Forecasting Excessive Rainfall with Random Forests and a Deterministic Convection-Allowing Model

AARON J. HILL[a] AND RUSS S. SCHUMACHER[a]

[a] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*

(Manuscript received 24 February 2021, in final form 7 July 2021)

ABSTRACT: Approximately seven years of daily initializations from the convection-allowing National Severe Storms Laboratory Weather Research and Forecasting Model are used as inputs to train random forest (RF) machine learning models to probabilistically predict instances of excessive rainfall. Unlike other hazards, excessive rainfall does not have an accepted definition, so multiple definitions of excessive rainfall and flash flooding—including flash flood reports and 24-h average recurrence intervals (ARIs)—are used to explore RF configuration forecast sensitivities. RF forecasts are analogous to operational Weather Prediction Center (WPC) day-1 Excessive Rainfall Outlooks (EROs) and their resolution, reliability, and skill are strongly influenced by rainfall definitions and how inputs are assembled for training. Models trained with 1-yr ARI exceedances defined by the Stage-IV (ST4) precipitation analysis perform poorly in the northern Great Plains and Southwest United States, in part due to a high bias in the number of training events in these regions. Increasing the ARI threshold to 2 years or removing ST4 data from training, optimizing forecast skill geographically, and spatially averaging meteorological inputs for training generally results in improved CONUS-wide RF forecast skill. Both EROs and RF forecasts have seasonal skill——poor forecasts in the late fall and winter and skillful forecasts in the summer and early fall. However, the EROs are consistently and significantly better than their RF counterparts, regardless of RF configuration, particularly in the summer months. The results suggest careful consideration should be made when developing ML-based probabilistic precipitation forecasts with convection-allowing model inputs, and further development is necessary to consider these forecast products for operational implementation.

SIGNIFICANCE STATEMENT: Machine learning (ML) models can deduce statistical relationships between a set of predictors and meteorological events. In this work, ML models are developed to predict excessive rainfall events. Since excessive rainfall is difficult to uniformly define across the United States, multiple ML models are built from a variety of rainfall datasets with predictors gathered from output of a high-resolution numerical weather prediction model and forecasts are made from each model. Forecasts made from these models are highly sensitive to both the definitions of excessive rainfall (e.g., 100 mm of rain in a day may cause flooding in a usually dry area, but not in a wet area) and the predictors used. Forecast skill can increase when excessive rainfall events are rarer and when predictors synthesize the surrounding environment rather than characterize specific geographical points. ML-based models have great potential for excessive rainfall prediction, but careful attention to the configuration of these models is required.

KEYWORDS: Rainfall; Numerical weather prediction/forecasting; Operational forecasting; Machine learning; Decision trees; Decision trees

1. Introduction

Flash flooding and excessive rainfall pose significant risks to society, combining for over $60 billion (U.S. dollars) in damage and 212 deaths in the United States from 2010 to 2019 (NCEI 2020), and contributing to significant increases in property damage over the last two decades (Ahmadalipour and Moradkhani 2019). To provide timely forecasts that alert the general public to the threat of excessive rainfall, the Weather Prediction Center (WPC) issues excessive rainfall outlooks (EROs; NOAA/Weather Prediction Center 2021) at approximately 0900 and 2100 UTC daily that cover the 3–27-h (day 1), 27–51-h (day 2), and 51–75-h (day 3) periods across the contiguous United States (CONUS). Forecasting excessive rainfall and accompanying flooding events operationally is made difficult by numerical weather prediction (NWP) model guidance, which must accurately depict the location, intensity, and duration of rainfall events along with accompanying environmental ingredients (e.g., Doswell et al. 1996; Schumacher 2017), that often exhibit substantial quantitative precipitation forecast (QPF) biases (Herman and Schumacher 2016). Whereas forecasters have a plethora of observational data available to aid their day-1 EROs, day-2 and day-3 EROs are primarily dependent on NWP model guidance and previously issued outlooks (Novak et al. 2014; Erickson et al. 2021). Despite recent enhancements to NWP models that have improved precipitation forecasts, including the reduction of horizontal grid spacing to convection-allowing resolutions (e.g., Done et al. 2004; Clark et al. 2007, 2009; Ikeda et al. 2013), the important processes responsible for long- and short-duration precipitation events that lead to flooding are still unresolvable in convection-allowing NWP models (CAMs). CAMs are not immune to QPF biases either (e.g., Romine et al. 2013; Herman and Schumacher 2016; Wong et al. 2020), and recent work has suggested that NWP model horizontal grid spacing must be reduced (to 1 km) to further improve precipitation forecasts (Schwartz and Sobash 2019).

Post-processing methods applied to model output, including machine learning (ML), have shown promise in producing calibrated, probabilistic forecasts of extreme precipitation

Corresponding author: Aaron J. Hill, aaron.hill@colostate.edu

DOI: 10.1175/WAF-D-21-0026.1

(e.g., Gagne et al. 2014; Scheuerer and Hamill 2015; Herman and Schumacher 2018b; Whan and Schmeits 2018; Loken et al. 2019) and other convection-based hazards like tornadoes, hail, and severe wind (e.g., Gagne et al. 2017; McGovern et al. 2017; Burke et al. 2020; Hill et al. 2020; Loken et al. 2020; Sobash et al. 2020; Flora et al. 2021). These postprocessing examples have used NWP model forecasts as inputs to random forests (RFs; Breiman 2001) and artificial neural networks to produce both quantitative and probabilistic event-based forecasts of weather hazards (e.g., Herman and Schumacher 2018b; Loken et al. 2019). As an example, Herman and Schumacher (2018b) leveraged the National Oceanic and Atmospheric Administration (NOAA) Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) dataset to train RFs to probabilistically predict the occurrence of excessive rainfall events over 24-h periods analogous to day-2 and day-3 WPC EROs. Their statistical model forecasts demonstrated improved skill over raw probabilistic QPFs from global operational models and have been successfully transitioned to operations at the WPC (Schumacher et al. 2021) as "first-guess" guidance fields for operational forecasters. Unfortunately, whereas the GEFS-driven RF forecasts are more skillful than humans on days 2 and 3, they are unable to routinely outperform the human WPC forecasters on day 1 (Schumacher et al. 2021). This study seeks to advance ML-based day-1 forecasts of excessive rainfall by exploring deterministic CAM-based RF models for extreme precipitation forecasting and examining sensitivities of RF configurations in order to determine their benefit to operational forecasters in guiding their day-1 EROs and where improvements can be made in the future to operationalize CAM-based RF products.

CAMs are a viable alternative to global models for ML-based precipitation prediction because they offer higher resolution spatiotemporal information, and they effectively characterize QPF climatologies (Herman and Schumacher 2016; Goines and Kennedy 2018), which suggests they can more accurately simulate environments supportive of excessive rainfall compared to coarser global models. On the other hand, the application of CAMs with ML comes with two primary limitations: 1) CAMs have limited temporal range, often restricted to less than 60 forecast hours due to computational constraints; and 2) CAMs often undergo development upgrades routinely, which alter their biases. The latter limitation is important because ML models are effective at learning input biases—e.g., high temperature bias during the daytime—when the input biases remain static (Loken et al. 2019). These limitations have likely hampered the development of CAM-based ML models for hazard forecasting. A recently study conducted by Loken et al. (2019) used high-resolution CAM ensemble forecasts as inputs to RF models and forecast accumulated precipitation over forecast hours 12–36, effectively calibrating the CAM ensemble QPFs. They showed that the RFs were able to correct systemic biases in QPF, and the RF forecasts outperformed the dynamic model QPF guidance.

However, the fixed threshold precipitation accumulation used by Loken et al. (2019) may not sufficiently characterize the threat of flash flooding locally, which can be heavily influenced by topography, antecedent conditions (e.g., Brocca et al. 2008), land type and use (e.g., Ogden et al. 2000), and urban impediments (e.g., Smith et al. 2005). Rainfall accumulations are also hard to measure and observe in some locales, as radar estimates are often relied on, which have inherent range biases in complex terrain (Nelson et al. 2016) and sensitivities to attenuation and mixed-phase precipitation regimes (Zhang et al. 2020). The difficulty in accurately observing rainfall events feeds into proper ML-model training and verification procedures as well (e.g., Schumacher et al. 2021). Schumacher et al. (2021) used average recurrence intervals (ARIs) with RF models and noted that regional excessive rainfall predictions could be calibrated by using slightly different definitions of excessive rainfall – unique to the forecast region—and different quantitative precipitation estimation (QPE) datasets to define ARI exceedances (e.g., Stage-IV), which have known regional biases (Herman and Schumacher 2018a).

CAM-based RF models are explored in this work, utilizing the experimental and nonoperational National Severe Storms Laboratory Weather Research and Forecast (NSSL WRF) model and locally varying climatologies of excessive rainfall (i.e., ARIs). The NSSL WRF used in this work is a deterministic, 4-km horizontal grid spacing CAM that produces forecast output over the CONUS with 36 h of lead time.[1] The primary advantage of the NSSL WRF, apart from the convection-allowing resolution, is the nearly static configuration for over a decade, which provides a long training dataset to develop robust ML models. The ML models and analysis methods are discussed in section 2. Operational forecast skill is assessed in section 3. ML-model forecast sensitivities to excessive rainfall definitions and predictor assembly, along with an example forecast, are presented in section 4. Concluding remarks and a summary of the results is reserved for section 5.

2. Data and methods

In this work, RFs are trained to forecast excessive rainfall as a binary event using different sets of features and labels. An important consideration in this work is how to define "excessive," as no widely accepted definition exists. The definition of a flood, flash flood, or extreme rainfall event varies considerably across the CONUS (Gourley et al. 2013; Gourley and Vergara 2021; Schumacher and Herman 2021), and NOAA National Weather Service Weather Forecast Offices use different thresholds to determine when to issue flash flood warnings, flood watches, as well as record local storm reports (LSRs; Gourley et al. 2013; Marjerison et al. 2016). Flash flood guidance (FFG; Sweeney 1992) is used as a proxy for flash flooding to define the WPC ERO forecasts (NOAA/Weather Prediction Center 2021), but the computed values of FFG can produce sharp discontinuities across NWS River Forecast Center boundaries (e.g., Clark et al. 2014). An alternative proxy for excessive rainfall is frequency-based thresholds such

[1] An alternative NSSL WRF model runs out to 60 h with 3-km horizontal grid spacing. However, that 3-km experimental model was developed for comparison to other convection-allowing models under development in late 2018, and thus, does not have a long enough record desired for this study.

as ARIs, which are used in this work, but specific ARIs and accumulation periods that best correspond to flash flooding vary regionally across the CONUS and across datasets (Herman and Schumacher 2018a; Gourley and Vergara 2021; Schumacher and Herman 2021). These considerations and the lack of a formal definition motivate the use of more than one dataset to define excessive rainfall.

Excessive rainfall events in this work are encoded using a combination of up-to three datasets: 1) flash flood reports (FFRs), 2) NCEP Stage-IV precipitation analysis over a 24-h accumulation period (ST4; Nelson et al. 2016), and 3) 24-h climatology calibrated precipitation accumulation (CCPA; Hou et al. 2014). ST4 QPE analyses are generated from rain gauge observations and radar-derived rainfall estimates, whereas CCPA QPE fields are derived from ST4 QPE and the Climate Prediction Center's (CPC) unified global daily gauge analysis. Effectively, the CCPA dataset represents a gauge-corrected ST4 QPE; the CPC-based analysis is generally considered more accurate due to more rigorous quality control. Differences between the two QPE datasets are covered extensively by Herman and Schumacher (2018a). ST4 and CCPA QPE accumulation grids are used to define instances of 1- and 2-yr 24-h ARI exceedances. The specific dataset combinations used in training for sensitivity experiments are discussed in section 2b. The trained RF models produce probabilistic forecasts analogous to day-1 WPC ERO outlooks, that is, the probability of an excessive rainfall event within 40 km of a point over the 24-h 1200–1200 UTC period. WPC EROs are compared to RF forecasts over the same verification period (discussed in section 2a), and all discussion of EROs is restricted to the day-1 products.

a. *Random forests*

RFs constructed herein are made up of decision trees that individually make unique classification predictions (e.g., 0 or 1) of a particular event based on features (i.e., inputs) to the tree. Trees are trained by considering a historical set of labels (e.g., rainfall events) and corresponding features, and each tree considers a random subsample of training examples to generate unique trees. Beginning with the root node of a tree, branches are traversed based on the outcomes of criteria specified at each node. A random subset of features is evaluated at each node to select a criterion that minimizes node impurity for all remaining training examples. In other words, subsequent nodes in the tree, ideally, become more pure and aligned with a specific classification label (e.g., flood or no flood). Once the subset of events in a node becomes pure or is no longer large enough to split, a "leaf" node is produced, which makes a declaration of the event classification. To make a forecast, new inputs—e.g., from real-time NWP model output—are supplied to the tree, and the tree is traversed to a leaf node. All decision-tree predictions are then aggregated to produce a probabilistic forecast of excessive rainfall based on the inputs provided.

