and resolution. Comparison with other CSU RFs reveals that the HRRR-based system is competitive with that based on the NSSL-WRF, but both CAM-based RFs are inferior to the WPC ERO and to the GEFS-based RF.

For our summer 2023 evaluation period, we found some interesting version-to-version differences in performance in the SW region. In particular, the HRRR2023 RF, despite having a strictly HRRRv4 based training period, using spatial aggregation of predictors across all HRRR gridpoints, and using a shorter predictor time step, exhibited a lower BSS in the SW region than the earlier versions of the HRRR-based RF. Examination of the verification revealed that the lower BSS was due to an overprediction bias, especially for the lower risk categories. A weekly time series of BSS between the different RF versions revealed variability in the BSS of the HRRR2023 version, with lower BSS during most of the NAM season from mid-July to mid-August. However, the unusual nature of the 2023 NAM season, with relatively few excessive rainfall events during most of the season except for the exceptional few days surrounding the arrival of Hurricane Hilary, brings into question the representativeness of these results. Further verification is needed to definitively evaluate the HRRR2023 version.

These results help to highlight several promising areas of future work. The beneficial impact of using the UFVS to define events for training an RF increases the importance of further refining the UFVS. The existence of heavy snowfall events in the UFVS in the western US is a continuing challenge for verification of both operational and ML-based excessive rainfall predictions; some of the higher skill seen in the RF-based systems compared to the WPC ERO in the PCST region is due to these events (see, for example, Fig. 30a).

The improved RF skill stemming from use of FFG as an additional predictor raises the question of what other additional predictors could be brought into an RF framework? We have

explored only the 6-h FFG in this work; it is possible that additional benefit could be gained from using other durations, perhaps varying by region. Also, we have not explored varying the meteorological predictor fields from HRRR; it may be that other HRRR fields, or combinations of fields, could provide valuable information for predicting excessive rainfall. For example, one could envision an integrated water vapor flux predictor, or a warm advection predictor which is constructed from 925-hPa winds and the horizontal temperature gradient. Future work should explore the incorporation of information from additional predictors, in tandem with taking advantage of increasingly sophisticated deterministic and ensemble modeling systems, to improve the tools available to forecasters for predicting the occurrence of excessive rainfall.

CHAPTER 6: CONCLUSIONS

Flash flooding is a critical societal problem, and will remain one for the foreseeable future. It is important to evaluate current tools available for accurate prediction of excessive rainfall, for time scales ranging from the next couple of hours to sub-seasonal forecasts, as well as to advance the skill of available tools.

As described in chapter 2, there are many outstanding issues with existing QPE with regards to how well they represent excessive rainfall leading to flash flooding. Other work has validated QPE against high-quality, independent observations (e.g., Nelson et al. 2016; Cocks et al. 2016); in this work, we follow the approach of Herman and Schumacher (2018c) to evaluate the correspondence of QPE exceedances of various thresholds with observed flash flooding events. We find that dramatic uncertainty persists among the various available QPE datasets, even in relatively well-observed regions. In agreement with previous studies, there is greater correspondence between FFRs and QPE exceedances in the eastern CONUS than in the western CONUS. Stage IV QPE has the best correspondence with observed FFRs in almost all regions of the CONUS, and 6-h duration QPE exceedances correspond better with FFRs than 1-h duration QPE exceedances everywhere except in the western CONUS. Overall, FFG is the best threshold for correspondence with FFRs, which is an encouraging result highlighting the value of the FFG dataset.

Evaluating HRRR QPF in this same framework reveals generally inferior correspondence to FFRs in most regions, which is an expected result. However, in some regions of the coverage with sparse rain gauges and poor radar coverage, such as the SW region, HRRR 1-h QPF

exceedances actually correspond better with FFRs than any QPE exceedances, suggesting that model forecasts should play a role in QPE in poorly-observed regions.

Since prediction of excessive rainfall remains so challenging, it is important to advance the skill of tools available to operational forecasters. In the remainder of this work, we describe an RF based on predictors from the deterministic HRRR model. Sensitivity experiments with this RF complement recent work on RFs using predictors from global and high-resolution ensembles. In chapter 3, we described sensitivity experiments exploring the spatial aggregation of predictors (how best to reduce the dimensionality of high-resolution predictor information), as well as the predictor time step, and use of a "poor man's ensemble" of predictor information constructed from time-lagged HRRR initializations. We find that it is important to consider information from every high-resolution predictor grid point, and we find best performance when using a spatial maximum or minimum for "storm attribute" type predictors. We find minimal impact of using a shorter (1-h) time step compared to a longer (3-h) time step. The spatial aggregation of predictor information benefits forecasts most in cases of small-scale precipitation maxima, reflecting the importance of providing the RF with an accurate representation of the magnitude of the expected precipitation.

