Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011

ELLEN M. SUKOVICH

Cooperative Institute for Research in Environmental Sciences, and NOAA/Earth System Research Laboratory, Boulder, Colorado

F. MARTIN RALPH*

NOAA/Earth System Research Laboratory, Boulder, Colorado

FAYE E. BARTHOLD

I.M. Systems Group, Inc., and NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland

DAVID W. REYNOLDS

Cooperative Institute for Research in Environmental Sciences, and NOAA/Earth System Research Laboratory, Boulder, Colorado

DAVID R. NOVAK

NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland

(Manuscript received 31 May 2013, in final form 15 March 2014)

ABSTRACT

Extreme quantitative precipitation forecast (QPF) performance is baselined and analyzed by NOAA's Hydrometeorology Testbed (HMT) using 11 yr of 32-km gridded QPFs from NCEP's Weather Prediction Center (WPC). The analysis uses regional extreme precipitation thresholds, quantitatively defined as the 99th and 99.9th percentile precipitation values of all wet-site days from 2001 to 2011 for each River Forecast Center (RFC) region, to evaluate QPF performance at multiple lead times. Five verification metrics are used: probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), frequency bias, and conditional mean absolute error (MAE_{cond}). Results indicate that extreme QPFs have incrementally improved in forecast accuracy over the 11-yr period. Seasonal extreme QPFs show the highest skill during winter and the lowest skill during summer, although an increase in QPF skill is observed during September, most likely due to landfalling tropical systems. Seasonal extreme QPF skill decreases with increased lead time. Extreme QPF skill is higher over the western and northeastern RFCs and is lower over the central and southeastern RFC regions, likely due to the preponderance of convective events in the central and southeastern regions. This study extends the NOAA HMT study of regional extreme QPF performance in the western United States to include the contiguous United States and applies the regional assessment recommended therein. The method and framework applied here are readily applied to any gridded QPF dataset to define and verify extreme precipitation events.

1. Introduction

Extreme precipitation events (i.e., those events associated with the tail end of the precipitation probability distribution) are highly impactful and can cause loss of life, damage to property, and significant disruption to local, regional, and national economies. Extreme precipitation events can range from short and intense

* Current affiliation: Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California.

Corresponding author address: Ellen Sukovich, NOAA/Earth System Research Laboratory/Physical Sciences Division, 325 Broadway, Boulder, CO 80305.
E-mail: ellen.sukovich@noaa.gov

DOI: 10.1175/WAF-D-13-00061.1

periods of rainfall that result in flash flooding (e.g., mesoscale convective rainstorms) to prolonged periods of precipitation that result in the flooding of rivers and streams (e.g., landfalling atmospheric rivers); however, correctly forecasting extreme precipitation events remains one of the most difficult challenges in operational meteorology.

The demand for accurate quantitative precipitation forecasts (QPFs), particularly for extreme events, is echoed by many user communities [e.g., water resources management, industry, agriculture, transportation, government, emergency management; Ralph et al. (2005)] who require these forecasts for preparation, decision making, and management to effectively mitigate the subsequent impacts of extreme precipitation events. In fact, recent forecast-use studies have shown that precipitation forecasts are the most heavily utilized part of standard forecasts (Lazo et al. 2009). However, the forecast skill for these events is often insufficient to accurately and consistently predict the location, amount, type, and timing of precipitation (e.g., Olson et al. 1995; Cherubini et al. 2002; Charba et al. 2003; Ralph et al. 2003; Morss and Ralph 2007; Brennan et al. 2008; Ralph et al. 2010; Schumacher and Davis 2010; Novak et al. 2011).

A major example of the need for accurate extreme QPFs is the recent case of the Howard Hanson Dam flood risk management crisis (White et al. 2012). This crisis arose in 2009 when the Howard Hanson Dam, a flood-control dam above Seattle, Washington, showed signs of potential dam failure and was declared unsafe. The ensuing preparations for the following winter included identifying the heavy precipitation thresholds that would trigger the evacuation of a heavily populated region downstream of the dam and improving QPF products to more accurately predict those heavy precipitation thresholds at various lead times. Although insufficient time has passed to fully quantify the results and impacts of this effort, these improvements have allowed the National Weather Service (NWS) to provide much improved flood watches and warnings for this flood mitigation effort.

Improvement of QPFs, particularly extreme QPFs, is a high priority in meteorological research (e.g., National Weather Service 1999; Fritsch and Carbone 2004; Ralph et al. 2005). Accurate QPFs, particularly for extreme precipitation events, are a forecast challenge for both humans and numerical weather prediction (NWP) models due to the high temporal and spatial variability of precipitation (Ebert et al. 2003). While QPF improvements can be made in various ways (e.g., better data assimilation techniques, improved understanding of the dynamical and physical processes associated with heavy precipitation events, better physical parameterizations in the forecast models, etc.), a crucial component is forecast verification. QPF verification is critical in assessing forecast trends and biases, identifying forecast errors, monitoring forecast improvement, and providing users with information to most effectively use QPFs (Murphy and Winkler 1987; Jolliffe and Stephenson 2003).