To build robust statistical models in this excessive rainfall context, it is important to employ a long training period with a nearly static NWP model for inputs. ML models are capable at learning robust statistical relationships when input biases remain consistent, and it is typical for CAMs to receive regular upgrades every year or two, which may alter their forecast biases. Therefore, inputs are gathered from the 4-km grid spacing National Severe Storms Laboratory (NSSL) Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008) CAM (NSSL WRF), which has remained nearly static since 2009. Example meteorological input variables include accumulated precipitation, convective available potential energy, precipitable water, and 2-m temperature (see Table 1 for a list of all variables included), which are presumed to have meaningful relationships with precipitation that can be deduced by the random forests.

TABLE 1. Meteorological predictors input to the RF models during training and forecasting.

Symbol	Variable description
APCP	3-hourly accumulated precipitation
CAPE	Convective available potential energy
CIN	Convective inhibition
PWAT	Precipitable water
MSLP	Mean sea level pressure
U10	10-m latitudinal horizontal wind speed
V10	10-m longitudinal horizontal wind speed
T2M	2-m temperature
Q2M	2-m specific humidity
UPHL	2–5-km updraft helicity
Z500	500-hPa geopotential height
U6000	0–6-km average latitudinal horizontal wind speed
V6000	0–6-km average longitudinal horizontal wind speed
W3000	0–3-km average vertical wind speed

Inputs are assembled in a forecast-point perspective by gathering all p variables from a radius of n nearby NSSL WRF grid points in both the longitudinal and latitudinal directions (e.g., Herman and Schumacher 2018b; Hill et al. 2020); variables are spaced by 48 km (12 grid points) and up to 240 km from the forecast point ($n = 5$). Temporally, inputs are selected every three hours over 1200 to 1200 UTC (forecast hours 12–36) from the 0000 UTC initialized NSSL WRF forecast, corresponding to the forecast period of interest. The number of predictors per training label can be expressed mathematically as $N = pt(2n + 1)^2$, where t is the number of forecast times ($t = 9$ in this work). The prescribed feature assembly amounts to 1089 predictors per variable p, and $N = 15\,246$ predictors per training example. Additionally, static inputs related to the climatological 1- and 10-yr ARIs near a forecast point are included, along with latitude, longitude, day, and a seasonality estimate; these static predictors are identical to those described by Herman and Schumacher (2018b). The predictor assembly process is fundamentally the same as that used by Herman and Schumacher (2018b) and Hill et al. (2020), is motivated by a number of sensitivity tests conducted by Herman and Schumacher (2018b), and is particularly useful for capturing any spatiotemporal biases in the underlying NWP model used for predictors. Furthermore, the spatiotemporal predictor assembly (e.g., $n = 5$) was chosen to closely match that of models developed by Schumacher et al. (2021) for subjective comparisons. It is also important to note that training points (i.e., labels) are defined on a coarser, half-degree grid, such that predictors are gathered

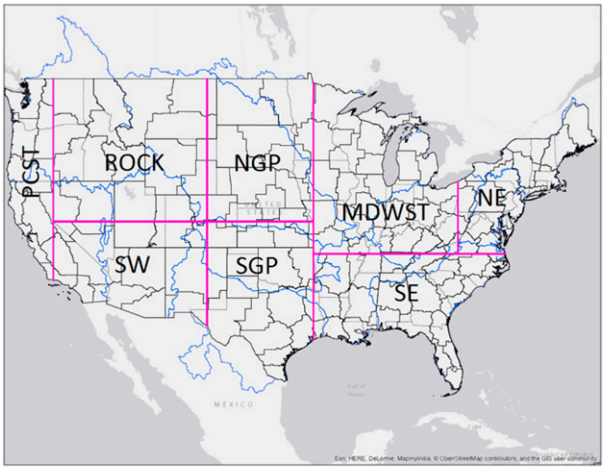
PCST
ROCK
NGP
MDWST
NE
SW
SGP
SE
MEXICO
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

FIG. 1. Geographical regions over which RF models are trained and forecasts are issued.

relative to the closest NSSL WRF grid point; this coarsening is necessitated by computational constraints.

RFs are trained with excessive rainfall labels and CAM-based features over an approximately 7-yr training period from 9 June 2009 to 31 August 2016 using daily initializations of the NSSL WRF for inputs (2290 days in total); occasional days within this period are omitted from training as NSSL WRF data were not available. Regionally varying climatologies of excessive rainfall suggest that the RFs should be trained separately across the CONUS (Fig. 1). Only one RF configuration parameter (i.e., minimum number of examples required to split a node) was varied across each regionally trained model (Table 2); the number of decision trees was set to 1000, the maximum number of predictors evaluated at each node was the $\sqrt{N}$, and entropy was used as the splitting criterion. The values prescribed to each parameter were largely determined from prior work (e.g., Herman and Schumacher 2018b) and subjective comparisons, and no formal tuning procedure was applied. For the purposes of this work, the static RF parameters should not be seen as a detriment as the focus is on how to employ CAMs with ML models to forecast excessive rainfall in a way that would be useful for operational forecasters. Practically, a validation and testing procedure would involve tuning each RF against the training dataset, which is not used for verification purposes. Additional tests (not shown) verify that optimally tuning the RF parameters had negligible impact on forecast skill. Each of the parameters is passed to the Python Scikit-learn RandomForestClassifier package (Pedregosa et al. 2011) to train the RFs. As a result of the regionalization, forecasts generated from the regional RFs are stitched together using a sigmoid function to reduce discontinuities between regional boundaries. A 2-yr evaluation period from 1 January 2017 to 31 December 2018 is used to assess model forecast performance and compute aggregate skill statistics, discussed in section 2c.

TABLE 2. RF regional model configuration parameters: number of decision trees is set to 1000, entropy is the splitting criterion, and maximum number of features at a node is set to $\sqrt{N}$.

Region	Min No. of samples
PCST	16
SW	30
ROCK	30
NGP	120
SGP	120
MDWST	120
SE	120
NE	120

b. Sensitivity experiments

Because excessive rainfall is poorly defined as discussed above, several experiments are designed to test the sensitivity of RF performance to different definitions of excessive rainfall. Four combinations of excessive rainfall datasets are used to train RFs in each geographic region: 1) FFR and 1-yr ARI exceedances from either CCPA or ST4 (FCS1); 2) FFR and 1-yr ARI exceedances from only CCPA (FC1); 3) FFR and 2-yr ARI exceedances from either CCPA or ST4 (FCS2); and 4) FFR and 2-yr ARI exceedances from only CCPA (FC2). Previous work has suggested ST4 has substantial biases in the U.S. Southwest, which helps motivate the construction of these four datasets (Schumacher et al. 2021). Forecasts generated from each regional model can then be stitched together across the CONUS and evaluated against the corresponding WPC ERO and an independent verification dataset (described in the following section); hereafter, model forecasts generated from these models are referenced by their labels (e.g., FCS1-trained model forecasts). The first set of model forecasts use consistent label definitions across all regions, e.g., FCS1 for all eight regions. The second forecast system (OPT) considers the regional skill of each separately trained RF, and the RFs that maximize regional skill contribute to a CONUS-wide forecast. In other words, OPT forecasts are made from regionally skillful RFs.

The previously discussed RF forecast systems specifically use raw grid points as predictors as described in section 2a (hereafter denoted as RAW). Previous work has demonstrated skillful ML forecasts by using spatially averaged meteorological predictors (e.g., Loken et al. 2020; Sobash et al. 2020). To explore an optimal procedure for gathering predictors from CAM-based output, a third forecast system (OPT_AVG) spatially averages the $(2n + 1)^2$ predictors for each p variable and time t ($N = pt$); hereafter, these model forecasts are referred to as SPT-predictor forecasts. The same iterative regional skill optimization procedure is conducted with OPT_AVG, i.e., forecast skill is computed for the four separately trained models (FCS1, FC1, FCS2, and FC2 with SPT predictors) and the most skillful regional models are retained for CONUS-wide forecasts. The regional label datasets used by OPT and OPT_AVG are presented in Table 3. Because the trained RFs use the same, static hyperparameters, it is reasonably assured that any sensitivities in forecasts and skill will be due to training differences in either the label datasets or predictor assemblies.

c. Analysis metrics

An independent dataset of excessive rainfall occurrence is used to evaluate both ERO and RF-based forecast skill to

TABLE 3. Label datasets used in each region of the OPT and OPT_AVG models.

Region	OPT	OPT_AVG
PCST	FC2	FC2
ROCK	FCS2	FC2
SW	FC2	FC2
NGP	FC1	FC1
SGP	FC1	FC1
MDWST	FC1	FC1
SE	FCS1	FCS1
NE	FCS1	FC2

avoid favorably skewed results when verifying RF-based forecasts with the same dataset used to train the models. The WPC has developed the Unified Flood Verification System (UFVS; Erickson et al. 2019, 2021) to combat excessive rainfall reporting deficiencies from individual sources; the UFVS comprises flash flood reports, flooding reports from United States Geological Survey (USGS) stream gauges, and exceedances of the 5-yr ARI or FFG all valid within the 1200–1200 UTC period. The UFVS was not used to train RF models because it has a limited record, only dating back to 2016. A 40-km radius is applied around each observed event to match the neighborhood specified by the ERO and RF forecast products. The UFVS dataset is used to compute daily forecast skill through the Brier score (BS) as well as aggregate Brier skill score (BSS) statistics. UFVS observations are aggregated and then spatially smoothed over 1 October 2016–30 September 2020 to quantify the climatological occurrence of excessive rainfall across the CONUS (Fig. 2a) in order to calculate BSSs. The smoothing procedure is similar to that used by Schumacher et al. (2021) and Krocak and Brooks (2018) for severe reports (e.g., tornadoes, hail, and wind).

To appropriately compare the discrete WPC ERO probabilities and continuous RF forecast distribution, the RF-based probabilities are discretized to match the probability bins of the EROs. The EROs have categorical definitions corresponding to 5%–10% (marginal), 10%–20% (slight), 20%–50% (moderate), and >50% (high) probability bins. RF-based probabilities within these bins are mapped to the bin midpoints (7.5%, 15%, 35%, and 75%, respectively) to calculate skill scores; similarly, the WPC EROs are remapped to the midpoint probability values for quantitative skill comparisons. However, the full RF forecast distribution is used to construct reliability diagrams and formally evaluate RF forecast reliability and resolution against the UFVS observations. Forecast resolution is assessed via area under the relative operating characteristic curve (AuROC), which characterizes how well a forecast system discriminates between events and nonevents. Spatial reliability is also considered separately from BSSs by computing the area coverage of UFVS events in specific probability contours that correspond with the ERO categories (e.g., Erickson et al. 2019; Hill et al. 2020; Erickson et al. 2021; Schumacher et al. 2021).

3. Operational forecasting skill

The climatological observed frequency of excessive rainfall events from the UFVS (Fig. 2a) illustrates routine areas that experience excessive rainfall, including the southern Great Plains, southeast United States, and mid-Atlantic region. The Cascade mountain range in western Washington, the Sierra Mountains in California, and more generally the Rocky Mountains in Montana, Wyoming, Colorado, and New Mexico also experience frequent excessive rainfall. Notably, across the Pacific Northwest, four corners region of Utah, Colorado, Arizona, and New Mexico, and to some extent the Great Lakes region and upper northeast United States, excessive rainfall events are less common. Generally speaking, excessive rainfall during the verification period closely matches that of climatology (cf. Figs. 2a,b) with localized maxima due to a shorter period of record.

Operational ERO forecasts are issued by the WPC three times daily (0100, 0900, and 1600 UTC). Only the 0900 UTC issued forecasts are valid 1200–1200 UTC the following day, consistent with the RF-issued forecasts; therefore, for the purposes of this work, the 0900 UTC ERO forecast skill will be assessed. East of the Rocky Mountains, the WPC EROs exhibit considerable skill (i.e., BSS > 0), with a broad swath of positive BSSs overlapping climatological events extending from southern Texas to northern Wisconsin (Fig. 2c). Statistically significant skill is also present across the eastern Carolinas, southern Georgia, and the coast of California. Due to a limited verification period, poor skill (i.e., BSS < 0) exists in small, localized pockets (e.g., Florida peninsula), which are likely due to one or a few intense rainfall events that were missed by EROs and generated a number of excessive rainfall observations. The WPC EROs exhibit little to no skill across the western CONUS and Pacific Northwest, where few excessive rainfall events occur in the UFVS observations but forecasts were issued.

