Verification Results from the 2017 HMT–WPC Flash Flood and Intense Rainfall Experiment

MICHAEL J. ERICKSON,[a,d] JOSHUA S. KASTMAN,[a,d] BENJAMIN ALBRIGHT,[b,d] SARAH PERFATER,[c] JAMES A. NELSON,[d] RUSS S. SCHUMACHER,[e] AND GREGORY R. HERMAN[f]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *Systems Research Group, Inc., College Park, Maryland*
[c] *NOAA/OAR/Office of Weather and Air Quality, and Cherokee Nation Businesses, Silver Spring, Maryland*
[d] *NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland*
[e] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*
[f] *The Climate Corporation, Seattle, Washington*

(Manuscript received 26 April 2019, in final form 26 August 2019)

ABSTRACT

The Flash Flood and Intense Rainfall (FFaIR) Experiment developed within the Hydrometeorology Testbed (HMT) of the Weather Prediction Center (WPC) is a pseudo-operational platform for participants from across the weather enterprise to test emerging flash flood forecasting tools and issue experimental forecast products. This study presents the objective verification portion of the 2017 edition of the experiment, which examines the performance from a variety of guidance tools (deterministic models, ensembles, and machine-learning techniques) and the participants' forecasts, with occasional reference to the participants' subjective ratings. The skill of the model guidance used in the FFaIR Experiment is evaluated using performance diagrams verified against the Stage IV analysis. The operational and FFaIR Experiment versions of the excessive rainfall outlook (ERO) are evaluated by assessing the frequency of issuances, probabilistic calibration, Brier skill score (BSS), and area under relative operating characteristic (AuROC). An ERO first-guess field called the Colorado State University Machine-Learning Probabilities method (CSU-MLP) is also evaluated in the FFaIR Experiment. Among convection-allowing models, the Met Office Unified Model generally performed optimally throughout the FFaIR Experiment when using performance diagrams (at the 0.5- and 1-in. thresholds; 1 in. = 25.4 mm), whereas the High-Resolution Rapid Refresh (HRRR), version 3, performed best subjectively. In terms of subjective and objective ensemble scores, the HRRR ensemble scored optimally. The CSU-MLP overpredicted lower risk categories and underpredicted higher risk categories, but it shows future promise as an ERO first-guess field. The EROs issued by the FFaIR Experiment forecasters had improved BSS and AuROC relative to the operational ERO, suggesting that the experimental guidance may have aided forecasters.

1. Introduction

a. Background

Flash flooding is defined as a rapid and extreme flow of high water into a normally dry area, or a rapid water-level rise in a stream or creek above a predetermined flood level within 6 h of the causative event (NOAA 2012). Between 2015 and 2017, flash flooding has resulted in more fatalities than lightning, hail, tornadoes, and straight-line wind damage from thunderstorms combined (NWS 2017). Between October 2015 and October 2016, there have been approximately $19 billion in losses from flooding disasters over the contiguous United States (CONUS; Novak 2017).

Significant predictability issues remain with potential flash flooding events due to challenges associated with quantitative precipitation forecasts (QPF) during the warm season (Yu et al. 2013; Clark et al. 2016; Gowan et al. 2018), which are likely driven by smaller-scale forcing when compared to the cool season (Fritsch and Carbone 2004; Sukovich et al. 2014). Unfortunately, the treatment of flash flooding is inconsistent beginning with the definition of flash flooding and continuing into the forecasting, reporting, and verification of these events (Barthold et al. 2015). For instance, there are a variety of disparate forecast products issued by National Weather

Corresponding author: Michael Erickson, mjaerickson@gmail.com

DOI: 10.1175/JAMC-D-19-0097.1

TABLE 1. Datasets used in the objective verification portion of the 2017 FFaIR Experiment.

Data name	Provider	Data type	Total members	Cycles used	Period evaluated	Total sample size
NAM Nest	EMC	Model	1	0000 UTC	Day 2	35
FV3-GFDL	GFDL	Model	1	0000 UTC	Days 2 and 3	25
FV3-GFS	EMC	Model	1	0000 UTC	Days 2 and 3	26
FV3-CAPS	OU/CAPS	Model	1	0000 UTC	Days 2 and 3	18
UM-Oper.	Met Office	Model	1	0000 UTC	Days 2 and 3	21
HRRR-Exp	ESRL/GSD/EMC	Model	1	1200 UTC	Day 2	19
HREFv2	EMC	Ensemble	8	0000 UTC	Day 1 (1800–0000 UTC)	32
SSEFx	OU/CAPS	Ensemble	11	0000 UTC	Day 1 (1800–0000 UTC)	16
HRRRe	ESRL/GSD	Ensemble	9	0000 UTC	Day 1 (1800–0000 UTC)	24
CSU-MLP	CSU	First-guess field	1	—	Days 2 and 3	40

Service offices to alert the public to different types of flood risk. Adding to the complexity, the relationship between flood response and QPF is not linear and depends on complex upstream basin characteristics unique to that locality. Ideally, the hydrologic component to flash flooding should be better integrated into the forecast process when making a flash flood forecast (Gochis et al. 2015; Gourley et al. 2017; Li et al. 2017). The Weather Prediction Center (WPC) Flash Flood and Intense Rainfall (FFaIR) Experiment allows for the exploration of new flash flooding products with the future goal of utilizing a more integrated system.

b. Motivation

During the FFaIR Experiment, participants from across the weather enterprise can work together to explore the utility of emerging model guidance and tools for improving flash flood forecasts in a real-time pseudo-operational environment. The FFaIR Experiment was developed in 2013 within WPC's Hydrometeorology Testbed (HMT) and is typically held for four weeks from mid-June to mid-July (Barthold et al. 2015). During the FFaIR Experiment, a WPC forecaster and 10–12 participants (consisting of meteorologists, hydrologists, developers, and researchers) utilize a variety of experimental statistical and dynamical products to generate experimental forecasts. Historically, the majority of guidance evaluated and utilized in the FFaIR Experiment are experimental dynamical models from a variety of agencies [e.g., Center for Analysis and Prediction of Storms (CAPS) Geophysical Fluid Dynamics Laboratory (GFDL), Environmental Modeling Center (EMC), and Earth System Research Laboratory (ESRL)]. During the experiment, the participants refrain from looking at WPC operational forecasts and largely restrict their guidance to the products shown in Table 1, but they are free to analyze current in situ and remote observations.

As mentioned in section 1a, the flash flood forecasting paradigm needs a consistent and reliable flash flood database for verification. This is particularly crucial for WPC's excessive rainfall outlook (ERO), which in its current form is a probabilistic forecast of precipitation exceeding flash flood guidance (FFG; Schmidt et al. 2007) within 40 km of a point over the CONUS. In previous FFaIR Experiments (Barthold et al. 2015; Perfater and Albright 2017), instances of rainfall exceeding FFG created by River Forecast Centers was used as a proxy for flash floods. FFG provides an estimate for the amount of rain over an area and time period that may cause small streams to flood, given local soil moisture and streamflow conditions. However, FFG is subject to error and may not properly capture the complexity of observed flooding (Clark et al. 2014). No single source of flooding observations can be considered fully comprehensive across all of CONUS (Herman and Schumacher 2018c). For instance, relying solely on observations from National Weather Service local storm reports can result in missed observations and inaccurate reporting associated with the difficulty of separating regular and flash floods (Barthold et al. 2015; Gourley et al. 2013).

To address some of the inconsistencies identified in Barthold et al. (2015), WPC is taking a holistic approach to the flash floods paradigm in the 2017 FFaIR Experiment by looking at a variety of atmospheric and hydrologic models, statistical models, and verification methods. The future core of the Unified Forecast System known as the GFDL Finite-Volume Cubed-Sphere Dynamical Core (FV3; Lin and Rood 1997; Lin 1997) has been introduced to the FFaIR Experiment for evaluation. In addition, emerging hydrologic guidance from the National Water Model (Cohen et al. 2018) and the Flooded Locations and Simulated Hydrographs (Gourley et al. 2017) system has been evaluated by participants. For the first time in the FFaIR Experiment, an ERO first-guess field is explored called the Colorado State University Machine-Learning Probabilities method (CSU-MLP). Also new to the 2017 FFaIR Experiment,

additional proxies and observations for flooding are being explored with the goal of creating a more comprehensive verification.