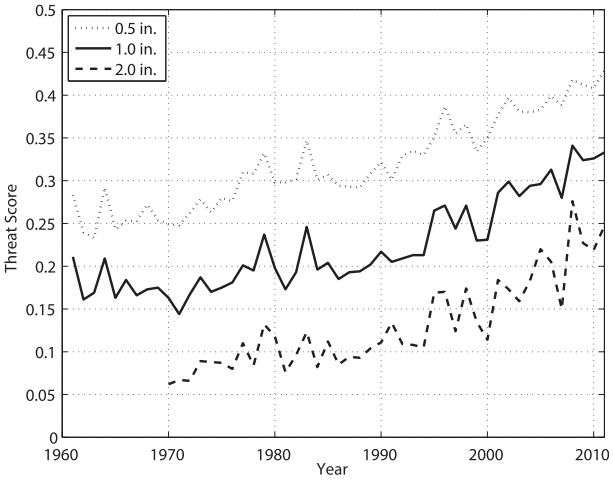

FIG. 1. Annual threat scores for the WPC's 0.50-in. (12.7 mm), 1.0-in. (25.4 mm), and 2.0-in. (50.8 mm) forecasts for day 1 from 1961 through 2011.

Since 1961, the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Prediction (NCEP) Weather Prediction Center (WPC) has monitored annual QPF performance using the threat score for various 24-h precipitation thresholds for day-1 (24 h) lead times over the contiguous United States (CONUS; Olson et al. 1995.) The threat score is a simple calculation that indicates relative forecast accuracy by measuring the degree of coincidence between forecast and observed events for a given precipitation threshold (Stanski et al. 1989). Over the last 50 yr, this metric has shown gradual but steady improvement over multiple thresholds for the day-1 precipitation forecasts created by the WPC and verified with a manually, quality-controlled quantitative precipitation estimation (QPE) product (Fig. 1; Anthes 1983; Olson et al. 1995). Today, the yearly QPF metric tracked by the Government Performance and Results Act (GPRA) of 1993 (www.whitehouse.gov/omb/mgmt-gpra/gplaw2m) is the 1.0 in. (25.4 mm) $(24\,\text{h})^{-1}$ threat score over the CONUS calculated from the WPC day-1 QPF product. While the threat score is useful in assessing relative forecast improvements (e.g., Reynolds 2003; Charba et al. 2003), this verification metric does not adequately assess the performance of extreme events, which occur less frequently and over smaller areas than lesser precipitation events. The threat score

metric is also sensitive to forecast hits, penalizes both misses and false alarms, and it does not distinguish the source of forecast errors. In addition, the threat score by itself (or any single metric for that matter) cannot provide a complete picture of forecasting successes and errors (Murphy and Winkler 1987; Doswell et al. 1990).

The Hydrometeorology Testbed (HMT; hmt.noaa.gov) was established in 2003 to address both the scientific and practical challenges associated with forecasting extreme precipitation to help improve operational forecasting skill (Ralph et al. 2005, 2013b). A key driver of HMT was the recognition that the current metric (i.e., threat score) and precipitation threshold [1.0 in. (25.4 mm) $(24\,\text{h})^{-1}$] used to assess annual national forecast skill is inadequate for monitoring and analyzing extreme QPF performance and does not represent regional variations.

Previous HMT work by Ralph et al. (2010) developed a method for monitoring and verifying extreme QPF performance by analyzing 4-km QPF data from the California–Nevada River Forecast Center (CNRFC) and the Northwest River Forecast Center (NWRFC) at 41 sites along the U.S. West Coast over the 2005–06 cool season. Using five verification metrics [probability of detection (POD), false alarm ratio (FAR), critical success index (CSI; i.e., threat score), frequency bias, and conditional mean absolute error (MAE_{cond})], Ralph et al. (2010) found that most of the extreme precipitation events [i.e., defined in that paper as events with precipitation ≥3.0 in. (76.2 mm) $(24\,\text{h})^{-1}$ and ≥5.0 in. (127.0 mm) $(24\,\text{h})^{-1}$] that occurred were underforecast, especially at longer lead times. In addition, the CNRFC region had significantly more extreme events, when defined by the ≥3.0 in. (76.2 mm) $(24\,\text{h})^{-1}$ and ≥5.0 in. (127.0 mm) $(24\,\text{h})^{-1}$ thresholds, than the NWRFC region. This led Ralph et al. (2010) to determine that one national threshold could not accurately define extreme precipitation events over different geographical areas. Based upon these results, Ralph et al. (2010) proposed monitoring and verifying extreme precipitation events over the CONUS by defining regionally relevant extreme precipitation thresholds by top percentiles (i.e., 1.0% and 0.1%) and then applying a simple verification framework using five performance metrics to ensure that the quality of the precipitation forecast is thoroughly assessed.

This paper represents an implementation of the Ralph et al. (2010) findings by using over a decade's worth of data rather than that of one winter season, by applying the Ralph et al. (2010) verification method nationally and regionally rather than to one test bed area, and by using gridded WPC QPF–QPE data rather than QPF–QPE from the CNRFC and NWRFC at 41 points. By applying the extreme QPF definitions and verification framework proposed by Ralph et al. (2010) to the 32-km WPC gridded operational QPF dataset over the CONUS, this paper benchmarks extreme QPF performance over an 11-yr period by evaluating the general change in extreme QPF performance over time. While the results presented here are not necessarily statistically significant due to the relatively short 11-yr analysis period, this work aims to verify and study the extreme QPF performance metrics necessary to quantify and improve the timing, location, and amount of predicted precipitation. Section 2 describes the QPF and QPE data used in this analysis. Section 3 provides an overview of the verification software and methodology used to analyze QPF performance, while section 4 objectively defines regional extreme precipitation thresholds for the data analyzed. Sections 5 and 6 present national and regional results, respectively, by lead time and season, and section 7 summarizes this research and describes future research and applications.