The WPC EROs feature up to four distinct categorical risk thresholds: marginal, slight, moderate, and high. Marginal contours were issued most frequently across the Appalachian mountains during the verification period (Fig. 3a), with nearly 15% of days featuring a probability contour in some locales. Slight categorical outlooks were issued far less frequently along the spine of the Appalachians (Fig. 3b), which generally also coincides with only slightly positive BSSs (Fig. 2c), indicating potentially low-predictability events or a lack of observations to verify the outlooks. In contrast, slight outlooks were issued frequently along the eastern slopes of the Appalachian range (Fig. 3b) coinciding with improved skill. Slight and moderate forecasts were issued most frequently across central Arkansas during this period as well (Figs. 3b,c), which coincides with high skill scores and a swath of observed events in central Arkansas (Fig. 2b). Qualitatively, the frequency of marginal and slight risk contours tends to map favorably to the WPC ERO skill scores, particularly from the Gulf Coast northward and northeastward.

The substantial skill of WPC EROs along the California coast is also accentuated within categorical risk frequencies, with a ribbon of marginal and slight risk contours being issued right along the coast line and Sierra Nevada mountain range (Figs. 3a,b). Limited climatological events in this area would suggest that EROs are being issued sparingly, corresponding to when rainfall events have occurred. Moderate risks issued west of the Great Plains are primarily reserved for mountain ranges

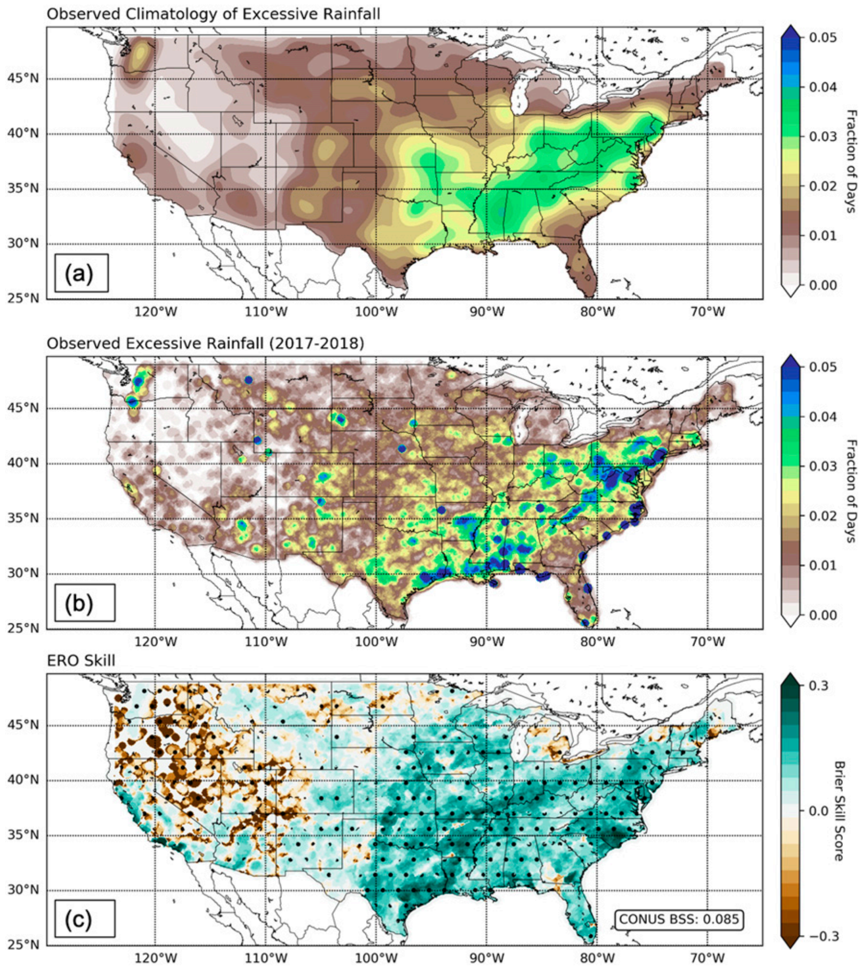

FIG. 2. (a) Fraction of days with excessive rainfall in the UFVS from 1 Oct 2016 to 30 Sep 2020. Frequencies have been spatially smoothed as discussed in the text. (b) Fraction of days with excessive rainfall in the UFVS over the 2017–18 verification period. (c) BSS of WPC EROs over the 2017–18 verification period. Stippling represents statistical significance at the 95% level obtained through bootstrap resampling of the approximately 2-yr forecast distribution 200 times. The value in the bottom-right corner of (c) is the CONUS-wide aggregate BSS.

in California and Arizona (Fig. 3c). There are noticeable hot spots of marginal and slight risk contours in New Mexico and Arizona, coincident with the U.S. monsoon region and mountainous terrain (Figs. 3a,b). However, observed events in this area are quite scattered (Fig. 2b), which necessarily degrades ERO skill (Fig. 2c) in localized areas. Coincident with the low climatological frequency of excessive rainfall events across the Pacific Northwest and western Rocky Mountains, few ERO contours of any magnitude are issued in the area (Fig. 3). ERO skill and categorical forecast frequencies highlight the skill of operational forecasters and provide a baseline for RF-based products in order to be operationally useful.

4. Results of sensitivity experiments

First, the frequency of events within FCS1, FCS2, FC1, and FC2 training datasets are examined. Both FCS1 and FCS2 feature numerous events within the Southwest (SW) region (Figs. 4a,b), which extend north and northwestward along the Rockies in FCS1 (Fig. 4a). The increased ARI threshold (i.e., changing to 2 years) in FCS2 results in substantially fewer rainfall events across the Great Plains and upper Rockies, and maintains a relative maximum across the mid-Atlantic (Fig. 4b). Similarities between FCS1 and FC1 in the intermountain west and Pacific Northwest suggest the higher frequencies of excessive rainfall in this region—compared to FCS2 and FC2—are the result of the 1-yr ARI exceedances from the CCPA. Additionally, comparing FCS1 with FC1 (cf. Figs. 4a,c), as well as FCS2 with FC2 (cf. Figs. 4b,d), it is evident that the high frequency of observations in the SW region is directly attributable to the ST4 dataset. The regional bias in New Mexico was similarly noted by Herman and Schumacher (2018a), and it was attributed to poor QPE estimation by ST4 in complex terrain. FC2 features the fewest number of

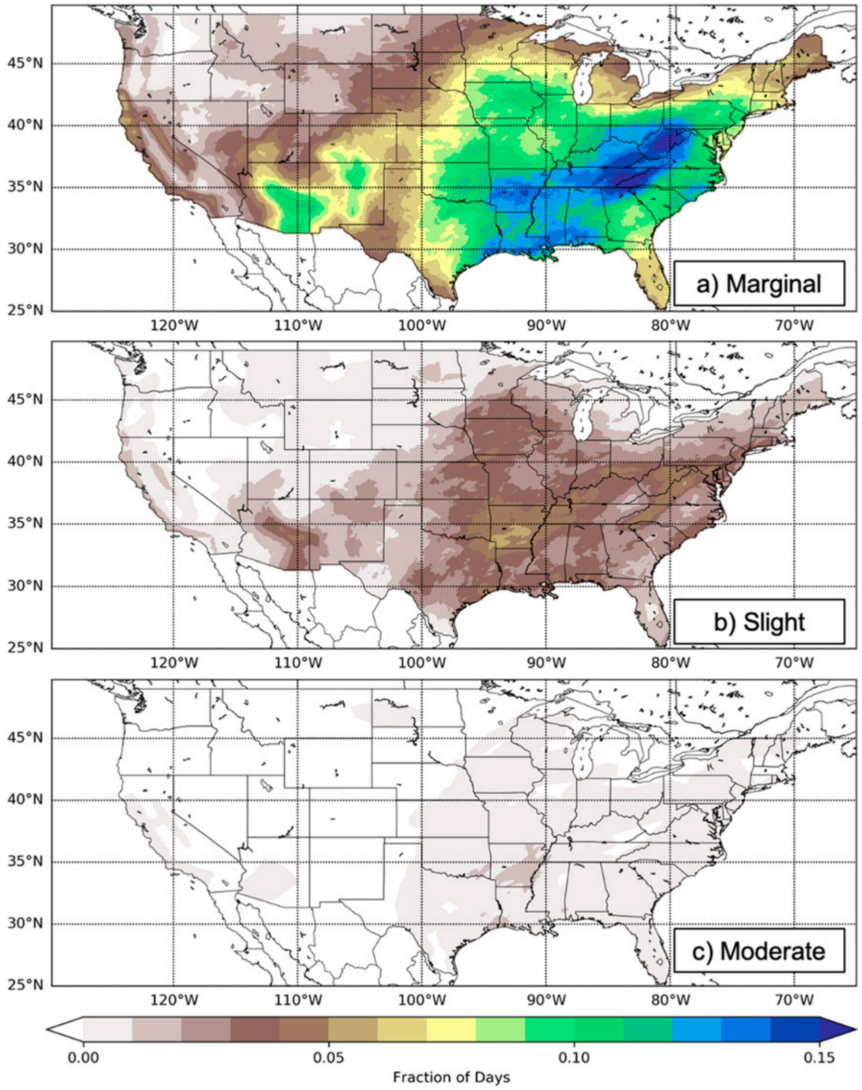

FIG. 3. Fraction of days within the verification period that featured at least a (a) marginal-, (b) slight-, and (c) moderate-risk categorical contour within the WPC EROs.

excessive rainfall events (Fig. 4d) due to removing ST4 and increasing the ARI threshold. The overall lower frequency of excessive rainfall events in FC1 and FC2 is partially attributable to the CCPA dataset, which can mute event frequencies through its linear calibration (Hou et al. 2014; Herman and Schumacher 2018a).

a. Model performance and forecast skill

BSs are calculated from daily forecasts made by the FCS1-, FCS2-, FC1-, and FC2-trained RAW models, which are then incorporated with climatological BSs to compute BSSs across the CONUS (Fig. 5). As mentioned previously, regional RF forecasts are stitched together from models trained with the same label dataset. Across all four systems, skill is worst across the western CONUS (Fig. 5), coinciding with the fewest excessive rainfall events (Figs. 2a, 4); the models do not learn critical statistical relationships when event sample sizes are small. BSSs are largely negative across the western CONUS, indicating RF-based forecasts are worse than a climatological forecast.