This study presents an overview of the objective verification system first established in the FFaIR 2017 Experiment. The primary focus of this study will be on objectively verifying the forecast products issued by the FFaIR Experiment participants, with a secondary focus on the deterministic and ensemble atmospheric models and statistical tools. The objective verification of forecaster-issued products allows for a comparison of FFaIR products with WPC operational products. When appropriate, the objective ratings will be compared to the participants' subjective ratings.

Section 2 details the data and methods, including the experimental guidance, issued forecast products, and verification used in the FFaIR Experiment. Section 3 presents the verification results, with an emphasis on differences in skill and bias between different forecast products and experimental guidance. Section 4 briefly summarizes the important lessons learned in the 2017 FFaIR Experiment and discusses future directions for the experiment. The appendix contains a list of the acronyms used in this paper.

2. Data and methods

In 2017, the FFaIR Experiment was conducted for four weeks spanning from Monday 19 June to Friday 21 July 2017, with no experiment running the week of Monday 3 July 2017. This year featured a massive multiagency collaboration effort between WPC and CAPS, GFDL, the Met Office, ESRL-Global Systems Division (GSD), EMC, CSU, Office of Water Prediction, Meteorological Development Laboratory, National Severe Storms Laboratory, and University of Oklahoma. In this paper, we focus mostly on the products that could be objectively verified. Details of the products used, collaborative efforts, forecasts issued, and verification efforts are detailed below.

a. Experimental model guidance

A variety of dynamical and statistical guidance is evaluated during the FFaIR Experiment for the day 1 (valid 1200 UTC on the current day to 1200 UTC 1 day into the future), day 2 (valid 1200 UTC 1 day into the future to 1200 UTC 2 days into the future), and day 3 (valid 1200 UTC 2 days into the future to 1200 UTC 3 days into the future) forecast periods. The 24-h accumulated QPF from all deterministic models are verified for days 2 and 3 (Table 1). Deterministic guidance includes the FV3-GFDL (uses GFDL microphysics), the FV3-Global Forecast System (FV3-GFS; uses GFS physics), FV3-CAPS (uses Thompson microphysics), and Met Office Unified Model Operational (UM-Oper.), all initialized at 0000 UTC preceding day 1. In addition, the 24-h QPF from the 1200 UTC experimental High-Resolution Rapid Refresh (HRRR-Exp) and the North American Mesoscale Forecast System model nest (NAM nest) are evaluated for day 2 only, since these models do not extend out to the day 3 time period. Additional details for the deterministic guidance, including sample size of model runs throughout the FFaIR Experiment, are shown in Table 1.

In addition to the deterministic guidance, ensembles are verified over the day 1 6-h QPF between 1800 and 0000 UTC for the High-Resolution Ensemble Forecast, version 2 (HREFv2), experimental Storm-Scale Ensemble Forecast (SSEFx), and experimental High-Resolution Rapid Refresh Ensemble (HRRRe). The 6-h QPF evaluated consists of a 50% blend of the probability-matched mean and the conventional mean (i.e., the arithmetic mean) from the ensemble. The probability-matched mean (Ebert 2001) sets the probability distribution function of the ensemble mean equal to that of the collective ensemble members. The localized probability-matched mean (Perfater and Albright 2017; Blake et al. 2018) calculates the probability-matched mean over small patches of the domain and then applies a Gaussian smoother to the data. The localized probability-matched mean provides many of the advantages of the probability-matched mean while retaining small-scale structures that may be of meteorological interest to forecasters. During the 2016 FFaIR Experiment, the probability-matched mean exhibited good spatial structure but produced values that were too high, while the conventional mean produced overly smoothed values that were too low (Perfater and Albright 2017). Combining the probability-matched mean and conventional mean into a blended mean preserved the best aspects of both means.

New to the 2017 FFaIR Experiment, the CSU-MLP is a machine-learning random-forest technique trained on 11 years of Global Ensemble Forecast System reforecasts to predict probability of QPF exceeding the 1-yr average recurrence intervals (ARIs; Herman and Schumacher 2018a,b). While the 1-yr ARI is not entirely consistent with the current operational WPC ERO definition of quantitative precipitation estimates exceeding FFG, ARIs are becoming an important flooding proxy as detailed in section 2c. The CSU-MLP technique is used to generate first-guess fields for days 2 and 3. This study defines a "first-guess field" as a tool that can be used as a starting point to aid forecasters in the creation of a forecast product. The performance of the first-guess fields are subjectively evaluated by forecasters and objectively evaluated using flooding observations and

proxies described in section 2c. Participants were asked to subjectively rank different products on a scale from 1 (poor) to 10 (great) by writing their score on a whiteboard and presenting it to the scorekeeper. Participants did not have to be present for the entire week to participate.

b. *Forecast issued products in the FFaIR experiment*

WPC participants used the deterministic and ensemble guidance, first-guess fields, and additional guidance (e.g., National Water Model) to create several probabilistic forecasts. Specifically, the forecasters issued a day 1 6-h probability of flash flooding forecast valid 1800 to 0000 UTC, a day 2 24-h experimental ERO, and a day 3 24-h experimental ERO. This paper will focus on the objective verification of the day 2 and day 3 experimental ERO, with comparisons with the operational EROs issued at 0900 UTC.

The operational ERO was defined until 13 October 2017 as the probability of QPF exceeding FFG at a point, while the experimental ERO was defined as the probability of flooding rains occurring within 40 km of a point. The ERO probabilities used in the FFaIR Experiment are marginal = 5%–15%, slight = 15%–30%, moderate = 30%–50%, and high = 50%–100%. These 40-km ERO probability thresholds are derived from a 1.5-yr retrospective verification of the operational ERO extrapolated to a 40-km radius (Erickson and Nelson 2018). Hence, the ERO probability categories used in the FFaIR Experiment can be applied to the operational ERO if a 40-km-neighborhood radius is used. Note that the definition of the operational ERO was changed to a 40-km radius effective 13 October 2017.

c. *Verification*

As discussed in section 1b, no single flooding dataset can be considered fully comprehensive to sampling all flooding events. Starting in 2017, the FFaIR Experiment began addressing the inconsistency in flash flood reports by creating a comprehensive flash flood verification dataset consisting of flooding observations from local storm reports (LSR) and U.S. Geological Survey river gauge measurements. However, LSRs exhibit significant spatial discontinuity due to regional reporting biases and U.S. Geological Survey river gauge observations only sample a very small number of possible flooding locations (Gourley et al. 2013; Clark et al. 2014).

In an attempt to capture flash flooding occurrences that may be missed with traditional observations, grid-based flooding proxies are computed by examining instances of Stage IV analysis exceeding FFG. The Stage IV analysis is a near-real-time product generated by River Forecast Centers by utilizing radar precipitation estimates and rain gauges and includes some bias correction and manual adjustment of data (Nelson et al. 2016). A 5-km CONUS mosaic of FFG is generated at WPC from the original FFGs created by regional River Forecast Centers (Barthold et al. 2015). Section 1b briefly discusses the assumptions that go into calculating the FFG product. Clark et al. (2014) found relatively slow skill values between FFG and LSRs, perhaps due to reporting biases in LSRs, with slightly higher values comparing FFG to U.S. Geological Survey river gauge observations.

Finally, instances of Stage IV analysis exceeding 5-yr ARI are also considered as a flooding proxy. Instances of exceeding the 1-, 2-, and 10-yr ARI are also analyzed (not shown), but the 5-yr ARI subjectively aligns best with LSRs and captures 80% of all floods (Lincoln and Thomason 2018). The combination of all flooding observations and observation proxies (FFG and ARI exceedances) are referred to as the Unified Flooding Verification System (UFVS) within WPC. A detailed cross verification of all the datasets within the UFVS is beyond the scope of this study and likely very difficult to perform given reporting biases, but their combination is meant to capture all potential flooding occurrences that are likely missed by using one or two datasets. An example of a participant issued ERO with UFVS verification is shown in Fig. 1 valid between 1200 UTC 23 June and 1200 UTC June 24 in 2017. This was a significant flooding event with the experimental ERO being displaced slightly south of where the main flooding occurred.

The majority of verification performed at WPC is accomplished by using the Model Evaluation Tools, version 6.0 (METv6.0), in conjunction with a series of Python programming language wrappers (Brown et al. 2009). MET is used to evaluate the deterministic- and probabilistic-model guidance, flooding observations and proxies, and the forecast products issued in the FFaIR Experiment. Most of the plots produced in this study use MET output and are plotted in Python.