2. Data

The primary forecast and verification data evaluated in this study were 32-km gridded QPF and QPE data products obtained from the now-defunct National Precipitation Verification Unit (NPVU).[1] These data cover the CONUS for the period from 1 January 2001 through 31 December 2011, and this 11-yr period is studied because it was the longest continuous period of data available from the NPVU. All verification was performed on a 32-km Lambert conic conformal grid. Grid points with missing QPF and/or missing QPE data were eliminated from the dataset. All precipitation amounts (observed and predicted) examined in this paper are in English units with International System of Units (SI) in parentheses, since precipitation forecasts in the United States are commonly issued in units of inches.

a. Operational QPF product

Forecast data obtained from the NPVU are 6-hourly, 32-km QPF grids generated by the NCEP's WPC. The WPC is an NCEP service center that produces daily QPFs for the CONUS and has done so since 1960. These QPFs rely heavily on NWP guidance available from the suite of NCEP and international operational forecast models. Details regarding the WPC QPF process are described in Olson et al. (1995) and NWS (1999). The WPC issues gridded CONUS QPFs for 6-hourly intervals

[1] The NPVU was suspended as of 23 February 2012 due to budget constraints.

out to 84 h, which were archived at the NPVU until February 2012. All of the WPC QPF amounts represent spatially averaged precipitation amounts over a 32-km grid.

This study evaluates 24-h precipitation amounts accumulated from the 6-hourly QPF grids for the 12–36-, 36–60-, and 60–84-h forecast periods, hereafter called day 1, day 2, and day 3, respectively. Day-1 QPFs are analyzed from 2001 through 2011, while day-2 and day-3 QPFs are analyzed from 2003 through 2011. Prior to 2003, day-2 and day-3 QPFs were available from the NPVU only for western RFC regions [i.e., CNRFC, NWRFC and the Colorado basin RFC (CBRFC)] during the cool season (i.e., October–March); thus, for the national yearly QPF analysis, day-2 and day-3 QPFs were not verified before 2003.

b. *Verification data*

The QPE data used for verification in this study are 24-h accumulated precipitation amounts produced by the NWS RFCs on the 4-km Hydrologic Rainfall Analysis Project (HRAP) grid (stage IV). In the eastern United States, the RFC QPE data come from the stage III precipitation dataset, which is a blend of quality-controlled gauge, radar, and satellite data to obtain a composite best estimate of the precipitation distribution (Breidenbach et al. 1999; Fread et al. 1995; Fulton et al. 1998; McDonald and Graziano 2001.) In the western United States, RFC QPE data are obtained using precipitation gauge measurements that are gridded using Mountain Mapper (MM; Henkel and Peterson 1996), which converts gauge-point data to the 4-km HRAP grid using the Parameter–elevation Regressions on Independent Slopes Model (PRISM) climatology (Daly et al. 1994; McDonald and Graziano 2001). After each RFC gathers the appropriate stage III data, RFC Hydrometeorological Analysis and Support (HAS) forecasters manually quality control the QPE fields to remove gross errors and to ensure spatial contiguity in the precipitation fields (Antolik 2000). Each NWS RFC then sends its adjusted QPE product (now referred to as stage IV) on the 4-km HRAP grid to the NPVU. At the NPVU, the 4-km QPE grid was upscaled to a 32-km grid using spatial averaging, where spatial averaging is a simple average of all of the original grid points within the area of the output grid. Further details on the NPVU QPE product can be found in McDonald and Graziano (2001) and Antolik (2000).

c. *WPC-calculated threat scores*

Since 1961, the WPC has used the threat score (i.e., CSI) as a metric to monitor the progress and quality of the WPC QPFs (Olson et al. 1995). With an in-house, consistently maintained threat score verification scheme, the WPC has archived over 50 yr of annual day-1 threat scores for precipitation events exceeding 0.5 in. (12.7 mm) and 1.0 in. (25.4 mm) $(24\,\text{h})^{-1}$, and over 40 yr of annual day-1 yearly threat scores for precipitation events exceeding 2.0 in. (50.8 mm) $(24\,\text{h})^{-1}$ (Fig. 1).

WPC forecasters compute annual threat scores by verifying WPC QPFs with human quality-controlled quantitative precipitation analyses. These analyses are based on a first-guess field from either the multisensor stage IV QPE mosaic analyses (Lin and Mitchell 2005) or the daily Climate Prediction Center (CPC) precipitation analysis (http://www.cpc.ncep.noaa.gov/products/Global_Monsoons/gl_obs.shtml). The WPC forecaster manually performs a quality control routine on the first-guess field by using additional data sources, such as radar precipitation estimates, Community Collaborative Rain, Hail and Snow Network (CoCoRaHS) observations, Cooperative Observer Network (COOP) data, and aviation routine weather report (METAR) observations. When the WPC forecaster is satisfied with the analysis, both the WPC gridded QPFs and the manually quality-controlled precipitation analysis (e.g., QPE) are mapped onto the 32-km HRAP grid and threat scores are computed. The WPC QPFs are verified for amounts equal to or greater than 0.5 in. (12.7 mm) $(24\,\text{h})^{-1}$ with amounts greater than 6.0 in. (152.4 mm) $(24\,\text{h})^{-1}$ rare since WPC QPFs represent 32-km spatial averages.