Calculation of feature contributions using the tree interpreter software sheds some light on the reasons for improved forecasts from the spatial aggregation experiment. In particular, taking a spatial maximum of the QPF predictor allows that predictor to make larger positive contributions to the resulting excessive rainfall forecast compared to considering the QPF predictor information only at sparse input grid points.

Use of time-lagged ensemble predictor information from HRRR allows us to explore whether the inclusion of ensemble information is an important factor in RF performance for

excessive rainfall prediction. We find a statistically significant improvement in HRRR-based RF predictions when we include time-lagged ensemble predictor information. Note that the time-lagged ensemble of the 0000, 0600, and 1200 UTC HRRR initializations is far from a formally designed ensemble, and likely does not provide the quality of predictor information as could be contained in a formal CAM ensemble; use of predictors from a formal CAM ensemble is left for future work, but would likely mirror successes with RFs predicting severe convective hazards based on the High Resolution Ensemble Forecast (HREF) system (Loken et al. 2020) or the Warn on Forecast System (WoFS; Clark and Loken 2022).

In chapter 4, we described a number of additional sensitivity experiments, exploring the impact of a version mismatch between the RF trained on the HRRRv3 and applied to the HRRRv4 for daily forecasts, the impact of a training length extension from two to three years, the impact of applying the RF trained on the 0000 UTC HRRR initialization to other HRRR initializations, and the impact of including an additional predictor based on Flash Flood Guidance (FFG). We find only a small degradation in forecast skill when there is a mismatch in HRRR version, but improved forecasts when using an extended training period; the latter experiment is complicated by the fact that the three-year training period included both HRRRv3 and HRRRv4. Use of the RF trained on the 0000 UTC HRRR initialization to issue forecasts based on the 0600 and 1200 UTC HRRR initializations also shows somewhat mixed results. In all regions but the NGP and MDWST, results are consistent across two different versions of the HRRR-based RF. Skill generally decreases with increasing initialization time difference in the eastern CONUS, but is improved relative to a consistent initialization time when the 0000 UTC RF is applied to the 1200 UTC HRRR initialization. In the NGP and MDWST, we found

dramatically degraded skill when applying the 0000 UTC RF to the 1200 UTC initialization with an earlier version of the HRRR-based RF, but not with a later version of the HRRR-based RF.

Finally, we evaluate three versions of the HRRR-based RF which were examined in realtime at the Flash Flood and Intense Rainfall (FFaIR) experiment. On the CONUS scale, we find statistically significant skill improvements with each subsequent year's HRRR-based RF, which reflects the knowledge gained from many of the sensitivity experiments described earlier. Evaluation of the most recent version of the HRRR-based RF is somewhat challenging due to the short period of record, but we find that the HRRR2023 version exhibits degraded BSS in the SW region. This is counter-intuitive due to the incorporation of the spatial aggregation approach and finer time step described in chapter 3. Further investigation reveals that the lower BSS stems from an overprediction of risk at the lower probability ranges (i.e., marginal risk predictions that are too numerous and spatially extensive). Future work should continue to investigate the reasons for this overprediction, and the associated meteorological situations.

Several other important avenues for future investigation are suggested by our findings. In particular, additional work examining the impact of operational model upgrades could be informative. The results shown here are not necessarily representative of other model upgrades, and there could be an important dependence on the nature of associated data assimilation and model changes. In addition, a cleaner comparison of training period length is needed, in which fixed model versions are used for training. Future work could explore how the forecast impacts of initialization time mismatch (e.g., applying an RF trained on the 0000 UTC HRRR to other HRRR initializations) are associated with the variability in raw model performance across initialization times. It would also be interesting to directly re-train an RF based on the 0600 and

1200 UTC HRRR initializations, although the latter may have limited operational utility due to the forecast latency.

The fundamental question remains: why are the CAM-based RFs still inferior to the GEFS-based RF? Is it because of the higher signal to noise ratio? Is it due to the fundamental challenge of predicting convective organization? Is it the lack of formal ensemble information? In the next few years, a formal ensemble based on the next-generation Rapid Refresh Forecast System (RRFS) will become operational in the US, providing rich opportunities for post-processing and machine learning. An additional opportunity coming in the RRFS era is the extension of convection allowing grid spacing beyond the CONUS to cover all of North America, Hawaii, and even the Arctic, raising the possibility of RF-based predictions of hazardous weather outside of CONUS. Of course, the challenge will be the availability of datasets of observed events for constructing a target vector. As more sophisticated modeling tools reach operations, RF systems, as well as more advanced ML approaches, will be able to take advantage of new products to provide improved forecasts for high-impact weather events, including excessive rainfall, which promises to remain a major societal challenge in the coming decades.