The skill of the deterministic and blended mean ensemble QPF forecasts are displayed using Roebber performance diagrams (Roebber 2009). The performance diagrams are a convenient way to simultaneously display probability of detection, false-alarm ratio, frequency bias (FB), and critical success index (CSI; Perfater and Albright 2017, their Fig. 5). Performance diagrams are computed for multiple precipitation thresholds, although the most relevant thresholds are presented in this study.

Experimental and operational ERO issuance probability is analyzed spatially throughout the experiment to determine the most active regions within CONUS and

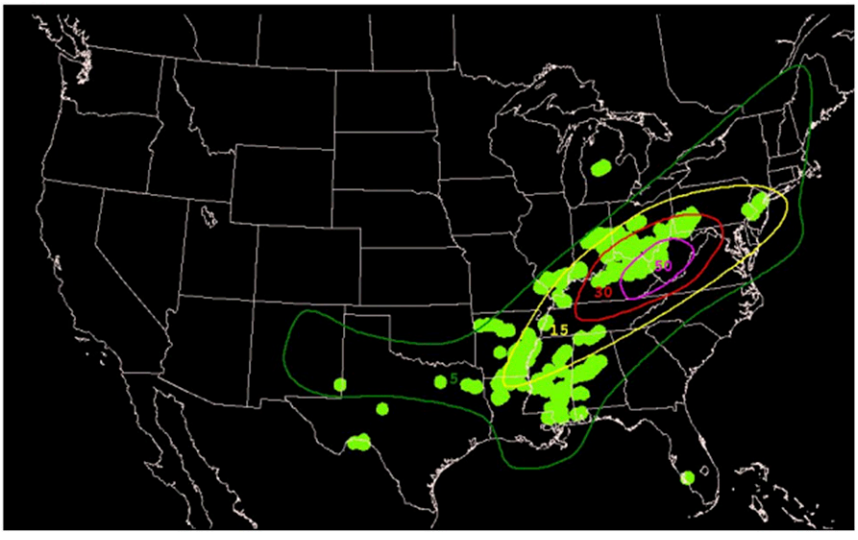

FIG. 1. The experimental ERO issued by participants valid between 1200 UTC 23 Jun and 1200 UTC 24 Jun 2017. Green dots denote instances of flooding observations and proxies in the UFVS with a 40-km-neighborhood radius applied.

highlight differences between the two types of forecasts. Since the ERO is a probabilistic forecast, calibration is assessed by computing the average fractional coverage of FFG exceeding Stage IV and all UFVS data within 40 km of a point for each probabilistic threshold (e.g., marginal, slight, moderate, and high). Fractional coverage for each ERO probabilistic threshold is also computed for the CSU-MLP.

Probabilistic forecast skill for the EROs and CSU-MLP is assessed by computing Brier scores (BS) and bulk area under the relative operating characteristic (AuROC). BS is a negatively oriented error metric and is analogous to mean-square error for probabilistic values while AuROC measures probability of detection versus false-alarm ratio with higher values representing more skill (Wilks 2011). Brier skill scores (BSS) are used to assess any potential improvement in probabilistic skill of the experimental ERO over the operational ERO throughout the FFaIR Experiment.

3. Results

a. Dynamical model verification

Performance diagrams for all deterministic-model 24-h accumulated QPF (i.e., NAM nest, FV3-GFDL, FV3-GFS, FV3-CAPS, UM-Oper., and HRRR-Exp) are plotted in Fig. 2 for days 2 and 3 exceeding both the 0.5-in. (12.7 mm) and 1-in. (25.4 mm) thresholds. In the performance diagrams, unbiased forecasts fall along the 1:1 line, while optimal forecasts approach unity in the top-right corner of the figure (Roebber 2009). All models exhibit a dry bias at the 0.5-in. threshold for days 2 and 3, with the FV3-CAPS exhibiting the greatest dry bias on day 2 (FB = 0.64) and the FV3-GFDL exhibiting the greatest dry bias on day 3 (FB = 0.65). There is greater model bias variability at the 1-in. threshold with the HRRR having the smallest day 2 dry bias (FB = 0.93) and the FV3-GFS having the greatest dry bias (day 2 FB = 0.40 and day 3 FB = 0.50). These results suggest that all models underpredict precipitation at both the 0.5- and 1-in. threshold. Participants generally viewed the FV3-GFS favorably, particularly for capturing the location of heavier precipitation, although it was noted that the model frequently underpredicted the higher amounts (Perfater and Albright 2017). In terms of CSI, the UM-Oper. exhibited the highest CSI for all days and thresholds, except for day 3 at the 0.5-in. threshold where the FV3-GFS was slightly higher. The HRRR-Exp (FV3-GFDL) had the lowest CSI for QPF exceeding 0.5 in. on day 2 (day 3), while the FV3-GFS had the lowest CSI for days 2 and 3 QPF exceeding 1 in.

The objective and subjective verification differs slightly when comparing CSI directly with the average ratings from the forecast participants. For instance, in the subjective verification boxplots (Figs. 3a,b), the HRRR-Exp performed best on day 2 (Fig. 3a) while the UM-Oper. performed best objectively on day 2 (Figs. 2a,b). On day 3, the UM-Oper. performed best objectively at 1 in. and close to best at 0.5 in. (Figs. 2c,d), which is consistent with the subjective results on day 3 (Figs. 2c,d). Participants commented that the HRRR-Exp

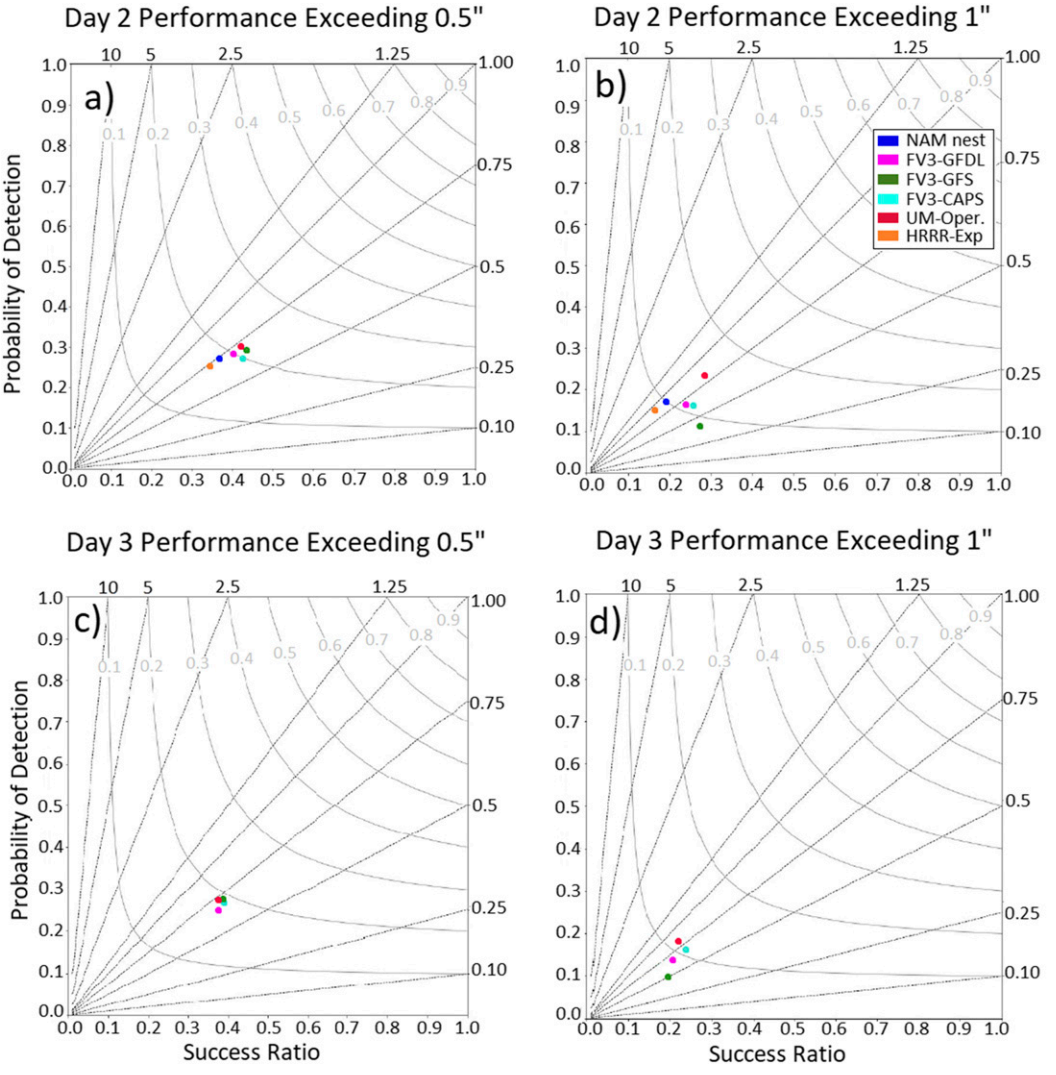

FIG. 2. Deterministic-model 24-h QPF performance diagrams for (a) day 2 exceeding 0.5 in., (b) day 2 exceeding 1 in., (c) day 3 exceeding 0.5 in., and (d) day 3 exceeding 1 in. Models shown are the NAM nest (blue), FV3-GFDL (magenta), FV3-GFS (green), FV3-CAPS (cyan), UM-Oper. (red), and HRRR-Exp (orange).