3. Verification methodology

In this study five metrics are calculated and evaluated: POD, FAR, CSI, frequency bias, and MAE_{cond}. These metrics are utilized because they are simple, easily understood scores with which most operational forecasters are familiar, particularly WPC forecasters. In addition, since 1) the CSI is the yearly GPRA standard and 2) the WPC has been monitoring the CSI for over 50 yr, continued analysis of the CSI, with the added information provided by the POD, FAR, frequency bias, and MAE_{cond}, allows for WPC QPF performance to be relatively evaluated by lead time, season, and region.

a. *Verification metrics*

Performance measures are verification metrics that focus on the correspondence between the forecasts and observations (Murphy 1993). In this study, an event is defined as a 24-h period (1200–1200 UTC) when the predicted P and/or observed O accumulated precipitation at a 32-km grid point matches or exceeds a specified precipitation threshold. When a grid point has both an observed and predicted event, a hit H occurs. When a grid point has a predicted event that is not observed,

TABLE 1. Contingency table of the four possible outcomes for categorical forecasts of a binary (yes–no) event. NO = not observed; NP = not predicted; CR = correct rejection.

Events	*O*	NO
P	*H*	*F*
NP	*M*	CR

a false alarm F occurs, and when a grid point has an observed event that is not predicted, a miss M occurs. Table 1 provides a summary of these variables displayed in a 2×2 contingency table.

The POD is the ratio of the number of correct forecasts H to the number of observed events $(H + M)$ and the FAR is the ratio of the number of false alarms F to the number of forecasts made $(H + F)$. The CSI (i.e., threat score) is the ratio of the correct forecasts H to all events either forecast or observed $(H + M + F)$. All three metrics range from 0 to 1, with 1 being perfect POD and CSI scores, and 0 being a perfect FAR score. Note that a perfect forecast system would produce only hits with no false alarms or misses.

To characterize the accuracy of QPFs (i.e., how close the forecasts, or predictions, are to the eventual outcomes), the $\mathrm{MAE}_{\mathrm{cond}}$, which is the average of the absolute errors, is utilized. In this study, the $\mathrm{MAE}_{\mathrm{cond}}$ is computed conditionally for QPE (observed) values greater than a specified precipitation threshold to answer the question: For the extreme events that were observed, how accurate is the forecasted amount of precipitation? Information about the variance of the forecast from the observed precipitation amount is given by $\mathrm{MAE}_{\mathrm{cond}}$, which does not indicate the direction of that error. However, the smaller the $\mathrm{MAE}_{\mathrm{cond}}$, the more accurate the QPF.

To understand how the QPFs overforecast–underforecast precipitation, the frequency bias (hereafter referred to as bias) is calculated from the ratio of the number of predicted events to observed events for all events (both observed and predicted) exceeding a given threshold. The bias ranges from 0 to infinity, with a bias equal to one being an unbiased forecast. By calculating the bias for all events, this metric can aid in identifying overforecasting and underforecasting properties of QPFs, where overforecasting occurs when the bias exceeds one and underforecasting occurs when the bias is less than one.

It is important to note that all five metrics, when analyzed individually, have limitations. The POD is sensitive to hits, but it ignores false alarms. This makes it a good measure for rare events, but it can be artificially improved by just issuing more "yes" forecasts to increase the number of hits. The FAR is sensitive to false alarms, but it ignores misses. Like the POD, it is very sensitive to the climatological frequency of the event (number of hits), and therefore it should be used in conjunction with the POD for optimal QPF performance feedback. The CSI is sensitive to hits and it penalizes both misses and false alarms. Rare events do not do well with this score as the CSI depends on the climatological frequency of events, thus resulting in poorer scores for rarer events (Mason 1989).

Although it is beyond the scope and aim of this paper, it should be noted that recent progress has been made in the verification of extreme events (e.g., Casati et al. 2008; Ebert et al. 2013). In particular, new verification metrics have been developed and are being tested, such as the extreme dependency score (EDS; Stephenson et al. 2008; Ghelli and Primo 2009), the stable extreme dependency score (SEDS; Hogan et al. 2009), the extremal dependence indices (EDIs; Ferro and Stephenson 2011), and the symmetric extremal dependence index (SEDI; Ferro and Stephenson 2011).

b. *Verification software*

Verifying large gridded datasets over extended periods of time, such as the 11 yr of data evaluated in this study, cannot be completed manually with any efficiency; thus, verification software is required. Model Evaluation Tools (MET) is a verification software package developed and maintained by the Developmental Testbed Center (DTC) for use by the meteorological community to help in assessing and evaluating the performance of numerical weather predictions (www.dtcenter.org/met/users/). MET is a highly configurable, state-of-the-art set of evaluation tools that can be applied to any gridded product. By comparing gridded forecast data to similarly gridded observational data, MET can be configured to calculate standard verification scores.