REFERENCES

Ahmadalipour, A., and H. Moradkhani, 2019: A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996-2017. *J. Hydrology,* **578,** 124106, https://doi.org/10.1016/j.hydrol.2019.124106.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.,* **47,** 806-818, https://doi.org/10.1175/2007JAMC1611.1.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. *Bull. Amer. Meteor. Soc.,* **96,** 1859-1866, https://doi.org/10.1175/BAMS-D-14-00201.1.

Benjamin, S. G., E. P. James, E. J. Szoke, P. T. Schlatter, and J. M. Brown, 2023: The 30 December 2021 Colorado Front Range windstorm and Marshall Fire: Evolution of surface and 3-d structure, NWP guidance, NWS forecasts and decision support. *Wea. Forecasting,* in press.

Benjamin, S. G., E. P. James, M. Hu, C. R. Alexander, T. T. Ladwig, J. M. Brown, S. S. Weygandt, D. D. Turner, P. Minnis, W. L. Smith, Jr., and A. K. Heidinger, 2021: Stratiform cloud-hydrometeor assimilation for HRRR and RAP model short-range weather prediction. *Mon. Wea. Rev.,* **149,** 2673-2694, https://doi.org/10.1175/MWR-D-20-0319.1.

Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2006: Version 3.0: Delaware, District of Columbia, Illinois, Indiana, Kentucky, Maryland, New Jersey, North Carolina, Ohio, Pennsylvania, South Carolina, Tennessee, Virginia, West Virigina. Vol. 2, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 295 pp., https://www.weather.gov/media/owp/oh/hdsc/docs/Atlas14_Volume2.pdf.

Breiman, L., Random forests. *Mach. Learn.,* **45,** 5-32, https://doi.org/10.1023/A:1010933404324.

Brothers, M. D., and C. L. Hammer, 2022: Random forest approach for improving non-convective high wind forecasting across southeast Wyoming. *Wea. Forecasting,* **38,** 37-67, https://doi.org/10.1175/WAF-D-21-0215.1.

Broxton, P., P. A. Troch, M. Schaffner, C. Unkrich, and D. Goodrich, 2014: An all-season flash flood forecasting system for real-time operations. *Bull. Amer. Meteor. Soc.,* **95,** 399-407, https://doi.org/10.1175/BAMS-D-12-00212.1.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning-based probabilistic hail predictions for operational forecasting. *Wea. Forecasting,* **35,** 149-168, https://doi.org/10.1175/WAF-D-19-0105.1.

Burke, P. C., A. Lamers, G. Carbin, M. J. Erickson, M. Klein, M. Chenard, J. McNatt, and L. Wood, 2023: The excessive rainfall outlook at the Weather Prediction Center: Operational definition, construction, and real-time collaboration. *Bull. Amer. Meteor. Soc.,* **104,** E542-E562, https://doi.org/10.1175/BAMS-D-21-0281.1.

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central US. *J. Adv. Model. Earth Syst.,* **7,** 1248-1264, https://doi.org/10.1002/2015MS000497.

Bytheway, J. L., C. D. Kummerow, and C. Alexander, 2017: A features-based assessment of warm season precipitation forecasts from the HRRR model over three years of development. *Wea. Forecasting,* **32,** 1841-1856, https://doi.org/10.1175/WAF-D-17-0050.1.

Bytheway, J. L., M. Hughes, K. Mahoeny, and R. Cifelli, 2020: On the uncertainty of the high-resolution hourly quantitative precipitation estimates in California. *J. Hydrometeor.,* **21,** 865-879, https://doi.org/10.1175/JHM-D-19-0160.1.

Calianno, M., I. Ruin, and J. J. Gourley, 2013: Supplementing flash flood reports with impact classifications. *J. Hydrology,* **477,** 1-16, https://doi.org/10.1016/j.jhydrol.2012.09.036.

Chapman, W. E., L. D. Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S.-P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.,* **150,** 215-234, https://doi.org//10.1175/MWR-D-21-0106.1.

Chen, S., J. J. Gourley, Y. Hong, P. E. Kirstetter, J. Zhang, K. Howard, Z. L. Flamig, J. Hu, and Y. Qi, 2013: Evaluation and uncertainty estimation of NOAA/NSSL next-generation national mosaic quantitative precipitation estimation product (Q2) over the continental United States. *J. Hydrometeor.,* **14,** 1308-1322, https://doi.org/10.1175/JHM-D-12-0150.1.

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

Clark, A., and Coauthors, 2021: Spring Forecast Experiment 2021: Preliminary findings and results. NSSL Report. Available online at https://hwt.nssl.noaa.gov/sfe/2021/docs/HWT_SFE_2021_Prelim_Findings_FINAL.pdf

Clark, R. A., J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, 2014: CONUS-wide evaluation of National Weather Service flash flood guidance products. *Wea. Forecasting,* **29,** 377-392, https://doi.org/10.1175/WAF-D-12-00124.1.