did well capturing the precipitation pattern over CONUS (Perfater and Albright 2017), which may not be reflected in the objective verification scores. The median of the FV3-CAPS generally performed the worst of all subjective evaluations for both days 2 and 3 (Figs. 3a,b), while objective verification results exhibited average performance (Fig. 2). Given the sample size (16–40 runs; Table 1) of the verification, statistical significance is difficult to deduce. Nonetheless, these results can be used to infer potential utility of the experimental guidance related to flash flooding forecasting and QPF during an active period. For instance, participants in the FFaIR Experiment mentioned that the FV3-CAPS generally did not produce enough precipitation and provided little utility (Perfater and Albright 2017). In several cases, the deterministic objective verification corroborates and quantifies the subjective evaluation from the participants.

Figure 4 shows the blended mean performance diagrams for the three ensembles evaluated in the FFaIR Experiment: the HREFv2, SSEFx, and the HRRRe. In terms of bias, the HREFv2 exhibits a dry bias at the 0.5-in. QPF threshold (FB = 0.60) and 1-in. threshold (FB = 0.30), while the HRRRe exhibits the smallest dry bias at both thresholds (FB = 0.85 at 0.5 in. and FB = 0.94 at 1 in.). The SSEFx also exhibits a dry bias generally in between the HRRRe and HREFv2. Despite the differing biases, the HREFv2 and HRRRe have similar CSI values for both thresholds, while the SSEFx exhibits reduced skill. Conversely, the SSEFx has the

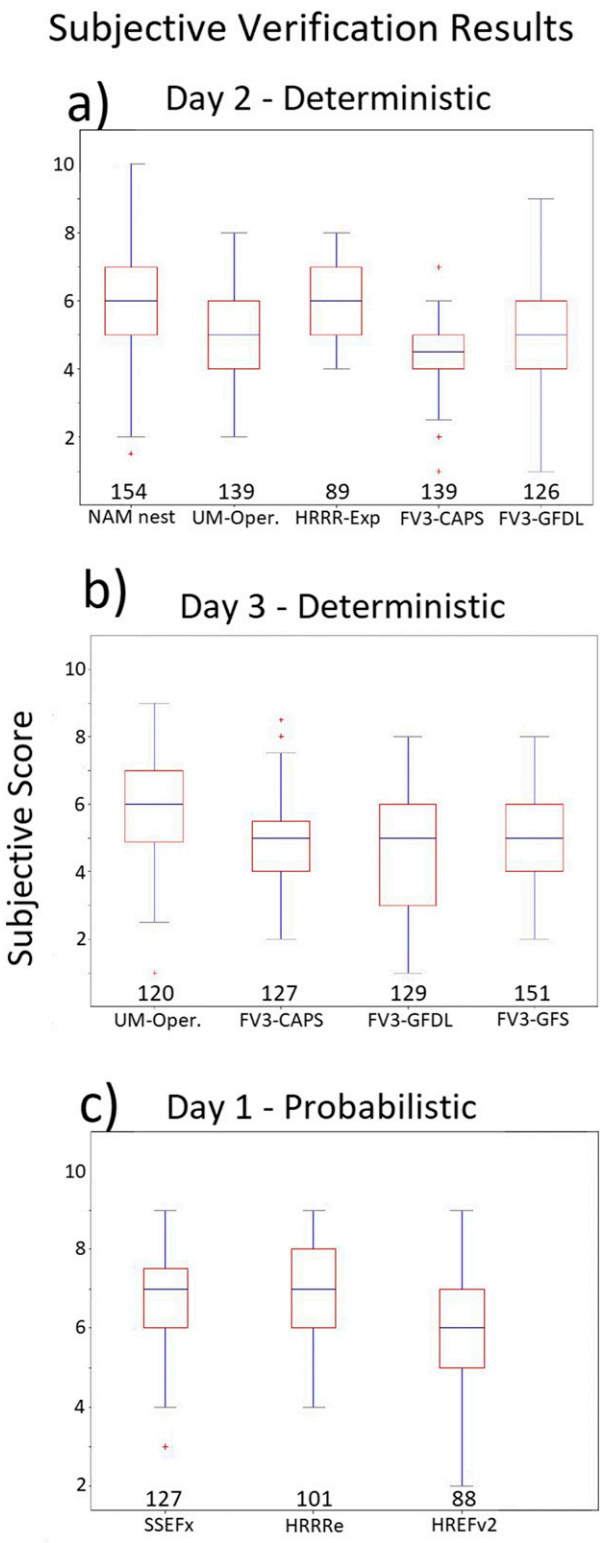

FIG. 3. Boxplot of the subjective verification results for the (a) day 2 and (b) day 3 deterministic models and the (c) day 1 1800–0000 UTC ensembles. Red symbols denote outliers, and the numbers below each boxplot show the total number of ratings.

highest rated average subjective score (6.73), with the HRRRe second (6.57), and the HREFv2 last (6.02; Fig. 3c).

The CSI values from Figs. 2 and 4 are very low and are generally consistent with the top 1% of all WPC QPF forecasts for July (Sukovich et al. 2014; their Fig. 8e). These results highlight that subjective and objective scores can differ, with the human eye recognizing convective structure and biases in the spatial patterns that remain ignored with grid-to-grid objective verification. Therefore, a human forecaster can still find value in a model forecast with a low CSI that produced displaced convection of a similar convective mode, intensity, or duration relative to what is observed. However, subjective ratings can vary from person to person depending on what object attributes and locations the forecasters value most in a QPF forecast. When evaluating the ensembles, participants were asked to subjectively rate the ensembles over a varying subdomain of CONUS, while objective verification was performed over all of CONUS. Nonetheless, inconsistencies between objective and subjective ratings is still useful feedback to model developers to address potential issues. For instance, model conditional biases related to geographical region or convective mode may not be captured by standard bulk verification metrics but noticed by forecasters. In general, the ensembles analyzed in this study were rated highly both subjectively and objectively, and the participant reactions to the ensembles were generally positive (Perfater and Albright 2017).

b. Average occurrence fields for the FFaIR experimental forecasts and first-guess products

The average spatial issuance probabilities of each ERO probabilistic risk category is analyzed during the 2017 FFaIR Experiment for the operational ERO, experimental ERO, and CSU-MLP. To compare the CSU-MLP to the categorical ERO field, the raw probability field is converted to the FFaIR Experiment defined risk categories. Note that the CSU-MLP is used as a first-guess field in the FFaIR Experiment while the EROs are human forecasts that may have considered the CSU-MLP as input, so a one-to-one comparison may not be particularly equitable. In addition, the CSU-MLP is designed to predict the probability of precipitation exceeding ARIs, rather than FFG or the entire UFVS (Herman and Schumacher 2018a). However, this study presents them side by side since they are all in a similar format.