To ensure that the MET verification software outputs verification results similar to the WPC-calculated verification scores (Fig. 1), yearly threat scores were calculated for 2001–11 using the MET software on the NPVU QPF and QPE datasets described in section 2. Figure 2 shows the MET-calculated threat scores overlaid with the WPC-calculated threat scores for the same 24-h precipitation thresholds used in Fig. 1: 0.5 in. (12.7 mm), 1.0 in. (25.4 mm), and 2.0 in. (50.8 mm). The shape of the MET time series is consistent with the shape of the WPC time series, although the MET scores are slightly lower for both the 0.5 in. (12.7 mm) and 1.0 in. (25.4 mm) thresholds and vary (both higher and lower) for the 2.0 in. (50.8 mm) threshold. The WPC and MET curves are not exactly the same because the QPE product used internally by the WPC differs from the NPVU QPE product used in this study. As described in section 2c, the

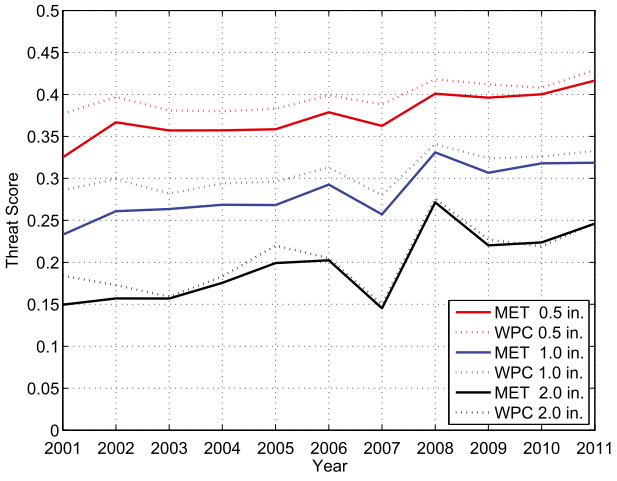

FIG. 2. Annual WPC-derived (WPC; dotted line) and MET-derived (MET; solid line) threat scores for 0.50 in. (12.7 mm; red), 1.0 in. (25.4 mm; blue), and 2.0 in. (50.8 mm; black) for day-1 precipitation forecasts from 2001 through 2011.

WPC QPE product is manually quality controlled. In addition, WPC QPE values do not exceed 6 in. (152.4 mm) $(24\,\text{h})^{-1}$, whereas the MET analysis allows for the verification of forecasts exceeding 6.0 in. (152.4 mm) $(24\,\text{h})^{-1}$. Finally, the WPC QPF verification methodology is different from the MET verification software. Fortunately, this study is focused on the incremental and relative improvement of only one forecast product over a period of time so although the MET-derived and WPC-calculated threat scores are not identical, the MET software does represent the WPC threat scores adequately enough to be used in this study to calculate other performance metrics (e.g., POD, FAR, bias, and MAE_{cond}) and analyze extreme QPF performance.

4. Extreme precipitation events

Extreme precipitation events can be defined in many ways; however, currently there is no accepted standard for defining such an event. To better quantify extreme precipitation events, this paper defines these cases as those events that exceed the 99th (top 1.0% events) or the 99.9th (top 0.1% events) percentile values of all precipitation events in a 24-h period.

a. Regional extreme precipitation thresholds

Due to the varying climatology in different regions of the CONUS (e.g., landfalling tropical storms in the Southeast, atmospheric rivers on the West Coast, mesoscale convective systems in the central plains, and monsoons in the Southwest), it follows that extreme event thresholds will vary by region. Per the methodology of Ralph et al. (2010), regional thresholds for the top 1.0% and top 0.1% events were calculated over each RFC area for the gridded 32-km QPE dataset over the 11-yr period (Fig. 3).

To calculate these extreme event thresholds, all observed wet-site days [i.e., grid points with precipitation

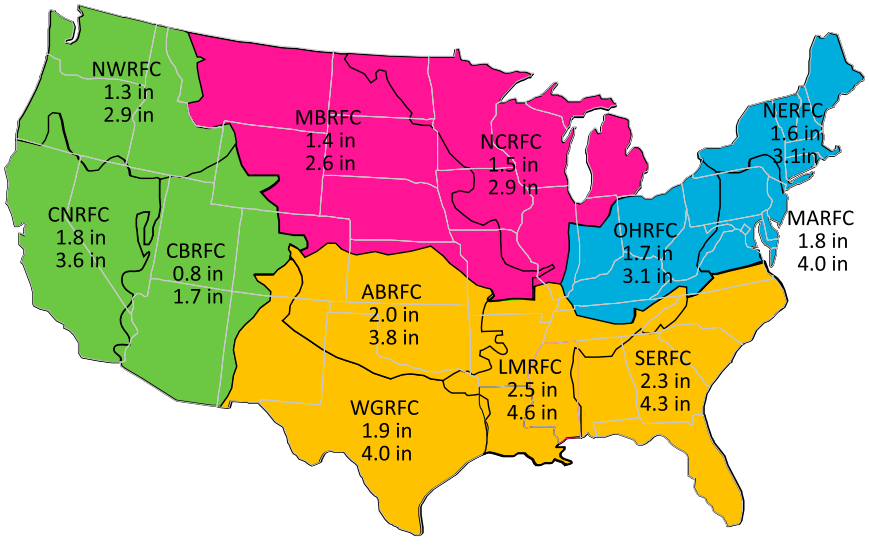

FIG. 3. CONUS NWS RFC regional thresholds for daily precipitation amounts calculated from the NPVU's 32-km gridded stage IV QPE data. In each RFC region, the top number is the threshold of the top 1.0% of precipitation events and the bottom number the threshold of the top 0.1% of precipitation events over the period from 2001 to 2011 (in.). RFC regions are color coded into four broad U.S. geographical regions: West (green), upper Midwest (red), South/Southeast (yellow), and East/Northeast (blue).