Regional skill differences across the forecasts are noteworthy along and east of the Rockies. FCS1-trained model forecasts perform statistically worse than climatology (i.e., skill is negative with 95% confidence based on bootstrap resampling) in the upper Great Plains (Fig. 5a), and forecast skill in this region substantially improves when ST4 labels are removed (Fig. 5c) and the ARI threshold is increased (Figs. 5b,d). This result indicates a potential high bias in the ST4 ARI exceedances and/or a better correspondence between longer recurrence intervals and the UFVS observations. Along the Rockies, where the ST4 dataset contributed to a higher frequency of training examples,

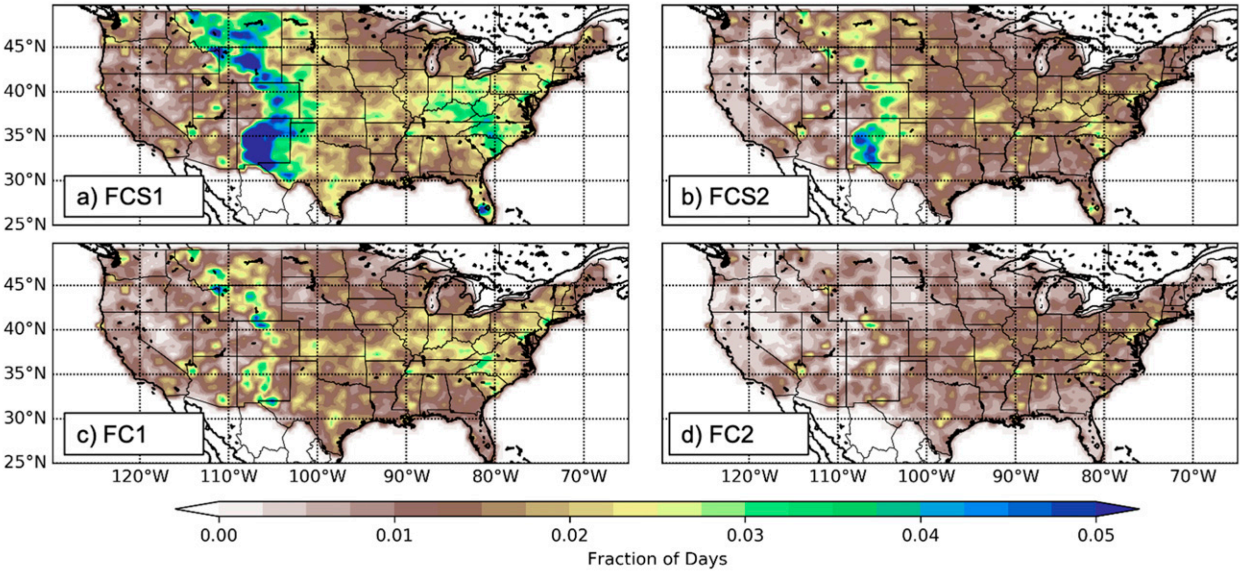

FIG. 4. Fraction of days with excessive rainfall events as defined by (a) FCS1, (b) FCS2, (c) FC1, and (d) FC2 datasets over the training period.

the FCS1- and FCS2-trained model forecasts have positive skill (Figs. 5a,b); skill decreases when models are not trained with ST4 (Figs. 5c,d). Because the UFVS includes ST4 5-yr ARI exceedances, it is plausible the positive skill seen in the complex terrain is artificial, given known ST4 QPE biases in these regions (e.g., Herman and Schumacher 2018a). Elsewhere, all forecasts exhibit positive skill across the southern Great Plains (SGP), Midwest (MDWST), Southeast (SE), and Northeast (NE) regions that is generally statistically significant; three exceptions include consistent areas of negative skill in western Texas, southern Michigan, and the Northern Great Plains.

BSSs are also calculated in a CONUS-wide perspective, i.e., aggregating all daily BSs from each grid point, which may alleviate some of the regional biases in ST4 or CCPA datasets (Herman and Schumacher 2018a). Removing ST4 1-yr ARI events from model training (FC1) increases forecast BSS from 0.014 to 0.036. Alternatively, increasing the ARI threshold to 2 years and retaining ST4 in training (FCS2) increases the CONUS-wide BSS to 0.033. Qualitatively, the FC2-trained

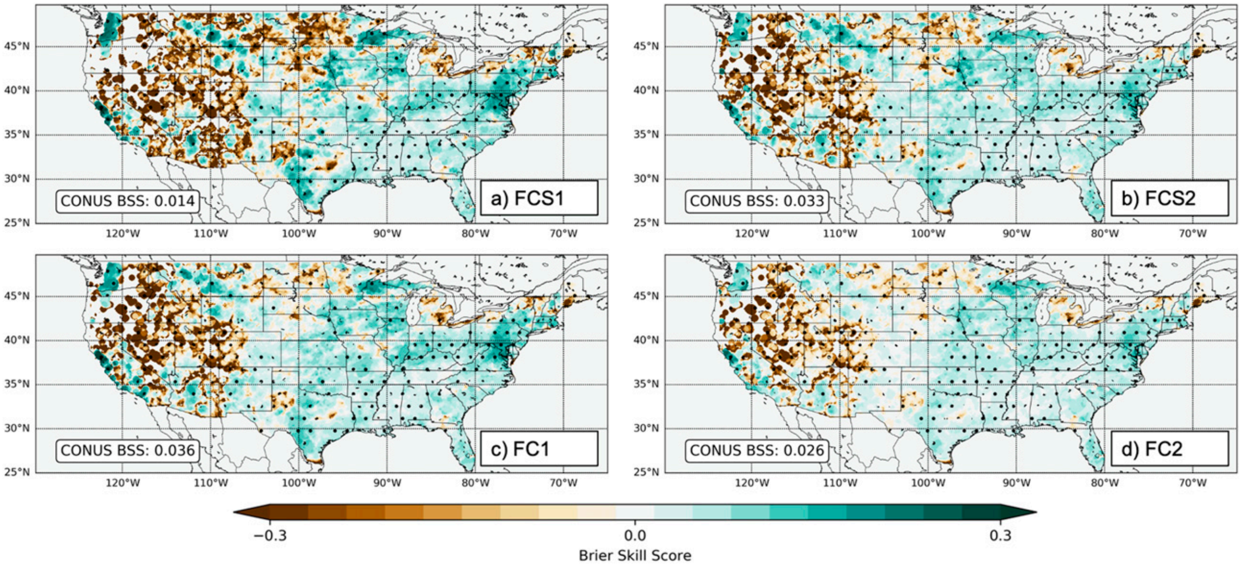

FIG. 5. BSS of RF-based model forecasts trained with (a) FCS1, (b) FCS2, (c) FC1, and (d) FC2 label datasets. Stippling represents statistical significance at the 95% level obtained through bootstrap resampling of the 2-yr forecast distribution 200 times. The value in the bottom-left corner of each panel is the CONUS-wide aggregate BSS of each model.

model clearly increases skill across the western regions [Pacific Coast (PCST), Rockies (ROCK), and SW], but also exhibits marginally poorer skill (but still positive) across the central and eastern United States; CONUS-wide BSS only increases to 0.026 from the baseline of 0.014. The CCPA dataset is known to curtail extreme events through the linear regression calibration procedure (Hou et al. 2014; Herman and Schumacher 2018a), which may reduce correspondence between FC2 observations and the UFVS resulting in less effective model training and poorer forecast skill. Spatially and qualitatively, the distribution of positive and negative skill is consistent with that of the EROs; less skill in the western United States and mostly positive skill across the central and eastern United States. However, ERO skill is still superior to the RF forecasts, qualitatively and in a CONUS-wide sense (0.085 BSS).

Regionally, EROs have mostly positive skill, with exceptions in PCST and ROCK; skill in SGP, MDWST, NE, and SE are markedly better than other regions (Fig. 6). RF-based forecasts have negative skill in the PCST, ROCK, and SW regions for all configurations, and negative skill in the NGP when trained with FCS1, consistent with the spatial BSSs assessed previously (Fig. 5a). In all geographic regions east of the Rockies, the best performing RF-based forecast system for a particular region has worse skill than the corresponding ERO. The FC2-trained RF model produces better forecasts than the EROs in PCST, whereas the FCS2-trained model has better skill in ROCK compared to the EROs. However, neither regional forecast is better than climatology in this instance. Interestingly, different label datasets favor specific geographical areas. For instance, FCS1-trained models have the best NE and SE regional forecast skill, FC1-trained models deliver the best northern Great Plains (NGP), SGP, and MDWST regional forecasts, consistent with results from Schumacher et al. (2021). Weaker connections exist in the west, where the best regional forecast in the PCST and SW regions comes from training with FC2, and the ROCK region benefits from increasing the ARI threshold but retaining ST4 in the FCS2 dataset. These results confirm the regional preferences of QPE datasets as discussed by Herman and Schumacher (2018a).

Regional forecast skill—in particular, areas of significant negative skill—can be partially attributed to the frequency of forecast issuance. The FCS1, FCS2, and FC1 RFs issue marginal categorical forecasts more frequently within the ROCK, SW, SGP, NGP, and MDWST regions compared to the EROs (cf. Figs. 7a,d,g and 3a). As a result, forecasts regionally are significantly worse (Fig. 6). Conversely, WPC EROs are issued more frequently than all RF forecasts in each outlook category across the SE (cf. Figs. 7 and 3), resulting in a BSS > 0.1 compared to the best RF-based BSS of <0.05 (Fig. 6). The statistical relationships learned by all NSSL WRF-based models in the SE do not sufficiently characterize the excessive rainfall threat in this region. These relatively high forecast frequencies are also prominent in the slight category for the FCS1, FCS2, and FC1-trained models (Figs. 7b,e,h) across the SW, NGP, and SGP regions; the far-too-frequent forecast issuance reduces skill. Spatial forecast improvements discussed previously can also be viewed through the forecast fraction lens, i.e., removing ST4 or increasing the ARI threshold significantly reduces forecast issuance in the problematic ROCK and SW regions (e.g., cf. Figs. 7b,e), which is reflected in improved spatial BSSs (Fig. 5). In other words, the high frequency of forecasts at lower probability thresholds (e.g., marginal and slight categories) contributes to poorer forecast skill, likely because low-probability forecasts are being issued when no observations exist (not shown). At the higher categorical thresholds (e.g., moderate), forecasts are issued far less often (Figs. 7c,f,i,l) across all configurations, which subjectively compares well with the ERO forecast fractions (Fig. 3c); one caveat exists related to the FCS1-trained RF model, which issues a relative maximum across the SW region due to training on a large number of excessive rainfall events in New Mexico.

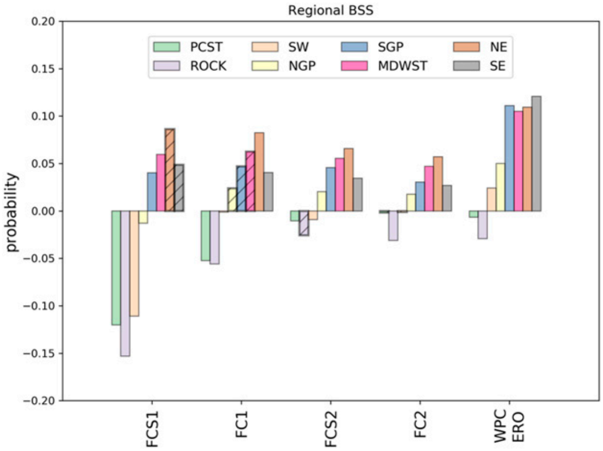

FIG. 6. Regional BSSs for WPC EROs and forecasts generated from RF-based models trained with different labels. Bar hatching and bolded edges depicts the BSS and corresponding label dataset that maximizes each region.

An additional aspect of the forecasts that is critical to evaluate is resolution, i.e., the ability of the forecasts to discriminate excessive rainfall events from nonevents. Regionally, forecast resolutions measured through the area under the ROC curve are comparable to ERO regional resolution for the various RF configurations (Fig. 8). For all regions, the highest forecast resolution arises from the FCS1-trained models (Fig. 8), likely due to the frequent issuance of probabilities that try to capture low-probability, complex terrain events, particularly in the Intermountain West. Additionally, the regional forecast resolutions from FCS1-trained models outperform the EROs, except in the SE region; this result coalesces with the lack of RF-based forecasts in the SE region (Fig. 7). Furthermore, all RF-based forecasts have better resolution than the ERO in PCST. Whereas the FCS1-model forecasts subjectively have the best resolution, FC2-model forecasts have fairly poor resolution, which is opposite of forecast skill (e.g., Fig. 5).