Cocks, S. B., S. M. Martinaitis, B. Kaney, J. Zhang, and K. Howard, 2016: MRMS QPE performance during the 2013/14 cool season. *J. Hydrometeor.,* **17,** 791-810, https://doi.org/10.1175/JHM-D-15-0095.1.

Doswell, C. A. III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting,* **4,** 560-581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Dougherty, K. J., J. D. Horel, and J. E. Nachmakin, 2021: Forecast skill for California heavy precipitation periods from the High-Resolution Rapid Refresh model and the Coupled Ocean-Atmosphere Mesoscale Prediction System. *Wea. Forecasting,* **36,** 2275-2288, https://doi.org/10.1175/WAF-D-20-0182.1.

Dowell, D. C., C. R. Alexander, E. P. James, S. S. Weygandt, S. G. Benjamin, G. S. Manikin, B. T. Blake, J. M. Brown, J. B. Olson, M. Hu, T. G. Smirnova, T. Ladwig, J. S. Kenyon, R. Ahmadov, D. D. Turner, J. D. Duda, and T. I. Alcott, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly-updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting,* **37,** 1371-1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Downton, M. W., J. Z. Barnard Miller, and R. A. Pielke, Jr., 2005: Reanalysis of U.S. National Weather Service flood loss database. *Nat. Hazards Rev.,* **6,** 13-22, https://doi.org/10.1061/(ASCE)1527-6988(2005)6:1(13).

English, J. M., D. D. Turner, T. I. Alcott, W. R. Moninger, J. L. Bytheway, R. Cifelli, and M. Marquis, 2021: Evaluating operational and experimental HRRR model forecasts of atmospheric river events in California. *Wea. Forecasting,* **36,** 1925-1944, https://doi.org/10.1175/WAF-D-21-0081.1.

Erickson, M. J., B. Albright, and J. A. Nelson, 2021: Verifying and redefining the Weather Prediction Center's Excessive Rainfall Outlook forecast product. *Wea. Forecasting,* 325-340, https://doi.org/10.1175/WAF-D-20-0020.1.

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. *J. Appl. Meteor. Climatol.,* **58,** 2591-2604, https://doi.org/10.1175/JAMC-D-19-0097.1.

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

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting,* **31,** 273-295, https://doi.org/10.1175/WAF-D-15-0134.1.

Ghasabi, Z., S. Karami, A. Vazifeh, and M. Habibi, 2023: Investigating the unprecedented summer 2022 penetration of the Indian monsoon to Iran and evaluation of global and regional model forecasts. *Dynamics and Atmospheres and Oceans,* **103,** 101386, https://doi.org/10.1016/j.dynatmoce.2023.101386.

Gourley, J. J., Z. L. Flamig, H. Vergara, P.-E. Kirstetter, R. A. Clark III, E. Argyle, A. Arthur, S. Martinaitis, G. Terti, J. M. Erlingis, Y. Hong, and K. W. Howard, 2017: The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States. *Bull. Amer. Meteor. Soc.,* **98,** 361-372, https://doi.org/10.1175/BAMS-D-15-00247.1.

Gourley, J. J., Y. Hong, Z. L. Flamig, A. Arthur, R. Clark, M. Calianno, I. Ruin, T. Ortel, M. E. Wieczorek, P.-E. Kirstetter, E. Clark, and W. F. Krajewski, 2013: A unified flash flood database across the United States. *Bull. Amer. Meteor. Soc.,* **94,** 799-805, https://doi.org/10.1175/BAMS-D-12-00198.1.

Hammond, N., 2018: A comparison between 2016 flash flood observations and rainfall ARIs across the north-central United States. Preprints, *32nd Conf. on Hydrology,* Austin, TX, Amer. Meteor. Soc., 42. [Available online at ams.confex.com/ams/98Annual/webprogram/Paper326494.html]

Harp, R. D., and D. E. Horton, 2022: Observed changes in daily precipitation intensity in the United States. *Geophys. Res. Lett.,* **49,** e2022GL099955, https://doi.org/10.1029/2022GL099955.

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

Herman, G. R., and R. S. Schumacher, 2016b: Extreme precipitation in models: An evaluation. *Wea. Forecasting,* **31,** 1853-1879, https://doi.org/10.1175/WAF-D-16-0093.1.

Herman, G. R., and R. S. Schumacher, 2018a: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.,* **146,** 1571-1599, https://doi.org/10.1175/MWR-D-17-0250.1.

Herman, G. R., and R. S. Schumacher, 2018b: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.,* **146,** 1785-1812, https://doi.org/10.1175/MWR-D-17-0307.1.