TABLE 2. Total number of wet-site days, or events, and total number of grid points by RFC region. Regional extreme precipitation thresholds are derived from 32-km gridded daily precipitation amounts as a function of RFC domain within the contiguous United States.

RFC	No. of wet-site days	No. of 32-km grid points	32-km grid 1.0% thresholds [in. (mm) (24 h)$^{-1}$]	32-km grid 0.1% thresholds [in. (mm) (24 h)$^{-1}$]
Arkansas-Red basin (AB)	769 379	539	2.0 (50.8)	3.8 (96.5)
Colorado basin (CB)	871 622	790	0.8 (20.3)	1.7 (43.2)
California–Nevada (CN)	571 051	641	1.8 (45.7)	3.6 (91.4)
Lower Mississippi (LM)	977 437	540	2.5 (63.5)	4.6 (116.8)
Middle Atlantic (MA)	401 716	192	1.8 (45.7)	4.0 (101.6)
Missouri basin (MB)	2 041 954	1312	1.4 (35.6)	2.6 (66.0)
North Central (NC)	1 483 737	837	1.5 (38.1)	2.9 (73.7)
Northeast (NE)	590 089	258	1.6 (40.6)	3.1 (78.7)
Northwest (NW)	1 438 307	777	1.3 (33.0)	2.9 (73.7)
Ohio (OH)	899 132	446	1.7 (43.2)	3.1 (78.7)
Southeast (SE)	1 258 745	659	2.3 (58.4)	4.3 (109.2)
West Gulf (WG)	1 272 814	1093	1.9 (48.3)	4.0 (101.6)
National sum (contiguous)	12 575 983	8084		

values greater than 0.0 in. (0.0 mm) (24 h)$^{-1}$] within an RFC region were collected over the 11-yr period and sorted from largest to smallest precipitation amount. Table 2 lists the number of wet-site days for each region. The 99th and 99.9th percentile values were then designated to be the extreme thresholds for the top 1.0% and the top 0.1% observed extreme events, respectively, for that region (Table 2).

It should be noted that since the extreme thresholds depend on both the resolution of the QPE dataset and the time period the QPE dataset covers, the top 1.0% and 0.1% thresholds for extreme events will vary by dataset (i.e., the regional extreme thresholds derived here should not be necessarily used for extreme QPF verification of other datasets.) Although the top 1.0% and the top 0.1% extreme event thresholds are unique to this dataset, the method for identifying extreme thresholds can be applied to any dataset.

b. *Frequency of extreme precipitation events*

As previously mentioned, each grid point with a QPE value exceeding 0.0 in. (0.0 mm) over a 24-h period within an RFC region is considered a wet-site day, or an event. The number of events per RFC region depends on the size of the RFC domain (e.g., the number of grid points that fall within the RFC region) and on the local climatological variations of the RFC region. In total, for this study, 8084 grid points were analyzed and more than 12.5 million events were recorded over the CONUS for the 11-yr period.

Figure 4 shows the distribution of the number of extreme precipitation events (top 1.0% and 0.1% events) and the extreme event frequency (i.e., fraction of all wet-site days) of all the RFC regions aggregated by year. For the top 1.0% events, there were minimally 10 000 events per year (Fig. 4a); for the top 0.1% events there were at least 700 events per year (Fig. 4b). Most of the peaks in the number of extreme events correspond to active or above normal Atlantic hurricane seasons (e.g., 2004, 2008, and 2010). In 2002, however, although there were an above normal number of named tropical storms, most of these storms were weak and short lived (Pasch et al. 2004). Instead, the peak in the number of extreme events during 2002 is most likely related to significant heavy rain and flooding events in the upper Midwest (Waple and Lawrimore 2003) and Texas (Nielsen-Gammon et al. 2005; Lowrey and Yang 2008; Zhang et al. 2006).

5. National extreme QPF performance results

Since 24-h QPFs are the commonly accepted standard in QPF verification programs (Olson et al. 1995) and since the 24-h time period generally provides the densest network of precipitation observations, the verification results presented here are for 24-h QPFs. It should be noted, however, that 24-h precipitation periods may not be the most relevant periods for hydrological model improvement, particularly for RFC models that use 6-h QPF amounts to drive their operational hydrological models. In addition, previous studies have found that longer-lasting events (e.g., >24 h) are key to creating floods in many major watersheds (Ralph and Dettinger 2012; Moore et al. 2012; Ralph et al. 2013a). Section 7 describes future work to expand this study to include both shorter precipitation intervals (e.g., 1, 3, and 6 h) aimed at assessing and improving hydrological runoff models and longer precipitation periods (e.g., 72 h) aimed at addressing long-lasting flood event prediction.