The average issuance probability of predicting marginal risk for all three products is shown in Fig. 5. Both the operational ERO (Fig. 5a) and the FFaIR experimental ERO (Fig. 5b) highlight three active areas: the mesoscale convective systems (MCS) in the Midwest,

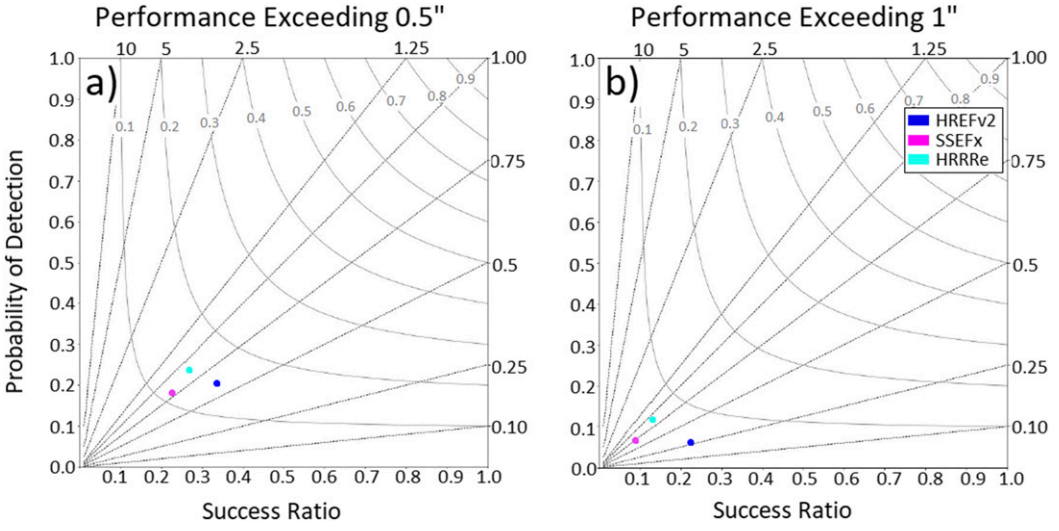

FIG. 4. Day 1 ensemble 6-h (between 1800 and 0000 UTC) blended mean QPF performance diagrams (a) exceeding 0.5 in. and (b) exceeding 1 in. Ensembles shown are the HREFv2 (blue), SSEFx (magenta), and HRRRe (cyan).

the convection and tropical activity in the Southeast, and the Southwest monsoon activity. The abundance of marginal issuances in the Southeast is the result of Tropical Storm Cindy during the first week of the experiment, with lingering convection for the second, third, and fourth weeks. The Midwest MCS activity is persistent throughout the experiment, whereas the Southwest monsoon began in the latter half of the experiment.

Both the operational and experimental EROs highlight the three ''hot spots'' mentioned earlier, although the fractional coverage of the experimental EROs is over 2 times the size (~207.7% greater) of the operational EROs. Of particular note, the experimental ERO increases the average issuance probabilities of marginal in portions of the Southeast and Southwest United States by over 4 times (e.g., certain locations that had two marginal issuances in the operational ERO had eight or greater issuances in the FFaIR ERO.) The CSU-MLP (Fig. 5c) is fairly similar to the operational ERO in the Midwest but shows much greater marginal risk probabilities in the Intermountain West, particularly in New Mexico and Colorado. In addition, CSU-MLP average issuance probability values are much lower in the Southeast compared to both ERO forecasts. These differences are noted by the FFaIR Experiment participants, particularly with regard to the large positive difference in the CSU-MLP from New Mexico to eastern Montana during monsoonal activity (Perfater and Albright 2017). The lower CSU-MLP values in the Southeast are likely caused by the greater abundance of FFG exceedances compared to ARI exceedances in that region.

Similar to Fig. 5, Fig. 6 shows the slight issuance probabilities for both EROs and the CSU-MLP. The three major marginal risk regions are highlighted in the slight category, with a fourth region stretching from Ohio to the central Appalachians. This region had several slight issuances from Tropical Storm Cindy and remnant MCS activity propagating in from the Midwest. The experimental ERO (Fig. 6b) slight risk covers over 2 times the size (~206.9%) relative to the operational ERO (Fig. 6a), with the greatest issuance frequency in the Midwest. The CSU-MLP (Fig. 6c) deviates greatly from the EROs, with the majority of slight issuances occurring over New Mexico, and very few issuances elsewhere in CONUS.

Issuances of the moderate risk are far rarer, occurring only with Tropical Storm Cindy and the most extreme MCS events extending from the Midwest to the Gulf Coast (Figs. 7a,b). Moderate risks are issued more abundantly with the experimental EROs and exhibited an average fractional coverage of almost 4 times as great an area (~394.9%) as the operational version. The CSU-MLP has one moderate issuance in the Florida Panhandle and a few issuances in southwestern New Mexico, but it fails to highlight any similar regions compared to the ERO products. Further investigation into the frequent prediction of high probabilities in New Mexico reveals that it is related to very frequent exceedance of ARI thresholds in the Stage IV precipitation analysis used to train the CSU-MLP model. This, in turn, results in the CSU-MLP model routinely predicting high probabilities for rainfall events that are not actually ''excessive'' or associated with flash flooding

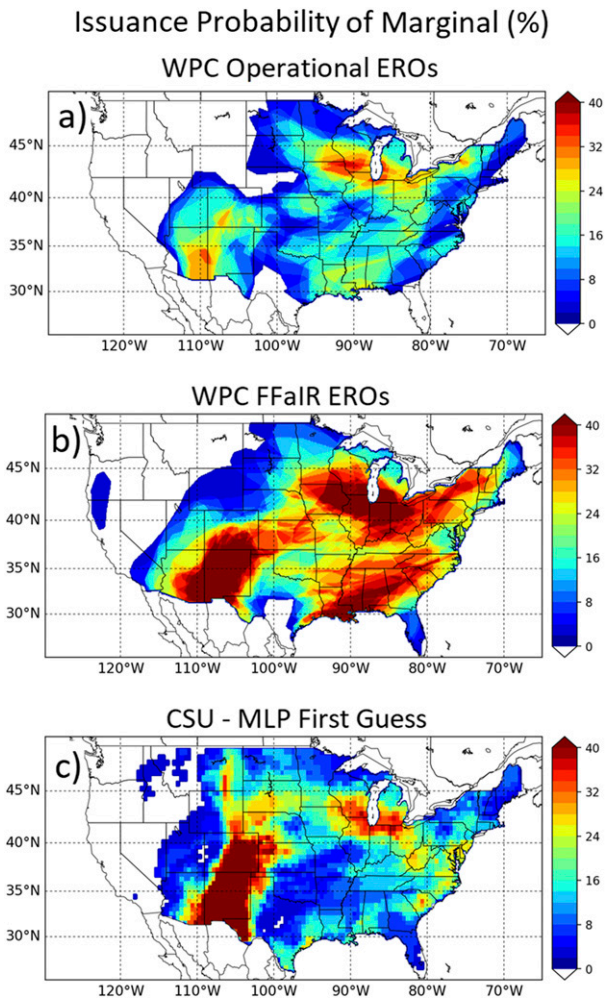

FIG. 5. Issuance probabilities (%) of marginal risk forecasts throughout the FFaIR Experiment period for (a) the operational ERO, (b) the FFaIR Experiment ERO, and (c) the CSU-MLP first-guess field.

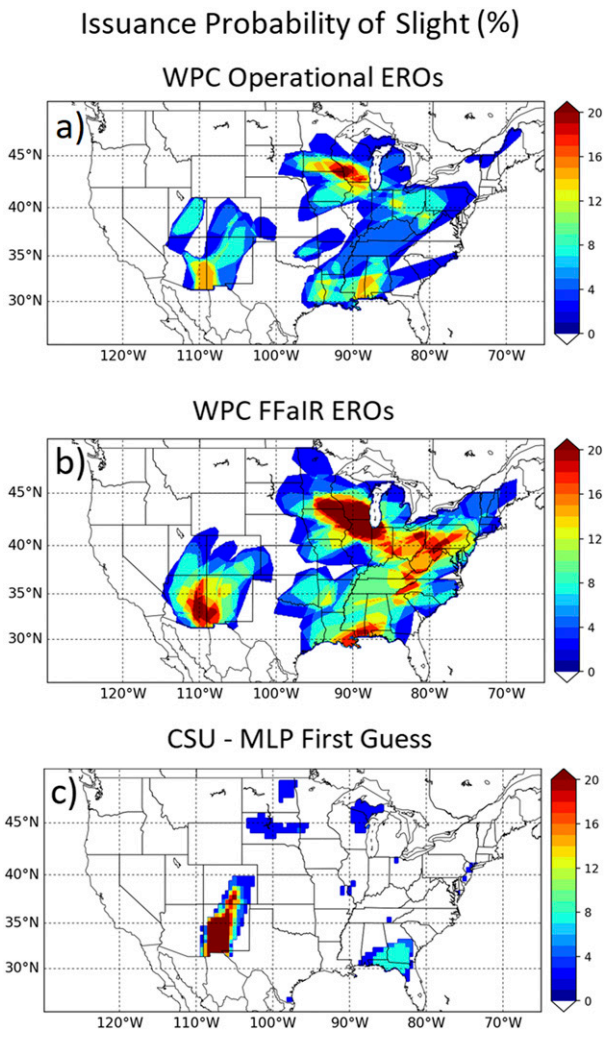

FIG. 6. As in Fig. 5, but for slight risk.