Herman, G. R., and R. S. Schumacher, 2018c: Flash flood verification: Pondering precipitation proxies. *J. Hydrometeor.,* **19,** 1753-1776, https://doi.org/10.1175/JHM-D-18-0092.1.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Wea. Forecasting,* **36,** 1693-1711, https://doi.org/10.1175/WAF-D-21-0026.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.,* **148,** 2135-2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hou, D., M. Charles, Y. Luo, Z. Toth, Y. Zhu, R .Krzysztofowicz, Y. Lin, P. Xie, D.-J. Seo, M. Pena, and B. Cui, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.,* **15,** 2542-2557, https://doi.org/10.1175/JHM-D-11-0140.1.

Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating High-Resolution Rapid Refresh model. *Wea. Forecasting,* **28,** 921-939, https://doi.org/10.1175/WAF-D-12-00085.1.

James, E. P., C. R. Alexander, D. C. Dowell, S. S. Weygandt, S. G. Benjamin, G. S. Manikin, J. M. Brown, J. B. Olson, M. Hu, T. G. Smirnova, T. Ladwig, J. S. Kenyon, and D. D. Turner, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Wea. Forecasting,* **37,** 1397-1417, https://doi.org/10.1175/WAF-D-21-0130.1.

Javier, J. R. N., J. A. Smith, K. L. Meierdiercks, M. L. Beack, and A. J. Miller, 2007: Flash flood forecasting for small urban watersheds in the Baltimore metropolitan region. *Wea. Forecasting,* **22,** 1331-1344, https://doi.org/10.1175/2007WAF2006036.1.

Justin, A. D., C. Willingham, A. McGovern, and J. T. Allen, 2023: Toward operational real-time identification of frontal boundaries using machine learning. *Artificial Intelligence for the Earth Systems,* **2,** 1-26, https://doi.org/10.1175/AIES-D-22-0052.1.

Kean, J. W., D. M. Staley, J. T. Lancaster, F. K. Rengers, B. J. Swanson, J. A. Coe, J. L. Hernandez, A. J. Sigman, K. E. Allstadt, and D. N. Lindsay, 2019: Inundation, flow dynamics, and damage in the 9 January 2018 Montecito debris-flow event, California, USA: Opportunities and challenges for post-wildfire risk assessment. *Geosphere* (2019) 15 (4): 1140-1163, https://doi.org/10.1130/GES0248.1.

Lawson, J. R., J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018: Advancing from convection-allowing NWP to Warn-on-Forecast: Evidence of progress. *Wea. Forecasting,* **33,** 599-607, https://doi.org/10.1175/WAF-D-17-0145.1.

Lehmkuhl, F., H. Schuttrumpf, J. Schwarzbauer, C. Brull, M. Dietze, P. Letmathe, C. Volker, and H. Hollers, 2022: Assessment of the 2021 summer flood in central Europe, *Environmental Sciences Europe,* **34,** 107, https://doi.org/10.1186/s12302-022-00685-1.

Li, Z., S. Gao, M. Chen, J. J. Gourley, C. Liu, A. F. Prein, and Y. Hong, 2022: The conterminous United States are projected to become more prone to flash floods in a high-end emissions scenario. *Nature Communications Earth and Environment,* **3,** 86, https://doi.org/10.1038/s43247-022-00409-6.

Lincoln, W. S., and R. F. L. Thomason, 2018: A preliminary look at using rainfall average recurrence intervals to characterize flash flood events for real-time warning forecasting. *J. Operational Meteor.,* **6** (2), 13-22, https://doi.org/10.15191/nwajom.2018.0602.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting,* **35,** 1605-1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Loken, E. D., A. J. Clark, and A. McGovern, 2022: Comparing and interpreting differently-designed random forests for next-day severe weather hazard prediction. *Wea. Forecasting,* **37,** 871-899, https://doi.org/10.1175/WAF-D-21-0138.1.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. *Wea. Forecasting,* **34,** 2017-2044, https://doi.org/10.1175/WAF-D-19-0109.1.

Lundquist, J., M. Hughest, E. Gutman, and S. Kapnick, 2019: Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bull. Amer. Meteor. Soc.,* **100,** 2473-2490, https://doi.org/10.1175/BAMS-D-19-0001.1.

Maddox, R. A., L. R. Hoxit, C. F. Chappell, and F. Caracena, 1978: Comparison of meteorological aspects of the Big Thompson and Rapid City flash floods. *Mon. Wea. Rev.,* **106,** 375-389, https://doi.org/10.1175/1520-0493(1978)106<0375:COMAOT>2.0.CO;2.

Maddox, R. A., C. F. Chappell, and L. Ro. Hoxit, 1979: Synoptic and meso-alpha scale aspects of flash flood events. *Bull. Amer. Meteor. Soc.,* **60,** 115-123, https://dx.doi.org/10.1175/1520-0477-60.2.115.