Annual national performance metrics are calculated for the top 1.0% and 0.1% extreme precipitation events

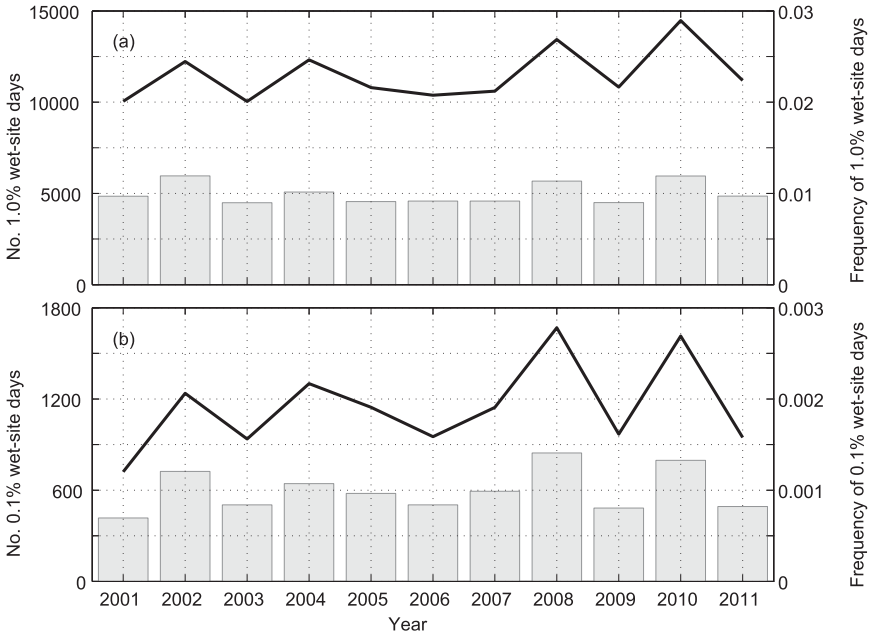

FIG. 4. Annual number of wet-site days (line) and frequency of wet-site days (bars) greater than or equal to the top (a) 1.0% and (b) 0.1% precipitation events from 2001 through 2011.

by applying the regional extreme precipitation thresholds defined in section 4a (Fig. 3; Table 2) and calculating the verification output (i.e., hits, misses, false alarms, and matched forecast–observation pairs) at each 32-km grid point in the WPC QPF dataset, aggregating the verification output by year over the CONUS, and then calculating the performance metrics of interest. The same procedure is also applied to events exceeding 1.0 in. (25.4 mm) over a 24-h time period to put the extreme event verification results in context with the current verification threshold in operational practice.

a. Yearly day-1 extreme QPF performance

Figure 5 displays the annual verification scores of the five metrics of interest for three precipitation thresholds [i.e., 1.0 in. (25.4 mm), top 1.0%, and top 0.1% events]. Bootstrapped 95% confidence intervals are computed from contingency tables for the POD, FAR, CSI, and bias. Beginning in 2001 and ending in 2011, the overall annual accuracy (i.e., CSI values) of the QPFs improved for all three thresholds, although the annual CSI values decreased with increased precipitation threshold (Fig. 5c).

Examination of annual POD and FAR values provides information about why the accuracy of the QPFs has increased from 2001 to 2011 (Figs. 5a,b). Figure 5a shows overall improvement in the annual POD scores for all three thresholds between 2001 and 2011, meaning that the hit rate (i.e., the proportion of events correctly forecast) increased for each precipitation threshold. While an increased hit rate can be accounted for by more events being correctly forecast, it can also increase when more events are predicted but do not occur (i.e., more false alarms occur). However, examination of the annual FAR scores (Fig. 5b) shows that over the same period, the FARs decreased, indicating that the improvement seen in the annual PODs is the result of more accurate QPFs rather than just more forecasts being made. This implies that the improvement in accuracy observed over the 11-yr period is real and not an artifact of more forecasts being issued or the verification metric chosen.

Examination of the annual biases by precipitation threshold (Fig. 5d) shows that the top 1.0% and 0.1% events have smaller annual bias values compared to the 1.0-in. events, which are just above and below a bias of one. The annual bias decreases (i.e., becomes more underforecast) with increased precipitation threshold, indicating that the most extreme events tend to be more underforecast than events of lesser precipitation magnitude. This result is similar to the results observed by Charba et al. (2003). In terms of improvement over the 11-yr period, 1.0% and 0.1% biases vary from year to year. The 1.0-in. events do not show marked improvement, fluctuating around one over the analysis period.