The average coverage of observations within categorical forecast outlooks is also considered, in order to measure how often probabilistic outlooks and forecasts are accurately conveying the spatial coverage of excessive rainfall (Fig. 9). For instance, a categorical slight risk conveys a 5%–10% risk of excessive rainfall within 40 km of a point, and therefore it is expected that each slight risk probabilistic contour has on

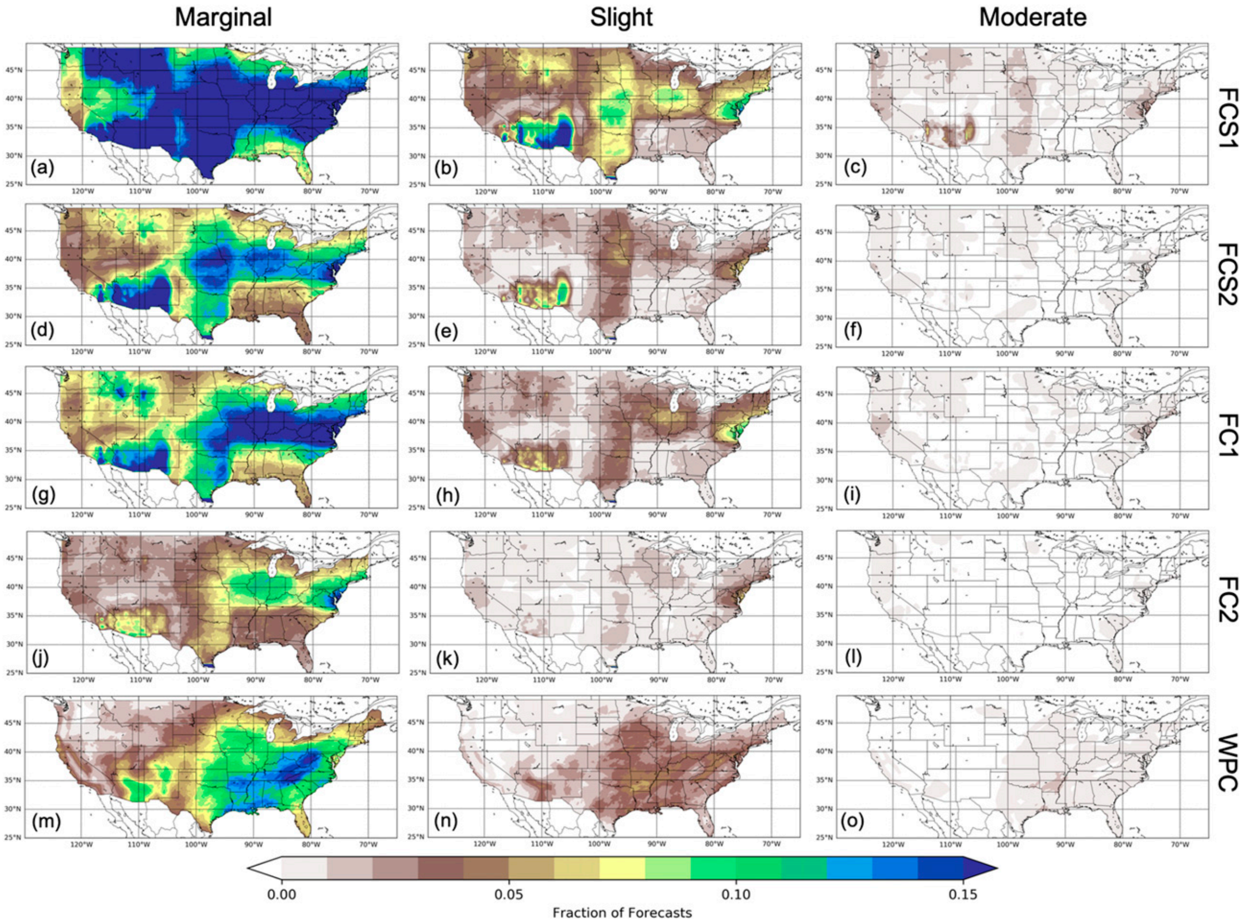

FIG. 7. Fraction of days within the verification period that featured at least a (left) marginal, (center) slight, and (right) moderate risk contour within RF-based forecasts generated by models trained with (a)–(c) FCS1, (d)–(f) FCS2, (g)–(i) FC1, and (j)–(l) FC2 label datasets. (m)–(o) Replicated from Fig. 3 for reference.

average a 5%–10% spatial overlap with observations (outlook area coverage lies between green and red lines in Fig. 9). Generally, EROs are too small regionally for each categorical outlook (above respective red lines in Fig. 9), which Erickson et al. (2021) and Schumacher et al. (2021) similarly showed. RF-based forecasts tend to be spatially calibrated when trained with FC1 and FCS2 for all outlooks (Fig. 9), and tend to be too large (below green lines in Fig. 9) when trained with FCS1 for the slight category (Fig. 9a). The FC2-trained models are oftentimes too small at the marginal and slight thresholds (Figs. 9a,b) and become better calibrated for the moderate categorical threshold (Fig. 9c), although some regions never experienced this probabilistic threshold during the verification period. RF-based forecasts in SE are nearly always too small when issued, which agrees with the lack of forecasts in this region (Fig. 7). Within each categorical outlook, the percent of outlook area covered by observations tends to increase as ARI thresholds increase and ST4 observations are withheld from model training, corresponding to forecast areas decreasing in size. This effect results in improved observation coverage when probability coverage is initially too large (below green lines), but poorer observation coverage (above red lines) when forecast areas initially are too small (e.g., SE region).

b. *Regional optimization and predictor sensitivities*

To further evaluate the sensitivities of CAM-based RF forecasts of excessive rainfall and their operational usefulness, two additional forecasts are generated as outlined in section 2. First, regional models maximizing BSS (hatched bars in Fig. 6) are used to generate forecasts across the CONUS (OPT forecasts), ideally maximizing regional forecast skill and resulting in improved CONUS-wide forecasts. Second, the same regional optimization procedure is conducted with the four label datasets but using SPT predictors (OPT_AVG forecasts). The skill of these forecasts is compared to EROs as well as the other RF-based forecasts.

The spatial skill patterns of OPT and OPT_AVG forecasts are qualitatively similar to both the ERO and other RF forecasts, i.e., negative skill to the west and positive skill east of the Rocky Mountains (Fig. 10). However, using skillful regional models to generate CONUS-wide forecasts results in improved CONUS BSS, rising from a maximum 0.036 in the FC1-trained RAW model (Fig. 5c) to 0.042 in OPT (Fig. 10a) and 0.052 in

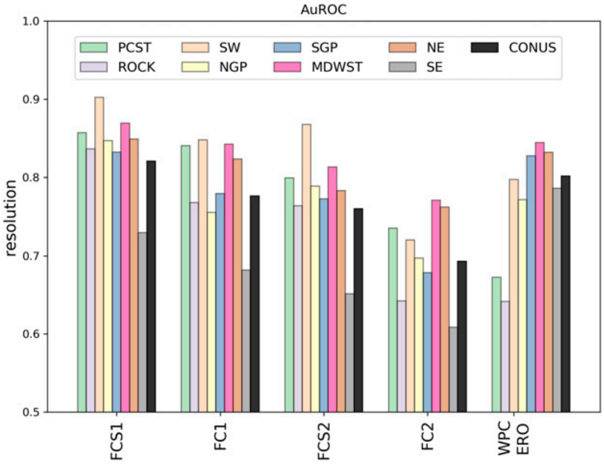

FIG. 8. As in Fig. 6, but for area under the ROC curve for each region and forecast system.

OPT_AVG (Fig. 10b). Whereas the spatial patterns between OPT and OPT_AVG are remarkably consistent, localized improvements in skill are responsible for the CONUS-wide BSS increase (0.042 to 0.052), including improved skill in Colorado, New Mexico, central Wyoming, and the mid-Atlantic region. Forecast degradation also exists—for instance, in West Texas—but it is often balanced in magnitude with an increase in skill—for example, in central and southern Texas. Overall, CONUS-wide skill from CAM-based RFs is improved when maximizing regional skill as well as simplifying predictors, but WPC skill for day-1 forecasts is still unmatched.

OPT and OPT_AVG forecasts defined by the categorical outlooks are qualitatively more aligned with the EROs, but they still exhibit high-frequency biases in large portions of the CONUS (Fig. 11). OPT marginal forecasts in ROCK are more frequent than the EROs, which correspond to better skill (cf. Figs. 11a and 3a). Across the southeastern United States, marginal-risk EROs are issued frequently across the Gulf Coast where there is a relative forecast frequency minimum in both the OPT and OPT_AVG forecasts (Figs. 11a,b). OPT_AVG marginal-risk forecasts are issued less frequently in ROCK and more frequently in the SW region, resulting in regionally degraded and improved skill, respectively. Given the robust positive ERO skill (Fig. 2c), it appears the marginal RF forecasts are issued too frequently. Moderate-risk forecasts are issued far less frequently, but an obvious maximum across Virginia, Maryland, and New Jersey (Fig. 11c) is not replicated by the EROs (Fig. 3c). Additionally, OPT_AVG has slightly improved skill relative to OPT in the SGP and NGP regions, partially attributed to increased slight-risk forecast frequencies (Fig. 11d), and a handful of moderate-risk forecasts (Fig. 11f).

OPT and OPT_AVG are better calibrated spatially than the EROs for marginal, slight, and moderate risk categories; EROs are on average too small (Fig. 12). The overall reliability of OPT and OPT_AVG suggests an underforecasting bias compared to observations above the 10% probability threshold (Fig. 13a). When comparing OPT and OPT_AVG to the other RAW predictor RF models (Fig. 13a), FC1 and FCS2-trained models have similar reliability, while FC2-trained models underforecast more severely and FCS1-based forecasts slightly overforecast events across all probability thresholds. When replacing RAW predictors with SPT predictors in model

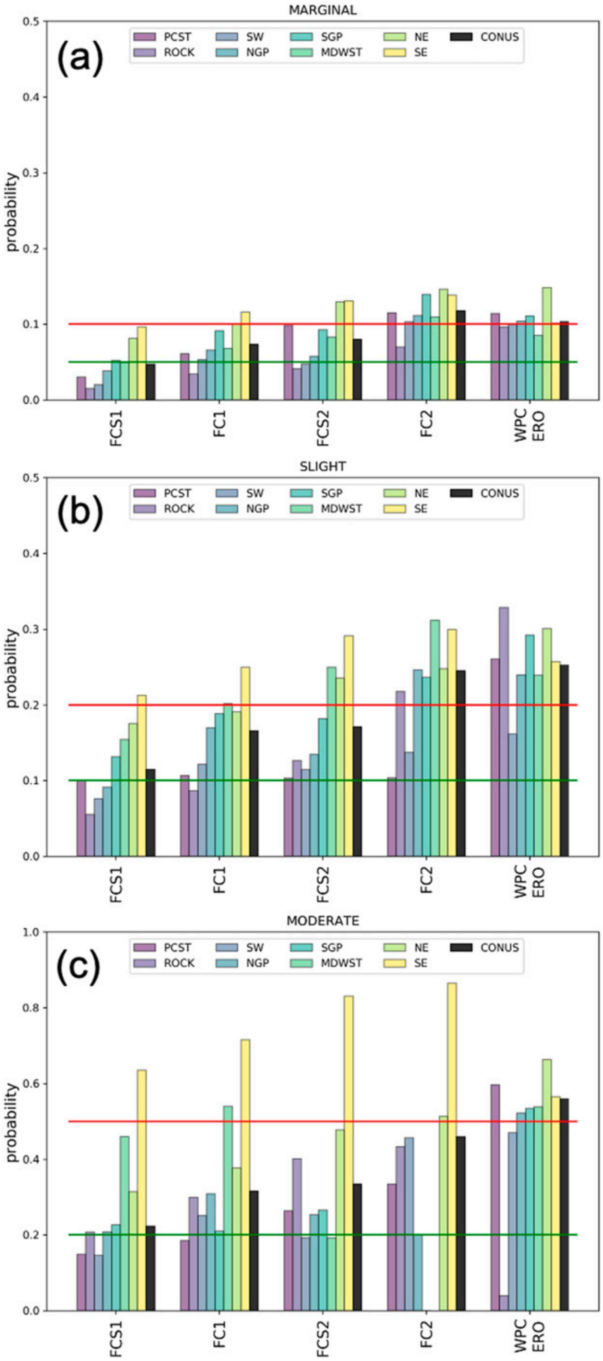

FIG. 9. As in Fig. 6, but fractional coverage of observations for the (a) marginal, (b) slight, and (c) moderate probability categories present in RF-based forecasts and WPC EROs. Green horizontal lines depict the lower bound of each probability bin; red horizontal lines depict the upper bound.

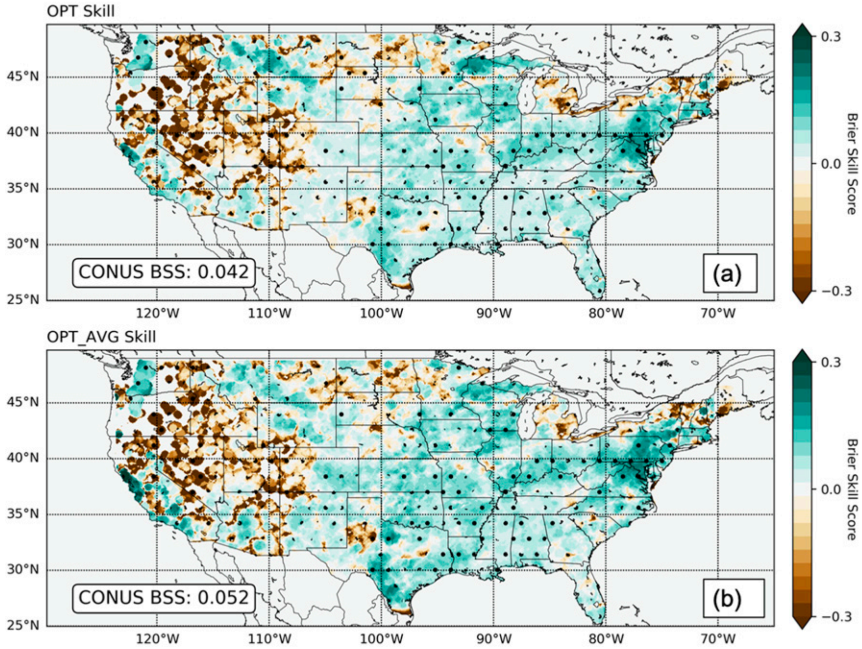

FIG. 10. As in Fig. 5a, but BSS of (a) OPT and (b) OPT_AVG forecasts as discussed in the text.

training, forecasts are issued more often (Fig. 13b). Models that were underforecasting events now move closer to perfect reliability (black dashed line in Fig. 13b), while the SPT-based model trained with FCS1 overforecasts excessive rainfall even more (brown line in Fig. 13b). The overall adjustment to more reliable forecasts when moving from RAW to SPT predictors explains why OPT_AVG is slightly more reliable than OPT, which directly translates into improved spatial BSSs.