in New Mexico (Herman and Schumacher 2018a,c). In the FFaIR Experiment, this can result in forecasters systematically ignoring high probabilities in this region, even when high probabilities are warranted. Although these biases are not a problem with the machine-learning technique per se, this result is still undesirable, and other precipitation products are being used for training of the CSU-MLP model to alleviate this problem in the 2018 FFaIR Experiment. At the high risk threshold, only the experimental ERO have any issuances, all of which were associated with Tropical Storm Cindy (not shown).

c. *Verification of the FFaIR experimental forecasts and first-guess field*

As discussed in sections 2b and 3b, the ERO is a probabilistic forecast product consisting of four risk categories. The calibration of the ERO and CSU-MLP probabilities are assessed to determine if average observed relative frequencies from the UFVS match forecast probabilities (Wilks 2011). Throughout the FFaIR Experiment, the average fractional coverage of flooding occurrence or proxy within 40 km of a point is computed for each ERO risk category. On average, fractional coverage approximates for bins of average forecast probability, allowing for an ERO reliability plot (Wilks 2011) to be created. A forecast is considered reliable if the average fractional coverage falls within the probabilistic definition for each ERO category (e.g., in the case of the marginal category, the average fractional coverage must be between 5% and 15%).

Figure 8 shows the reliability for the CSU-MLP, operational ERO, and experimental ERO using the UFVS. For each ERO risk category, the horizontal green or red line represents the lower or upper bound,

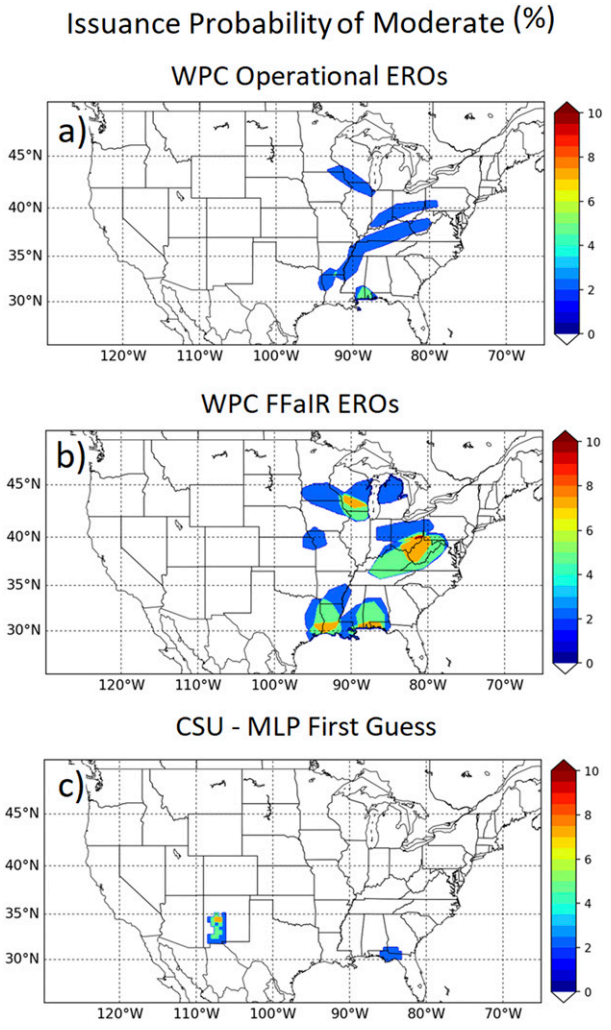

FIG. 7. As in Fig. 5, but for moderate risk.

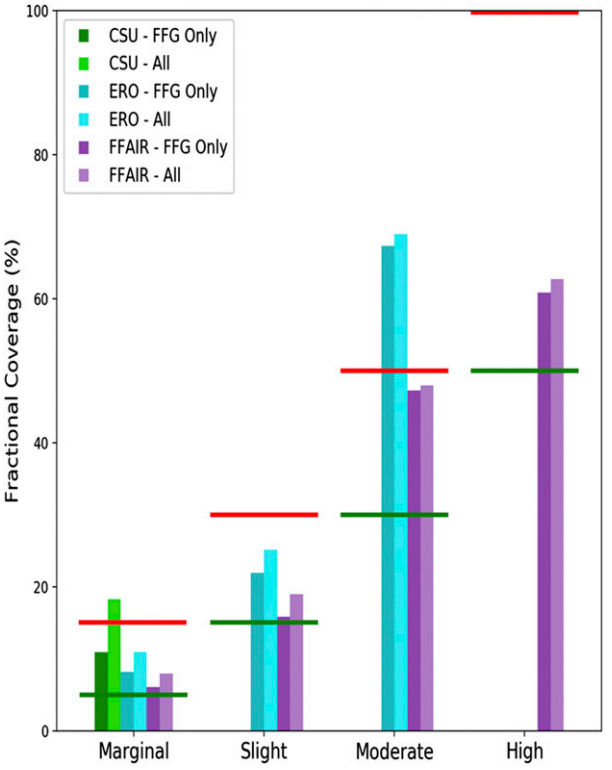

FIG. 8. Average fractional coverage of Stage IV exceeding FFG (label FFG Only) or all flooding observations and proxies (label All) by ERO risk category for the CSU-MLP (green), operational ERO (blue), and the FFaIR Experiment ERO (purple).

respectively, of the ERO definition. The CSU-MLP probabilities are calibrated only for marginal instances of precipitation exceeding FFG (i.e., the current operational ERO definition), but fails to identify slight, moderate, and high risk regions. The operational ERO is calibrated for the marginal and slight categories when considering FFG exceedances and all observations and proxies (i.e., the complete UFVS). However, the operational ERO exhibits fractional coverage that exceeds 60% for the moderate threshold, suggesting that forecasters should issue more moderate risks during borderline events or draw larger moderate areas to reduce fractional coverage. The experimental ERO is calibrated using all verification analyzed for all definitions of the ERO. This suggests that the larger and more frequent issuances of moderate and high risk categories in the experimental ERO was advantageous to improving calibration. The experimental ERO may have also benefited from the experimental guidance that was available to the FFaIR Experiment participants.

Probabilistic skill for the operational ERO, experimental ERO and CSU-MLP forecasts are assessed by computing the BS. Using BS, the BSS is computed to assess potential improvements in the experimental ERO by referencing the experimental ERO to both the operational ERO (Fig. 9a) and CSU-MLP (Fig. 9b). In this framework, the experimental ERO improves upon the referenced ERO if the BSS value is greater than zero. In general, there is significant day-to-day variability in the BSS values for the experimental ERO compared to the operational ERO (Fig. 9a). In bulk, the experimental ERO improves upon the operational ERO probabilistic values for day 2 and day 3 verified against everything in the UFVS. In all cases, there was a greater improvement in the experimental ERO compared to the operational ERO during the first two weeks of the FFaIR Experiment. This may have been caused by larger-scale forcing for precipitation during the two weeks of the FFaIR Experiment (e.g., Tropical Storm Cindy and synoptically forced MCS activity) before more weakly forced convection (e.g., monsoon, Gulf Coast, and weaker MCS activity) prevailed in the latter portion of the