Martinaitis, S. M., S. B. Cocks, A. P. Osborne, M. J. Simpson, L. Tang, J. Zhang, and K. W. Howard, 2021: The historic rainfall of Hurricane Harvey and Florence: A perspective from the multi-radar multi-sensor system. *J. Hydrometeor.,* **22,** 721-738, https://doi.org/10.1175/JHM-D-20-0199.1.

Martinaitis, S. M., K. A. Wilson, N. Yussouf, J. J. Gourley, H. Vergara, T. C. Meyer, P. L. Heinselman, A. Gerard, K. L. Berry, A. Vergara, and J. Monroe, 2022: A path towards short-term probabilistic flash flood prediction. *Bull. Amer. Meteor. Soc.,* **104,** E585-E605, https://doi.org/10.1175/BAMS-D-22-0026.1.

Marty, R., I. Zin, C. Obled, G. Bontron, and A. Djerboua, 2012: Toward real-time daily PQPF by an analog sorting approach: Application to flash-flood catchments. *J. Appl. Meteor. Cimatology,* **51,** 505-520, https://doi.org/10.1175/JAMC-D-11-011.1.

McGovern, A., R. Lagerquist, D. J. Gagne II, G. Eli Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.,* **100,** 2175-2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne II, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artificial Intelligence for the Earth Systems.* **2,** 1-24, https://doi.org/10.1175/AIES-D-22-0077.1.

NCEI, 2020: Billion-dollar weather and climate disasters (2023). https://www.ncei.noaa.gov/access/billions/, accessed 21 Aug 2023.

National Disaster Management Authority of Pakistan, 2022: NMDA Floods (2023) Situation Report, 18 Nov 2022. https://cms.ndma.gov.pk/storage/app/public/situation-reports/November2022/N2n1eEarMt6q6Rb8ZYwn.pdf, accessed 21 Aug 2023.

NWS, 2017: Summary of natural hazard statistics in the United States. National Weather Service, Office of Climate, Weather, and Water Services, http://weather.gov/hazstat/.

NWS, 2023: Colorado Flood Safety and Wildfire Awareness Week: Flash floods. Accessed online, 7 Sep 2023, https://www.weather.gov/pub/FSWPW4flashfloodswednesday.

NOAA, 2022: 80-Year List of Severe Weather Fatalities. https://weather.gov/media/hazstat/80years_2022.pdf, accessed 22 Aug 2023.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparison. *Wea. Forecasting,* **31,** 371-394, https://doi.org/10.1175/WAF-D-14-00112.1.

Nielsen, E. R., and R. S. Schumacher, 2018: Dynamical insights into extreme short-term precipitation associated with supercells and mesovortices. *J. Atmos. Sci.,* **75,** 2983-3009, https://doi.org/10.1175/JAS-D-17-0385.1.

Nielsen, E. R., and R. S. Schumacher, 2020: Dynamical mechanisms supporting extreme rainfall accumulations in the Houston "Tax Day" 2016 flood. *Mon. Wea. Rev.,* **148,** 83-109, https://doi.org/10.1175/MWR-D-19-0206.1.

Novak, D. R., 2023: The NOAA Precipitation Prediction Grand Challenge: Advancing predictions of extreme rainfall. 32nd Conference on Weather Analysis and Forecasting, 19 Jul 2023, Madison, WI.

Osborne, A. P., J. Zhang, M. J. Simpson, K. W. Howard, and S. B. Cocks, 2023: Application of machine learning techniques to improve multi-radar multi-sensor (MRMS) precipitation estimates in the western United States. *Artificial Intelligence for the Earth Systems.* **2,** e220053, https://doi.org/10.1175/AIES-D-22-0053.1.

Perica, S., S. Pavlovic, M. St. Laurent, C. Trypaluk, D. Unruh, and O. Wilhite, 2018: Version 2.0: Texas. Vol. 11, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 283 pp., https://www.weather.gov/media/owp/oh/hdsc/docs/Atlas14_Volume11.pdf.

Perica, S., S. Pavlovic, M. St. Laurent, C. Trypaluk, D. Unruh, D. Martin, and O. Wilhite, 2015: Version 2.0: Northeastern States (Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont). Vol. 10, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 265 pp., https://www.weather.gov/media/owp/oh/hdsc/docs/Atlas14_Volume10.pdf.

Perica, S., D. Martin, S. Pavlovic, I. Roy, M. St. Laurent, C. Trypaluk, D. Unruh, M. Yekta, and G. Bonnin, 2013a: Version 2.0: Midwestern States (Colorado, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Oklahoma, South Dakota, Wisconsin). Vol. 8, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 289 pp., https://www.weather.gov/media/owp/oh/hdsc/docs/Atlas14_Volume8.pdf.