Forecast resolution is also improved by using SPT predictors (Fig. 14). EROs exhibit an AuROC of 0.8 across the CONUS (Fig. 14a), whereas the best performing RAW model has an AuROC of 0.81 (FCS1-trained). All SPT-based models have improved resolutions over their RAW-based counterparts, and the best SPT model has an AuROC of 0.84. However, a drop in resolution for the other SPT-based models—0.81, 0.8, and 0.75 AuROC for the FC1, FCS2, and FC2-trained models, respectively—suggests that as the models progressively underforecast observed events (Fig. 13b) and get smaller (e.g., Fig. 9), forecasts are less able to discriminate excessive rainfall events from nonevents. The resolution improvements also manifest between OPT and OPT_AVG, which have 0.78 and 0.81 AuROC, respectively (Fig. 14).

Finally, the BSSs for each trained model (all RAW and SPT models) and the EROs are disaggregated monthly to consider any seasonal performance biases across the forecasts (Fig. 15). The monthly ERO BSSs are compared against RAW- and SPT-based models for each label dataset (Figs. 15a–d), as well as OPT (Fig. 15e) and OPT_AVG (Fig. 15f). A statistical significance test is applied using bootstrapping to determine whether BSSs between the forecasts are significantly different. EROs have lower relative skill in the Northern Hemisphere fall and winter months, and increased skill in the summer and early fall months (green lines in all panels of Fig. 15). Notably, ERO skill is only significantly better than the RFs during the summer and early fall months. The monthly ERO skill follows closely with that of the climatology of extreme precipitation across the CONUS (Stevenson and Schumacher 2014), and others have noted the positive skill of WPC QPFs for landfalling cyclones (e.g., Sukovich et al. 2014), which could partially contribute to increased skill during this time period. The RF-based forecasts follow a similar monthly skill pattern to the EROs, but raw skill (i.e., not considering significance testing) is always worse. RF-based skill, no matter the predictor set or label dataset used, is significantly worse than climatology in the late fall, winter, and early spring months (Figs. 15a–d). RF-based forecast skill improves into the summer months, but it is rarely distinguishable from climatological skill, and it is often statistically worse than the ERO skill between June and October. The least skillful RF-based forecast is generated from the FCS1 labels (Fig. 15a), and slight improvements in skill are observed over the summer months when SPT predictors are used in the FC1 and FC2 models (Figs. 15c–d). Compared to monthly ERO skill, the best forecast model is OPT_AVG, which is only statistically worse than ERO forecasts three months out of the year (Fig. 15f).

c. Forecast example and subjective evaluations

To provide event-based context for the results discussed herein, an ERO outlook and RF-based forecasts are presented from a high-impact flooding event in the Southeast United States (Fig. 16). The WPC issued a moderate risk of excessive rainfall at 0801 UTC 12 April 2020 spanning a narrow region extending from northern Mississippi to eastern Tennessee and western North Carolina (Fig. 16a); a corresponding marginal

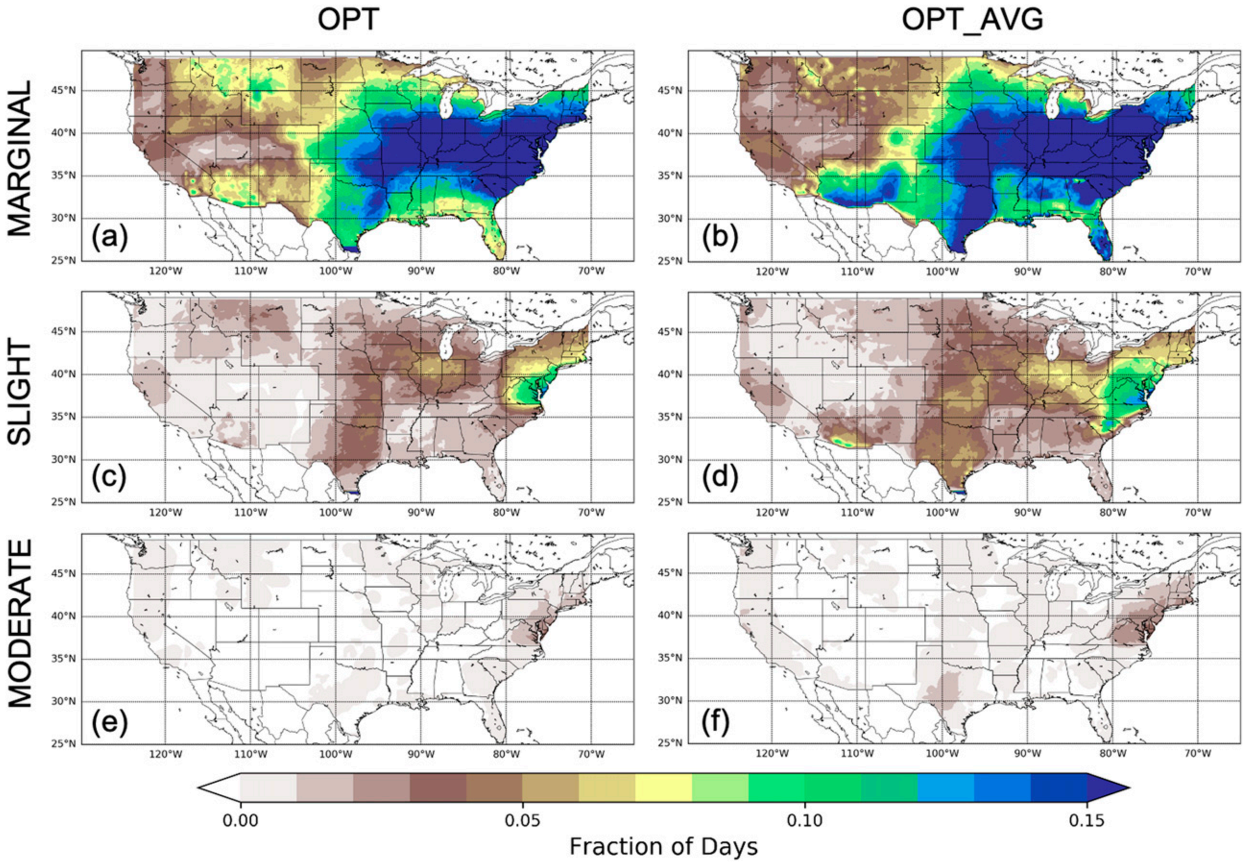

FIG. 11. As in Fig. 7, but forecast fractions are plotted for (left) OPT and (right) OPT_AVG in the (a),(b) marginal, (c),(d) slight, and (e),(f) moderate probability categories.

risk contour encompassed a large region from eastern Texas to southern Wisconsin eastward to the mid-Atlantic region. Excessive rainfall observations from the UFVS covered the entirety of the moderate risk area and large portions of the lower-probability risk areas (colored circles in Fig. 16a). A forecast from the FCS1-trained RAW predictor RF model demonstrates similar orientation to the ERO on this day, but the probabilities are, generally, lower than those from the ERO across the southeastern United States (Fig. 16b). In addition, the slight risk contour extends southwestward into Louisiana and Arkansas, but it is ultimately not as cohesive as the ERO slight risk contour, contributing to a poorer BSS for the day (0.08 for the RF, 0.152 for the ERO). A CONUS-wide forecast with optimized regional models in the OPT configuration slightly increases the forecast quality (0.0824 BSS) by refining the extent of the marginal contour and increasing the maximum probabilities in eastern Tennessee and northern Alabama and Georgia (Fig. 16c). In contrast, the OPT_AVG forecast enhances forecast probabilities considerably across the axis of observed excessive rainfall, and extends the slight risk northward and southward (Fig. 16d); OPT_AVG BSS jumps to 0.13 in this case.

Probabilistic excessive rainfall forecasts from ML models have been evaluated extensively since 2017 at the Hydrometeorological Testbed Flash Flood and Intensive Rainfall (FFaIR) experiment—a proving ground for new forecast guidance products (Barthold et al. 2015; Erickson et al. 2019)—in an effort to gauge the value of ML-based forecast products as operational forecasting tools. Participants span the weather enterprise—from NWS forecast offices to academia—and they evaluate new NWP models and guidance products that will be implemented in WPC forecast operations. During FFaIR experiments in 2019 and 2020, the three RF model forecast systems discussed in the previous paragraph were subjectively evaluated by FFaIR participants over four weeks spanning late June and early July; the FCS1-trained forecasts were evaluated in 2019, and the OPT and OPT_AVG forecasts were compared in the 2020 experiment. In the 2019 experiment, a global ensemble-based RF system was also in development (e.g., Schumacher et al. 2021) to produce day-1 excessive rainfall outlooks (hereafter called the GEFS model), and it was evaluated alongside the CAM-based RF forecasts as well as participant generated EROs. Over the four week evaluation period, participants subjectively ranked each product on a scale of 1–10, and unsurprisingly, the experimental ERO (i.e., produced by the participants) had the highest ranking (6.7), followed by the GEFS-based (6.07) and the CAM-based EROs (5.76) (Trojniak and Albright 2019). While objective measures

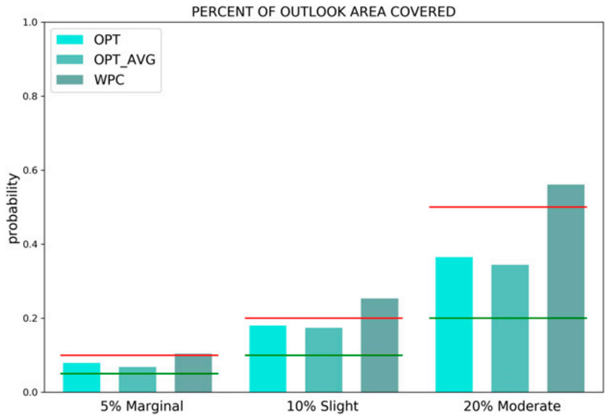

FIG. 12. Comparison of fractional coverage of observations for each probability category [marginal (5%–10%), slight (10%–20%), and moderate (20%–50%)] between the OPT and OPT_AVG forecasts, as well as the WPC EROs. Green horizontal lines depict the lower bound of each probability bin; red horizontal lines depict the upper bound.

of skill and resolution showed fairly similar results over the four-week period between the various EROs, participants noted “over forecasting seen in the NSSL ERO across the Northern Rockies and Northern Plains.” The final 2019 FFaIR report suggested participants “would reduce the [NSSL WRF] product by one probabilistic category (e.g., reduce a moderate to a slight)” when digesting the RF forecasts. These subjective interpretations are consistent with the objective results in this study, in which a categorical outlook frequency bias (Figs. 7a–c) and poor skill (Fig. 5a) was noted in the FCS1-trained RF forecasts.

In the 2020 FFaIR experiment, a new GEFS-based RF model along with the CAM-based OPT and OPT_AVG models were provided to participants to guide their experimental ERO, and participants graded each guidance product individually as in 2019. Again, the GEFS-based forecasts were deemed subjectively better (6.64 rating) compared to either the OPT (4.03) or OPT_AVG (5.22) (Trojniak et al. 2020). While OPT_AVG did outperform OPT in participant evaluation, the 2020 FFaIR report noted “the NSSL EROs had difficulty identifying areas of the higher magnitude excessive rainfall risks” (Trojniak et al. 2020), consistent with the underforecasting bias at higher probability thresholds noted previously in both forecast systems (Fig. 13). However, the report also noted “the change in training between OPT and OPT_AVG led to an increase in both the probability of being in Marginal and a Slight Risk,” effectively reducing the underforecasting bias in the OPT (Fig. 13). The 2020 report was also quick to point out “the [OPT_AVG] product was also able to identify the risk of excessive rainfall associated with the Southwest Monsoon,” which was also highlighted in the forecast fractions (Figs. 11b,d) and BSS (Fig. 10). In contrast, the participants found the OPT_AVG forecasts were too often focused in the Carolinas compared to the operational ERO (e.g., Fig. 11d). Despite the inferiority of CAM-based RF forecasts to the GEFS-based versions, FFaIR organizers felt OPT_AVG “should be the new configuration the [Colorado State University] team uses as they continue to refine the CAM-scale first guess ERO” (Trojniak et al. 2020). The results presented herein, along with FFaIR subjective evaluations, suggest there exists potential for CAM-based RFs to effectively forecast excessive rainfall and be an operational tool in the future.