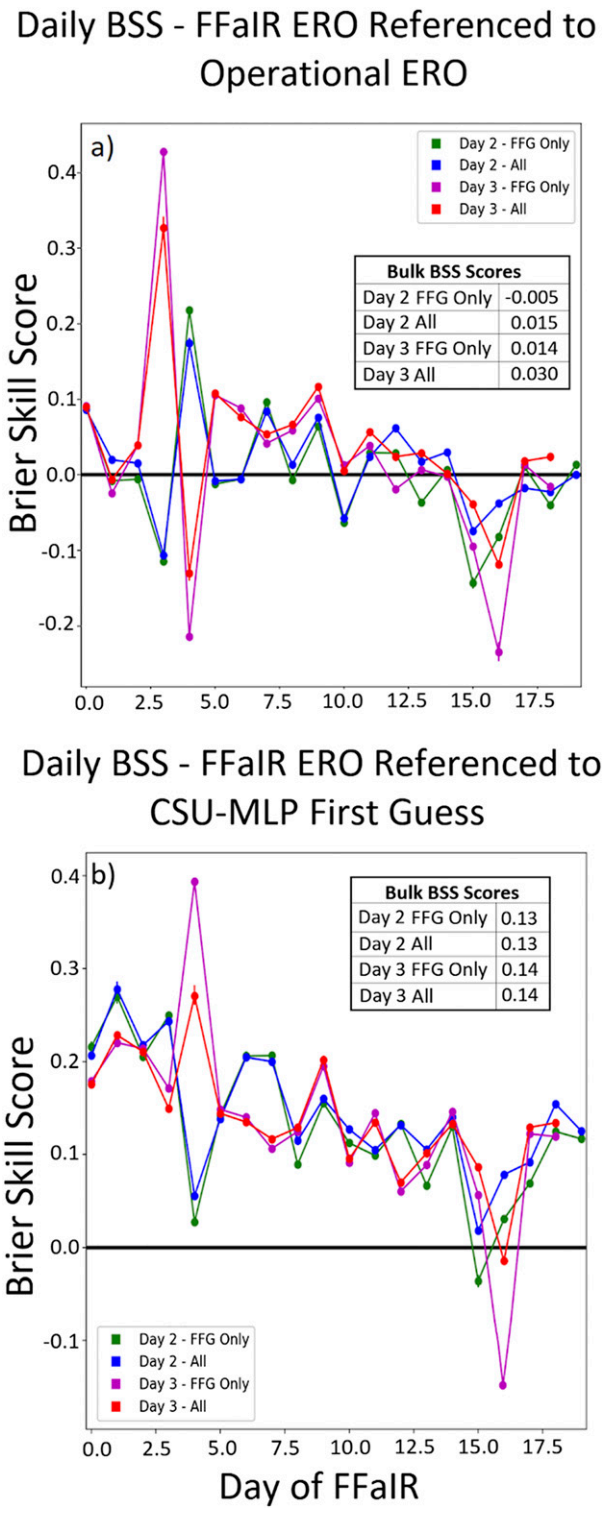

FIG. 9. Time series of daily BSS for the FFaIR Experiment ERO referenced to (a) the operational ERO and (b) CSU-MLP for day 2 verified against FFG (green), day 2 verified against all flooding observations/proxies (blue), day 3 verified against FFG only (magenta), and day 3 verified against all flooding observations/proxies (red).

experiment. The experimental ERO exhibited more probabilistic skill than the CSU-MLP (Fig. 9b) in 19 of 20 days analyzed, which is unsurprising considering that the participants were able to incorporate additional guidance beyond this automated product.

Another metric used to assess skill is the AuROC, which considers the hit rate versus the false-alarm ratio (Wilks 2011). Bulk AuROC values (higher values representing more skill) are shown for the CSU, operational ERO, and experimental ERO verified against both FFG and all of the UFVS (Fig. 10). For day 2, the CSU forecasts have the lowest skill followed by the operational ERO and the experimental ERO performing best. Interestingly, for day 3, the CSU forecasts have higher AuROC than the operational ERO and even the CSU day 2 forecast. This result suggests that the CSU-MLP may be useful as a first-guess field on day 3, which is a critical time period where WPC forecasters have no previous day 4 forecast to start from. However, the experimental ERO forecasts have the highest AuROC on both days 2 and 3.

When considering the fractional coverage, BSS, and AuROC values collectively, the experimental ERO performs slightly better and are slightly more calibrated than the operational ERO. These results suggest that the experimental guidance may have made a slight but important impact on the experimental ERO. This is particularly true on day 3, where BSS is consistently slightly improved (Fig. 9a) and AuROC is about 0.2 greater in the experimental EROs compared to the operational (Fig. 10). Other than the NAM nest, there is no guidance from convection-allowing models (CAM) available to the operational WPC forecasters that covers all of the day 2 period. Furthermore, there is no operational CAM guidance that covers all of the day 3 period. However, the FFaIR Experiment participants were able to utilize the day 3 UM-Oper., FV3-CAPS, and FV3-GFDL CAMs in addition to the non-CAM FV3-GFS. Hence, the experimental CAM guidance may have increased forecaster confidence, resulting in more frequent issuances of higher risk ERO categories with larger areas. During the FFaIR Experiment, not all CAMs are guaranteed to run on any given day, but the availability of any CAM guidance may have contributed to the greater probabilistic improvement between the FFaIR and operational ERO forecasts on day 3 relative to those on day 2.

4. Discussion and conclusions

The 2017 Flash Flood and Intense Rainfall Experiment was designed to test emerging experimental products with the goal of improving heavy rain and flash

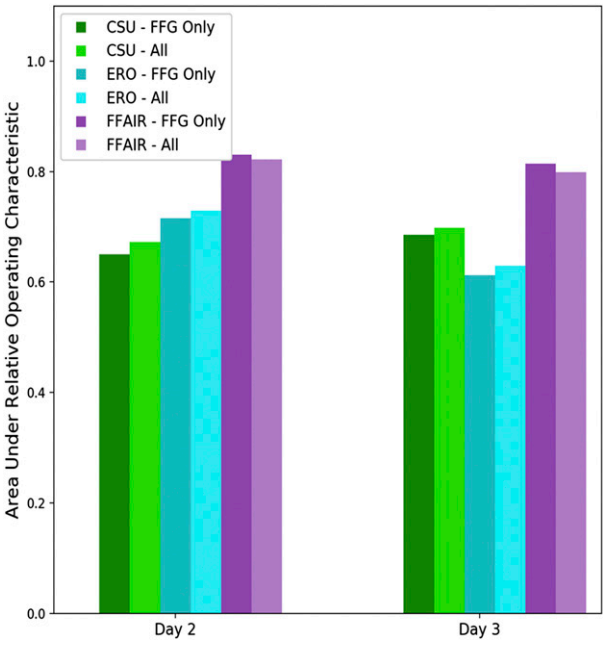

FIG. 10. Day 2 and day 3 AuROC values for the CSU (green), operational ERO (blue), and the FFaIR Experiment ERO (purple) verified against FFG (darker colors) and all flooding observations/proxies (lighter colors).

flooding forecasts in a collaborative pseudo-operational environment. Hence, the FFaIR Experiment provided a critical platform to evaluate products that may be transitioned from research to operations. This study focused on the objective verification of guidance (models, ensembles, and first-guess products) and forecasts issued during the FFaIR Experiment. Where applicable, the objective verification results were compared with the subjective rating given by the participants.

A simple evaluation of deterministic models and ensembles examined in the FFaIR Experiment were performed using Roebber performance diagrams (Roebber 2009). The Met Office Unified Model exhibited the best quantitative precipitation forecast (QPF) predictive skill in terms of critical success index for day 2 at all thresholds analyzed and day 3 at 1 in. (25.4 mm), while the FV3-GFS exhibited the best skill for day 3 at 0.5 in. (12.7 mm; Fig. 2). Subjectively, the HRRR-Exp model performed best on day 2 (Fig. 3a), and the Unified Model performed best on day 3 (Fig. 3b). In terms of ensemble guidance, the experimental Storm-Scale Ensemble Forecast generally had the lowest CSI, with comparable CSI values between the HRRRe and HREFv2 (Fig. 4). However, participants subjectively rated the HRRRe the best and the HREFv2 the lowest, albeit only with a small difference. (Fig. 3c). Discrepancies between subjective and objective verification scores were possible, since participants intuitively evaluated object-oriented biases. However, participants may have valued certain object-oriented attributes or geographical locations differently in comparison with others.

Operational and experimentally issued excessive rainfall outlooks were compared throughout the duration of the experiment. Both ERO products highlighted similar geographical regions (Figs. 5–7), with the experimental ERO exhibiting significantly greater fractional coverage (by at least a factor of 2) for all thresholds. It was possible that the experimental guidance increased forecaster confidence, resulting in the issuance of larger and more frequent higher risk categories in the FFaIR Experiment, which were more consistent with the coverage of flooding observations and proxies. The greatest instances of slight risk issuance occurred in the Southwest, portions of the Midwest, the central Appalachians, and along the central Gulf Coast. The ERO issuances in the Southwest United States were associated with the monsoon activity during the latter half of the FFaIR Experiment, while portions of the Midwest experienced several mesoscale convective systems in June and July. Most of the higher risk activity along the Gulf Coast is associated with Tropical Storm Cindy while the central Appalachians experienced a combination of MCS-induced flooding and the remnants of Cindy.