Perica, S., D. Martin, S. Pavlovic, I. Roy, M. St. Laurent, C. Trypaluk, D. Unruh, M. Yekta, and G. Bonnin, 2013b: Version 2.0: Southeastern States (Alabama, Arkansas, Florida, Georgia,

Louisiana, Mississippi). Vol. 9, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 163 pp., https://www.weather.gov/media/owp/oh/hdsc/docs/Atlas14_Volume9.pdf.

Perica, S., and Coauthors, 2011: Version 2.0: California. Vol. 6, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 233 pp., https://www.weather.gov/media/owp/oh/hdsc/docs/Atlas14_Volume6.pdf.

Pielke, R. A., Jr., and M. W. Downton, 2002: Precipitation and damaging floods: Trends in the United States, 1932-97. *J. Climate,* **13,** 3625-3637, https://doi.org/10.1175/1520-0442(2000)013<3625:PADFTI>2.0.CO;2.

Potvin, C. K., C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian hierarchical modeling framework for correcting reporting bias in the U.S. tornado database. *Wea. Forecasting,* **34,** 15-30, https://doi.org/10.1175/WAF-D-18-0137.1.

Prein, A. F., C. Liu, K. Ikeda, S. B. Trier, R. M. Rasmussen, G. J. Holland, and M. P. Clark, 2017: Increased rainfall volume from future convective storms in the US. *Nature Climate Change Letters,* **7,** 880-884, https://doi.org/10.1038/s41558-017-0007-7.

Qi., Y., S. Martinaitis, J. Zhang, and S. Cocks, 2016: A real-time automated quality control of hourly rain gauge data based on multiple sensors in MRMS system. *J. Hydrometeor.,* **17,** 1675-1691, https://doi.org/10.1175/JHM-D-15-0188.1.

Radhakrishnan, C., and V. Chandrasekar, 2020: CASA prediction system over Dallas – Fort Worth urban network: Blending of nowcasting and high-resolution numerical weather prediction model. *J. Atmos. Oceanic Technol.,* **37,** 211-228, https://doi.org/10.1175/JTECH-D-18-0192.1.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting,* **35,** 2293-2316, https://doi.org/10.1175/WAF-D-20-0069.1.

Roebber, P., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting,* **24,** 601-608, https://doi.org/10.1175/2008WAF2222159.1.

Saabas, A., 2016: Random forest interpretation with scikit-learn. Accessed 18 Sep 2023, https://blog.datadive.net/random-forest-interpretation-with-scikit-learn/.

Schumacher, R. S., and G. R. Herman, 2021: Reply to "Comments on 'Flash flood verification: Pondering precipitation proxies.'" *J. Hydrometeor.,* **22,** 749-752, https://doi.org/10.1175/JHM-D-20-0275.1.

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

Sharif, H. O., D. Yates, R. Roberts, and C. Mueller, 2006: The use of an automated nowcasting system to forecast flash floods in an urban watershed. *J. Hydromet.,* **7,** 190-202, https://doi.org/10.1175/JHM482.1.

Smith, J. A., M. L. Baeck, Y. Su, M. Lui, and G. A. Vecchi, 2023: Strange storms: Rainfall extremes from the remnants of Hurricane Ida (2021) in the northeastern US. *Water Resources Res.,* **59,** e2022WR033934, https://doi.org/10.1029/2022WR033934.

Sobash, R. A., G. S. Romin, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting,* **35,** 1981-2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 211: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting,* **26,** 714-728, https://doi.org/10.1175/WAF-D-10-05046.1.

Stensrud, D. J., and coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.,* **90,** 1487-1499, https://doi.org/10.1175/2009BAMS2795.1.

Stensrud, D. J., and coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.,* **123,** 2-16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Sun, J., Y. Zhang, J. Ban, J.-S. Hong, and C.-Y. Lin, 2020: Impact of combined assimilation of radar and rainfall data on short-term heavy rainfall prediction: A case study. *Mon. Wea. Rev.,* **148,** 2211-2232, https://doi.org/10.1175/MWR-D-19-0337.1.

Sweeney, T. L., 1992: Modernized areal flash flood guidance. NOAA Tech. Memo NWS HYDRO 44, 37 pp., https://repository.noaa.gov/view/noaa/13498.

Szoke, E. J., S. G. Benjamin, C. R. Alexander, E. P. James, J. M. Brown, D .T. Lindsey, and B. D. Jamison, 2015: HRRR model performance for the September 2013 northestern Colorado floods. *29th Conf. on Hydrology,* Phoenix, AZ, Amer. Meteor. Soc., 6.2.

Szoke, E. J., S. G. Benjamin, C. R. Alexander, J. M. Brown, T. I. Alcott, and E. P. James, 2018: Examination of the current and next version of the HRRR model for some recent heavy precipitation events. *32nd Conf. on Hydrology,* Austin, TX, Amer. Meteor. Soc., 6.6.