5. Summary and conclusions

RF models are built to predict the occurrence of excessive rainfall over 24-h periods, analogous to WPC EROs, and forecast sensitivities to training labels and inputs are evaluated in consideration of how best to develop a CAM-based RF forecast system. NSSL WRF convection-allowing model forecasts are processed as inputs to the RFs. Because excessive rainfall is an ill-defined event, multiple precipitation and flooding datasets are used to define labels for the RFs—including flash flood reports and 1- and 2-yr ARI exceedances—and separate models are trained regionally across the CONUS. Separately, predictor assembly procedures are also considered, one in which meteorological predictors are selected at raw grid points over a defined box around training points, and a second in which all the raw gridpoint predictors are spatially averaged for a particular time and input variable. Training occurs over an approximately 7-yr period from 9 June 2009 to 31 August 2016, and forecast verification is conducted over 1 January 2017–31 December 2018.

EROs follow a general pattern of skillful forecasts east of the Rockies and nonskillful forecasts in the western CONUS; forecasts are generally worse than climatology in the Intermountain West where excessive rainfall and flooding events are not frequent and forecasts are issued more often than rainfall events are observed. RF-based forecast skill generally follows the same spatial pattern, with a negative/positive skill demarcation along the Rocky Mountains. However, localized skill differences exist as a result of different labels. Using the ST4 precipitation analysis to define ARI exceedances has deleterious effects on CONUS-wide forecast skill for all but the NE and SE regions, as the ST4 product identifies far more excessive rainfall events across the country, which in turn drives the RFs to issue more frequent forecasts. Aggregate forecast skill is improved by considering a stricter ARI threshold (i.e., 2 years) as well as removing ST4 ARIs from the training labels. However, the opposite is true of resolution: forecast resolution degrades as training label definitions become more strict.

CONUS-wide forecast skill improvements are noted when forecasts are created using regionally skillful models (i.e., the OPT forecasts), which highlights the importance of regionally varying labels in training. Moreover, SPT-predictor model forecasts (OPT_AVG) that also use regionally skillful models are more accurate and have better resolution than their RAW-predictor model counterparts. The OPT_AVG forecast improves upon OPT by reducing an underforecast bias and forecast resolution also improves as a result. It is speculated that using RAW predictors from a CAM introduces noisy predictors into RF training, and much of that noise is reduced when spatially averaging the meteorological information,

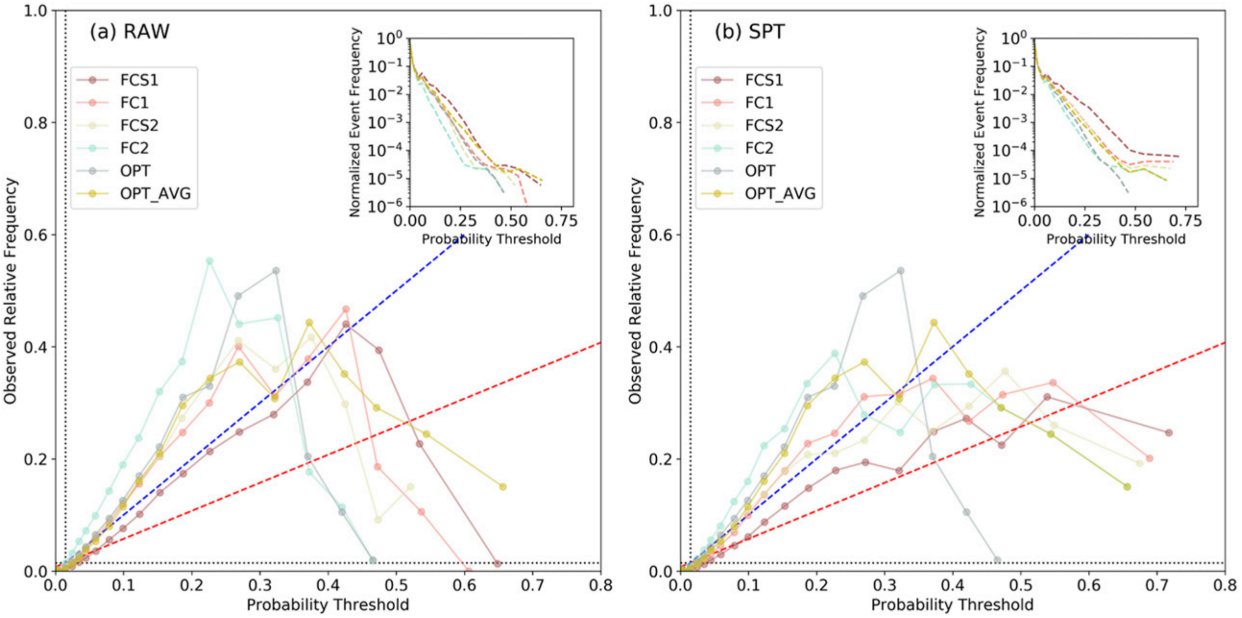

FIG. 13. Reliability diagrams of RF model forecasts generated with models trained using (a) RAW and (b) SPT predictors. Separate RF-based forecast models are described by different colors in the legend, and they are consistent with definitions provided in the text. Inset graph describes the normalized frequencies of forecasts at each probability threshold. Dashed blue and red lines represent perfect reliability and no skill, respectively. Dotted black vertical and horizontal lines denote no resolution (climatology).

although a thorough interrogation of this hypothesis is reserved for future work. In general, when the CAM-based forecasts incorporate regional skill optimization and spatially averaged predictors, forecast resolution and skill increases.

Both RF-based forecasts and EROs exhibit seasonal skill, with the worst skill in the late fall and winter months, and best skill in the late spring, summer, and early fall months; the latter follows closely with a seasonal maximum in excessive rainfall (Stevenson and Schumacher 2014). However, RF-based forecast skill is never statistically better than the EROs over the verification period, despite following the same skill patterns. The OPT_AVG model is subjectively the most skillful model seasonally, as there are only three months where it is statistically worse than the EROs. An example case study highlights

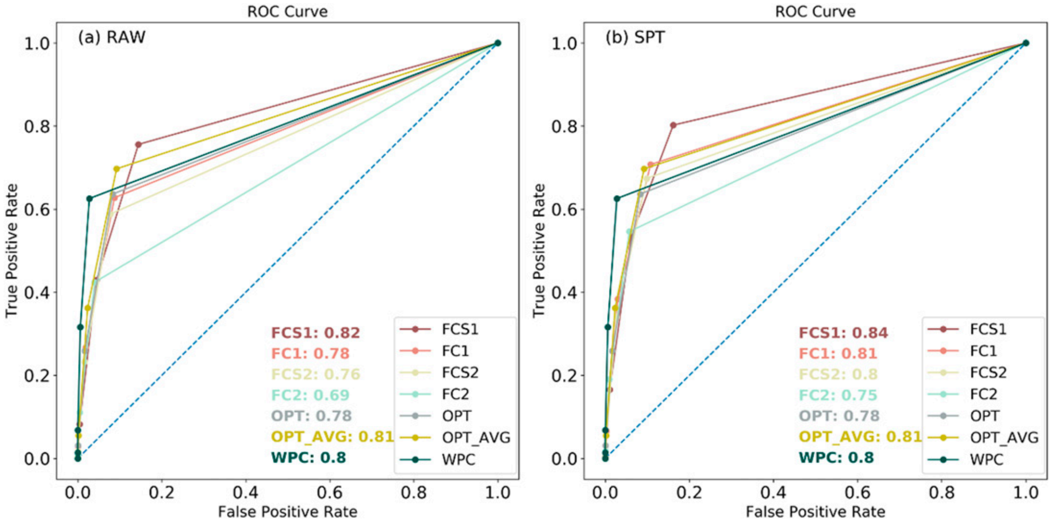

FIG. 14. ROC curves of RF model forecasts generated with models trained using (a) RAW and (b) SPT predictors. Separate RF-based forecast models are described by different colors in the legend, and consistent with definitions provided in the text. Values provided in the bottom right of (a) and (b) are the area under each colored ROC curve. Dashed blue line denotes no resolution.

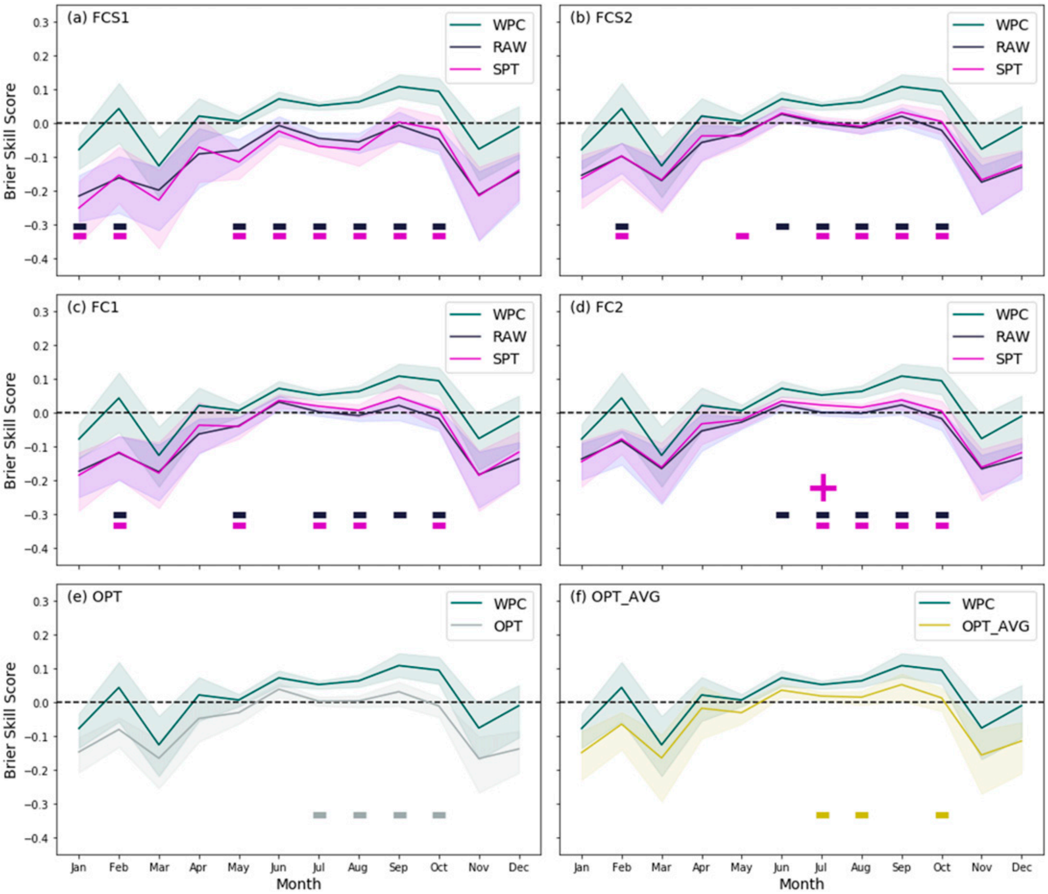

FIG. 15. Monthly BSS for (a) FCS1, (b) FCS2, (c) FC1, and (d) FC2-trained RAW and SPT models, as indicated in the legends. WPC ERO monthly BSS is also shown in green. (e) OPT and (f) OPT_AVG monthly BSS compared to WPC EROs. Colored shading represents 95% confidence interval obtained from bootstrapping each individual monthly distribution of forecasts. A colored plus in (a)–(d) delineates when RF skill of the corresponding color is statistically significantly greater than the other RF forecast; bootstrap samples do not overlap. A colored minus in (a)–(f) means RF forecast skill of the same color is statistically significantly worse than the WPC ERO skill.

many of the results of the study, including the propensity for the CAM-based models to have excessive areas of marginal risk (5% probability) compared to the operational EROs, and a general increase in probabilities as the models progress from statically trained to regionally optimized with SPT predictors (i.e., OPT_AVG). Reports from FFaIR as well as specific participant feedback and subjective evaluation synthesizes the results further: the CAM-based guidance products may be useful to operational forecasters in the future as ‘‘first-guess’’ ERO products, but significant improvements must be made to the products.