To assess the probabilistic calibration of the ERO, average fractional coverage of Stage IV precipitation exceeding flash flood guidance and all flooding observations and proxies (i.e., Stage IV exceeding FFG, Stage IV exceeding 5-yr average recurrence intervals, local storm reports, and river gauge observations) were included in the evaluation. In addition, the calibration of a new ERO first-guess field called the Colorado State University Machine-Learning Probabilities method was evaluated. In general, the operational and experimental EROs were well calibrated for all risk categories, except for moderate issuances of the operational ERO, which fell above the probabilistic definition. The CSU-MLP produced too many marginal instances and not enough slight, moderate, and high issuances.

Area under relative operating characteristic and Brier skill score were used to compare the skill of the experimental ERO, operational ERO, and CSU-MLP. The experimental ERO generally performed best, with the operational ERO exhibiting better skill than the CSU-MLP. Comparison of the EROs with the CSU-MLP was not necessarily fair since the CSU-MLP was used in the forecast process to produce the EROs. The improvement in the experimental ERO over the operational ERO was greatest on day 3, and it was possible that the increased availability of experimental convection-allowing models on day 3 contributed to this improvement.

One exception to the improvement of the experimental ERO over the operational ERO occurred for day 2 issuances during weeks 3 and 4, which were dominated by smaller-scale more weakly forced convective events than the first two weeks.

Several recommendations were made in the 2017 FFaIR Experiment final report (Perfater and Albright 2017). Relevant to the results presented in this study, it is recommended that CAMs be run to cover the entire day 3 period operationally (roughly out to forecast hour 72). Ideally, an ensemble of CAMs would be invoked in the future for this purpose to evaluate the probabilistic utility of an ensemble of QPF. The ensemble blended mean QPF's analyzed in the FFaIR Experiment were well received by participants, and there is justification for a transition of these products to operations.

The CSU-MLP was subjectively scored well by the participants in the FFaIR Experiment. The final FFaIR Experiment recommendation was that the CSU-MLP developers reduce some recurring spatial bias issues, particularly in New Mexico and Colorado, and reintroduce the technique in the 2018 FFaIR Experiment. The results of the CSU-MLP and other guidance products in the 2018 FFaIR Experiment will be discussed in a future paper.

Although not presented here, some experimental products were subjectively evaluated from the National Water Model. WPC recommended a more rigorous coupling of hydrologic and meteorological components to gather a more complete picture of the flash flooding paradigm. In the future, WPC hopes that a coupled probabilistic QPF-forced hydrologic output will become commonplace operationally, with new techniques to deduce probability of flooding and inundation from modeled streamflow.

Acknowledgments. The Weather Prediction Center portion of this work was supported by the U.S. Weather Research Program ''Probability of What?'' grant (NOAA-OAR-OWAQ-2015-2004230). Schumacher and Herman were supported by NOAA Joint Technology Transfer Initiative Award NA16OAR4590238. We thank three anonymous reviewers for their very helpful comments in improving this paper.

APPENDIX

List of Acronyms

ARI	Average recurrence interval
AuROC	Area under relative operating characteristic
BS	Brier score
BSS	Brier skill score
CAM	Convection-allowing model
CAPS	Center for Analysis and Prediction of Storms
CONUS	Contiguous United States
CSI	Critical success index
CSU	Colorado State University
EMC	Environmental Modeling Center
ERO	Excessive rainfall outlook
ESRL	Earth System Research Laboratory
FB	Frequency bias
FFaIR	Flash Flood and Intense Rainfall Experiment
FFG	Flash flood guidance
FV3	Finite-Volume Cubed-Sphere Dynamical Core
GFDL	Geophysical Fluid Dynamics Laboratory
GFS	Global Forecast System
GSD	Global Systems Division
HMT	Hydrometeorology Testbed
HREFv2	High-Resolution Ensemble Forecast, version 2
HRRR	High-Resolution Rapid Refresh
HRRRe	High-Resolution Rapid Refresh Ensemble
HRRR-Exp	HRRR experimental model
LSR	Local storm report
MCS	Mesoscale convective system
MLP	Machine-Learning Probabilities method
NAM	North American Mesoscale Forecast System
NOAA	National Oceanic and Atmospheric Administration
OU	University of Oklahoma
QPF	Quantitative precipitation forecasts
SSEFx	Storm-Scale Ensemble Forecast
UFVS	Unified Flooding Verification System
UM-Oper.	Unified Model Operational
WPC	Weather Prediction Center

REFERENCES

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.

Blake, B. T., J. R. Carley, T. I. Alcott, I. Jankov, M. E. Pyle, S. E. Perfater, and B. Albright, 2018: An adaptive approach for the calculation of ensemble gridpoint probabilities. *Wea. Forecasting*, **33**, 1063–1080, https://doi.org/10.1175/WAF-D-18-0035.1.

Brown, B. G., J. H. Gotway, R. Bullock, E. Gilleland, T. Fowler, D. Ahijevych, and T. Jensen, 2009: The Model Evaluation Tools (MET): Community tools for forecast evaluation. *25th Conf. on International Interactive Information and Processing*

Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Phoenix, AZ, Amer. Meteor. Soc., 9A.6, http://ams.confex.com/ams/pdfpapers/151349.pdf.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

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.

Cohen, S., S. Praskievicz, and D. R. Maidment, 2018: Featured collection introduction: National Water Model. *J. Amer. Water Resour. Assoc.*, **54**, 767–769, https://doi.org/10.1111/1752-1688.12664.

Ebert, E. E., 2001: Analysis of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Erickson, M. J., and J. A. Nelson, 2018: Verifying, calibrating, and redefining the excessive rainfall outlook at the Weather Prediction Center. *Eighth Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., 7.5, https://ams.confex.com/ams/98Annual/webprogram/Paper327404.html.

Fritsch, J. M., and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–964, https://doi.org/10.1175/BAMS-85-7-955.

Gochis, D. J., W. Yu, and D. N. Yates, 2015: The WRF-Hydro Model technical description and user's guide, version 1.0. NCAR Tech. Doc., 120 pp., https://www.ral.ucar.edu/projects/wrf_hydro.

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 Coauthors, 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.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

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**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

——, and ——, 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.

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

Li, J., Y. Chen, H. Wang, J. Qin, J. Li, and S. Chiao, 2017: Extending flood forecasting lead time in a large watershed by coupling WRF QPF with a distributed hydrological model. *Hydrol. Earth Syst. Sci.*, **21**, 1279–1294, https://doi.org/10.5194/hess-21-1279-2017.

Lin, S.-J., 1997: A finite-volume integration method for computing pressure gradient force in general vertical coordinates. *Quart. J. Roy. Meteor. Soc.*, **123**, 1749–1762, https://doi.org/10.1002/qj.49712354214.

——, and R. B. Rood, 1997: An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Quart. J. Roy. Meteor. Soc.*, **123**, 2477–2498, https://doi.org/10.1002/qj.49712354416.

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

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

NOAA, 2012: Hydrologic Services Program, definitions and general terminology. National Weather Service Manual 10-950, 5 pp., https://www.nws.noaa.gov/directives/sym/pd01009050curr.pdf.

Novak, D. R., 2017: The $1 billion dollar flood disasters of 2016. Major Weather Impacts of *2016*, Seattle, WA, Amer. Meteor. Soc., 2.2A, https://ams.confex.com/ams/97Annual/webprogram/Paper317351.html.

NWS, 2017: Summary of Natural Hazard Statistics in the United States. NOAA, accessed 15 October 2018, http://www.nws.noaa.gov/os/hazstats/sum17.pdf.

Perfater, S., and B. Albright, 2017: 2017 Flash Flood and Intense Rainfall Experiment. NOAA/NWS/WPC Weather Prediction Center Rep., 95 pp., https://www.wpc.ncep.noaa.gov/hmt/2017_FFaIR_final_report.pdf.

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

Schmidt, J. A., A. J. Anderson, and J. H. Paul, 2007: Spatially-variable, physically-derived flash flood guidance. *21st Conf. on Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 6B.2, https://ams.confex.com/ams/pdfpapers/120022.pdf.

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.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Yu, X., S. K. Park, Y. H. Lee, and Y. S. Choi, 2013: Quantitative precipitation forecast of a tropical cyclone through optimal parameter estimation in a convective parameterization. *SOLA*, **9**, 36–39, https://doi.org/10.2151/sola.2013-009.