Tang, L., J. Zhang, M. Simpson, A. Arthur, H. Grams, Y. Wang, and C. Langston, 2020: Updates on the radar data quality control in the MRMS quantitative precipitation estimation system. *J. Atmos. Oceanic Technol.,* **37,** 1521-1537, https://doi.org/10.1175/JTECH-D-19-0165.1.

Trojniak, S., and J. Correia, Jr., 2021: 2021 Flash Flood and Intense Rainfall Experiment: Findings and results. WPC report. Available online at https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2021_FFaIR_Experiment.pdf

Trojniak, S., and J. Correia, Jr., 2022: 2022 Flash Flood and Intense Rainfall Experiment Final Report: Results and findings. WPC report. Available online at https://wpc.ncep.noaa.gov/hmt/2022_FFaIR_Final_Report.pdf.

Trojniak, S., and J. Corrieia, Jr., 2023: 2023 Flash Flood and Intense Rainfall Experiment Operations Plan. WPC report. Available online at https://wpc.ncep.noaa.gov/hmt/hmt_webpages/2023_FFaIR_Operations_Plan.pdf.

Trojniak, S., J. Correia, Jr., and B. Albright, 2020: 2020 Flash Flood and Intense Rainfall Experiment: Findings and results. WPC report. Available online at https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.,* **125,** 527-548, https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

Weygandt, S. S., S. S. Benjamin, M. Hu, C. R. Alexander, T. G. Smirnova, and E. P. James, 2022: Radar reflectivity-based model initialization using specified latent heating (radar-LHI) within a diabatic digital filter or pre-forecast integration. *Wea. Forecasting,* **37,** 1419-1434, https://doi.og/10.1175/WAF-D-21-0142.1.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.,* **95,** 51-70, https://doi.org/10.1007/s10994-013-5346-7.

Williamson, M., K. Ash, M. J. Erickson, and E. Mullens, 2023: Damages associated with excessive rainfall outlooks (ERO) and missed flash floods. *Wea. Forecasting,* **38,** 971-984, https://doi.org/10.1175/WAF-D-22-0035.1.

Wix, J. M, 2023: Eastern Kentucky flooding overview. Flash Flood and Intense Rainfall (FFaIR) Experiment seminar, 8 Jun 2023. https://wpc.ncep.noaa.gov/hmt/hmt_webpages/seminars/2023/June82023_2022_East_Kentucky_Flooding_Overview.pdf.

Yussouf, N., and K. H. Knopfmeier, 2019: Application of the Warn-on-Forecast system for flash-flood-producing heavy convective rainfall events. *Q. J. Royal Met. Soc.,* **145,** 2385-2403, https://doi.org/10.1002/qj.3568.

Yussouf, N., J. S. Kain, and A. Clark, 2016: Short-term probabilistic forecasts of the 31 May 2013 Oklahoma tornado and flash flood event using a continuous-update-cycle storm-scale ensemble system. *Wea. Forecasting,* **31,** 957-983, https://doi.org/10.1175/WAF-D-15-0160.1.

Yussouf, N., K. A. Wilson, S. M. Martinaitis, H. Vergara, P. Heinselman, and J. J. Gourley, 2020: The coupling of NSSL Warn-on-Forecast and FLASH systems for probabilistic flash flood prediction. *J. Hydrometeor.,* **21,** 123-141, https://doi.org/10.1175/JHM-D-19-0131.1.

Zhang, J., L. Tang, S. Cocks, P. Zhang, A. Ryzhkov, K. Howard, C. Langston, and B. Kaney, 2020: A dual-polarization radar synthetic QPE for operations. *J. Hydrometeor.,* **21,** 2507-2521, https://doi.org/10.1175/JHM-D-19-0194.1.

Zhang, J., K. Howard, C. Langston, S. Vasiloff, B. Kaney, A. Arthur, S. Van Cooten, K. Kelleher, D. Kitzmiller, F. Ding, D.-J. Seo, E. Wells, and C. Dempsey, 2011: National Mosaic and multi-sensor QPE (NMQ) system. *Bull. Amer. Meteor. Soc.,* **92,** 1321-1338, https://doi.org/10.1175/2011BAMS-D-11-00047.1.

Zhang, J., K. Howard, C. Langston, B. Kaney, Y. Qi, L. Tang, H. Grams, Y. Wang, S. Cocks, S. Martinaitis, A. Arthur, K. Cooper, J. Brogden, and D. Kitzmiller, 2016: Multi-Radar Multi-

Sensor (MRMS) quantitative precipitation estimation. *Bull. Amer. Meteor. Soc.,* **97,** 612-638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zhang, Y., M. Long, K. Chen, L. Xing, R. Jin, M. I. Jordan, and J. Wang, 2023: Skilful nowcasting of extreme precipitation with NowcastNet. *Nature,* **619,** 526-532, https://doi.org/10.1038/s41586-023-06184-4.