To summarize, CAM-based forecasts of excessive rainfall presented herein have reasonable resolution, that is they can distinguish excessive rainfall events from nonevents, and the resolution is on par with WPC EROs under many configurations. However, forecast probabilities are generally not well calibrated from models trained with static labels, resulting in poorer skill relative to the EROs. When regional rainfall climatologies are varied across the CONUS, model forecasts become better calibrated (i.e., OPT and OPT_AVG). The results presented herein, which suggest superiority of human-based forecast guidance over statistical guidance in the day-1 time frame, have been noted by other studies as well (e.g., Hill et al. 2020; Schumacher et al. 2021). Day-1 human-generated forecasts are hard to beat when forecasters are utilizing a plethora of other available information—including many numerical weather model forecasts as well as observations—to generate their forecasts. Furthermore, while the focus of this work was on examining excessive rainfall events, EROs explicitly forecast the exceedance of FFG within 40 km of a point. One limitation of this work is that FFG is not considered as an input or label for the RFs, yet RF forecasts are being verified against FFG within the UFVS. Also, WPC EROs may not be completely separated from other ML guidance during the verification period; operationalized day-2 and day-3 ML models based on inputs from the GEFS have been used at the WPC by forecasters as first-guess fields since 2017. It is possible that ML guidance products were used to generated day-2 and day-3 EROs, and that information filtered into the day-1 outlooks; ERO skill is not independent of other ML guidance and may

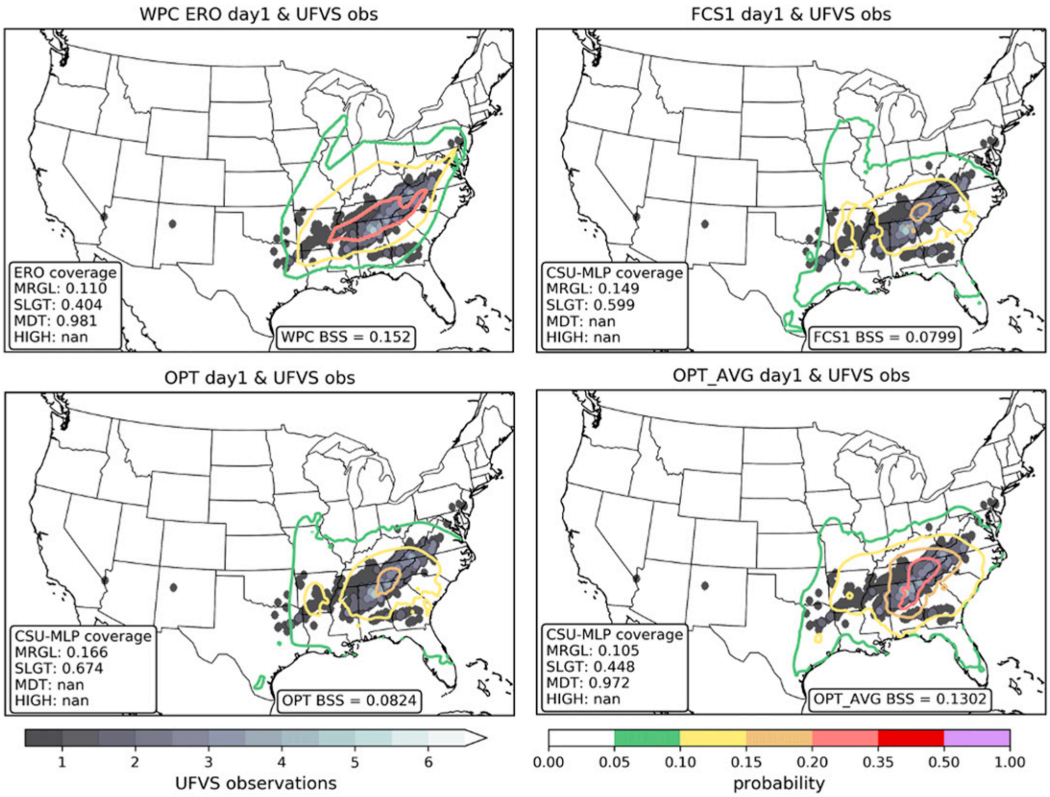

FIG. 16. (a) WPC day-1 ERO issued at 0801 UTC 12 Apr 2020 valid for the 24-h period ending at 1200 UTC 13 Apr 2020 (contoured) with UFVS observations defining excessive rainfall within 40 km of the grid point in shaded circles. (b) Forecast from RF-model trained with FCS1 and RAW predictors valid for the same period in (a) and UFVS observations overlaid. Intermediate probability contour of 15% is included. (c),(d) As in (b), but forecasts from OPT and OPT_AVG, respectively.

have benefited from products that are comparable to the CAM-based products presented (e.g., Trojniak and Albright 2019; Trojniak et al. 2020).

Future work will explore training new RF models with labels defined by the UFVS so that forecasts are more consistent with WPC verification procedures; this procedure would allow RFs to additionally learn about instances of FFG exceedance during training. However, a shorter training period would need to be used, and the impact this would have on forecast skill is unknown. Furthermore, the value of ensemble information gained by CAM-ensemble inputs, as well as the perceived superiority of global ensemble-based RFs (i.e., GEFS RFs) over their CAM-based counterparts—noted by FFaIR participants—will be thoroughly investigated. The most important consideration moving forward is how useful these RF products will be to operational forecasters at the WPC. This work is just a first step in producing an operationally ready product and continued evaluation at future FFaIR experiments will be critical to expose researchers and WPC forecasters to the products and gather subjective interpretations on a case-by-case basis, which are difficult to tease out of aggregate forecast statistics.

Acknowledgments. The authors are indebted to Dr. Greg Herman who developed the initial NSSL WRF-based RF models, and left behind an infrastructure for future model development. The authors also thank Ben Albright, Sarah Trojniak, Jimmy Correia, Mike Erickson, and Sarah Perfater for organizing FFaIR, producing robust final reports, and synthesizing too many experimental products to count. The authors are extremely grateful to all of the FFaIR participants and WPC forecasters that have been evaluating these products for the past few years; you have provided valuable insight into how to improve our products. The authors also thank Dr. Adam Clark for access to NSSL WRF data. We would also like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. We appreciate the valuable comments from three anonymous reviewers whose reviews enhanced the quality of our manuscript. This work is supported by NOAA Joint Technology Transfer Initiative Grant NA18OAR4590378.

Data availability statement. All ML forecasts are stored on local data servers and are available upon request from Dr. Russ Schumacher. Additionally, they can be accessed online at http://schumacher.atmos.colostate.edu/hilla/csu_mlp/nsslwrf_paper/. Excessive rainfall datasets are available from public-facing data servers at NOAA or available upon request from NOAA employees. FFRs available online at https://mesonet.agron.iastate.edu/request/gis/lsrs.phtml. UFVS data are available from WPC at https://ftp.wpc.ncep.noaa.gov/ERO_verif/.

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. Hydrol.*, **578**, 124106, https://doi.org/10.1016/j.jhydrol.2019.124106.

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.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brocca, L., F. Melone, and T. Moramarco, 2008: On the estimation of antecedent wetness conditions in rainfall–runoff modelling. *Hydrol. Processes*, **22**, 629–642, https://doi.org/10.1002/hyp.6629.

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.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*, **135**, 3456–3473, https://doi.org/10.1175/MWR3467.1.

——, ——, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

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.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Erickson, M. J., J. Kastman, B. Albright, S. Perfater, J. Nelson, R. Schumacher, and G. 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.

——, B. Albright, and J. A. Nelson, 2021: Verifying and redefining the Weather Prediction Center's Excessive Rainfall Outlook forecast product. *Wea. Forecasting*, **36**, 325–340, https://doi.org/10.1175/WAF-D-20-0020.1.

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

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, ——, 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.

Goines, D. C., and A. D. Kennedy, 2018: Precipitation from a multiyear database of convection-allowing WRF simulations. *J. Geophys. Res. Atmos.*, **123**, 2424–2453, https://doi.org/10.1002/2016JD026068.

Gourley, J. J., and Coauthors, 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.

——, and H. Vergara, 2021: Comments on "Flash flood verification: Pondering precipitation proxies." *J. Hydrometeor.*, **22**, 739–747, https://doi.org/10.1175/JHM-D-20-0215.1.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

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

——, and ——, 2018a: Flash flood verification: Pondering precipitation proxies. *J. Hydrometeor.*, **19**, 1753–1776, https://doi.org/10.1175/JHM-D-18-0092.1.

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

Hill, A. J., 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., and Coauthors, 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.

Krocak, M. J., and H. E. Brooks, 2018: Climatological estimates of hourly tornado probability for the United States. *Wea. Forecasting*, **33**, 59–69, https://doi.org/10.1175/WAF-D-17-0123.1.

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.

——, ——, 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.

Marjerison, R. D., M. T. Walter, P. J. Sullivan, and S. J. Colucci, 2016: Does population affect the location of flash flood reports? *J. Appl. Meteor. Climatol.*, **55**, 1953–1963, https://doi.org/10.1175/JAMC-D-15-0329.1.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

NCEI, 2020: Billion-dollar weather and climate disasters: Summary stats. U.S. billion-dollar weather and climate disasters (2020). NOAA/National Centers for Environmental Information (NCEI), accessed 31 August 2020, https://www.ncdc.noaa.gov/billions/summary-stats/US/2010-2019.

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 intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

NOAA/Weather Prediction Center, 2021: Excessive rainfall outlooks. NOAA, accessed 21 January 2021, https://www.wpc.ncep.noaa.gov/html/fam2.shtml#excessrain.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, https://doi.org/10.1175/WAF-D-13-00066.1.

Ogden, F., H. Sharif, S. Senarath, J. Smith, M. Baeck, and J. Richardson, 2000: Hydrologic analysis of the Fort Collins, Colorado, flash flood of 1997. *J. Hydrol.*, **228**, 82–100, https://doi.org/10.1016/S0022-1694(00)00146-3.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Romine, G. S., C. S. Schwartz, C. Snyder, J. L. Anderson, and M. L. Weisman, 2013: Model bias in a continuously cycled assimilation system and its influence on convection-permitting forecasts. *Mon. Wea. Rev.*, **141**, 1263–1284, https://doi.org/10.1175/MWR-D-12-00112.1.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

Schumacher, R. S., 2017: Heavy rainfall and flash flooding. *Nat. Hazard Sci.*, https://doi.org/10.1093/acrefore/9780199389407.013.132.

——, 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.

——, 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.*, https://doi.org/10.1175/BAMS-D-20-0186.1, in press.

Schwartz, C. S., and R. A. Sobash, 2019: Revisiting sensitivity to horizontal grid spacing in convection-allowing models over the central and eastern United States. *Mon. Wea. Rev.*, **147**, 4411–4435, https://doi.org/10.1175/MWR-D-19-0115.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Smith, J. A., A. J. Miller, M. L. Baeck, P. A. Nelson, G. T. Fisher, and K. L. Meierdiercks, 2005: Extraordinary flood response of a small urban watershed to short-duration convective rainfall. *J. Hydrometeor.*, **6**, 599–617, https://doi.org/10.1175/JHM426.1.

Sobash, R. A., G. S. Romine, 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.

Stevenson, S. N., and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea. Rev.*, **142**, 3147–3162, https://doi.org/10.1175/MWR-D-13-00345.1.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911, https://doi.org/10.1175/WAF-D-13-00061.1.

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

Trojniak, S., and B. Albright, 2019: 2019 flash flood and intense rainfall experiment: Findings and results. NOAA, 123 pp., https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2019_FFaIR.pdf.

——, J. Correia Jr., and B. Albright, 2020: 2020 flash flood and intense rainfall experiment: Findings and results. NOAA, 99 pp., https://www.wpc.ncep.noaa.gov/hmt/Final_Report_2020_FFaIR_Experiment_Nov13.pdf.

Whan, K., and M. Schmeits, 2018: Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Mon. Wea. Rev.*, **146**, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1.

Wong, M., G. Romine, and C. Snyder, 2020: Model improvement via systematic investigation of physics tendencies. *Mon. Wea. Rev.*, **148**, 671–688, https://doi.org/10.1175/MWR-D-19-0255.1.

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.