The annual MAE_{cond} values for precipitation events exceeding the 1.0 in. (25.4 mm), 1.0%, and 0.1% thresholds are shown in Fig. 5e. As previously stated, the MAE_{cond} is calculated conditionally for observed events that equal or exceed the specified threshold, meaning that forecasts

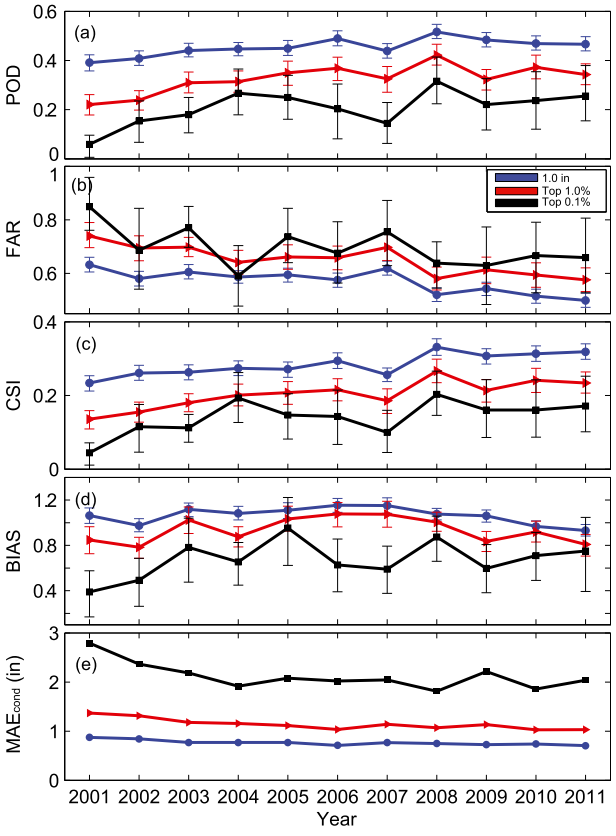

FIG. 5. Yearly day-1 (a) POD, (b) FAR, (c) CSI, (d) bias, and (e) MAE_{cond} values for the CONUS from 2001 to 2011 using the regional precipitation thresholds in Table 2. Blue, red, and black lines indicate events exceeding 1.0 in. (25.4 mm) (24 h)$^{-1}$, the top 1.0%, and the top 0.1% of all precipitation events, respectively. Brackets indicate 95% confidence intervals for the skill scores and bias.

included in the MAE_{cond} calculations have associated observed precipitation amounts equal to or less than the thresholds. As expected, the MAE_{cond} values are larger for larger precipitation thresholds although all three precipitation thresholds have decreasing errors over the analysis period.

b. *Yearly day-2 and day-3 extreme QPF performance*

Extreme QPF performance is further analyzed by examining day-2 and day-3 QPFs with the same verification metrics as for the day-1 QPFs. Although the longer lead times have similar patterns and trends as the day-1 QPFs, QPF performance decreases with increased lead time and there is more variability. It should be noted that the shorter analysis period (9 yr) of the day-2 and day-3 QPFs limits even further the statistical significance of defining a definite upward or downward trend to the data; however, looking at these results can provide preliminary information about extreme QPF performance at lead times longer than day 1.

Figure 6 shows the yearly performance metrics for the three lead times: day 1, day 2, and day 3 (i.e., 24, 48, and 72 h), and bootstrapped 95% confidence intervals are included for POD, FAR, CSI, and bias. For the top 1.0% QPFs, day 1 has the highest CSI values followed by day 2 and day 3, respectively (Fig. 6e). The top 0.1% events also follow this trend (Fig. 6f). The annual PODs show similar trends (Figs. 6a,b), but the yearly FARs are much more variable by lead time (Figs. 6c,d). Although the day-1 FARs have the lowest values (i.e., the better performance), there is no distinct delineation of whether the day-2 performance improves over the day-3 forecast during the 9-yr period. In addition, during the years 2009–10, the FARs for the top 0.1% events are essentially the same on day 1 as they were on day 3, showing little to no improvement with lead time. This could imply that perhaps the slight improvements seen in the POD values over the same time period are the result of overforecasting rather than better and more accurate precipitation forecasts; however, the 9-yr sample size is too small to make a definitive argument.

The bias scores (Figs. 6g,h) show that extreme events are underforecast for all lead times. Figure 6g shows that for the top 1.0% events, the day-1 annual bias is at or just below 1 (unbiased) but the day-2 and day-3 QPFs tend to be underforecast. The top 0.1% events have much more variability in their annual biases but again, the day-1 biases tend to be closer to one than the day-2 or day-3 annual biases (Fig. 6h). Meanwhile, the yearly MAE_{cond} values (Figs. 6i,j) increase with increasing lead time. The error for the top 1.0% events is approximately half of the error of the top 0.1% events. Over the 11-yr period, the top 1.0% events appear to have decreased in error by 0.25–0.5 in. (6.4–12.7 mm) while the MAE_{cond} values of the top 0.1% events have decreased by 0.5–0.75 in. (12.7–19.1 mm).

It should be noted that there are obvious improvements in the POD, FAR, CSI, and bias scores for 2008 and some improvements in 2005. These improvements may be related to strong Atlantic hurricane seasons. The hurricane season of 2005 is considered to be the most active Atlantic hurricane season ever (Beven et al. 2008), and the 2008 hurricane season was above the long-term mean (Brown et al. 2010).

c. *Seasonal national extreme QPF performance*

As noted by many studies (e.g., Junker et al. 1989; Olson et al. 1995; Fritsch and Carbone 2004), QPF performance and skill vary significantly by season. Generally, the highest CSIs occur during the cool season and the lowest CSIs occur during the warm season (Olson et al. 1995). To analyze seasonal variations in extreme QPF performance, 24-h extreme precipitation events were stratified by month. The top 1.0% and top 0.1% thresholds were

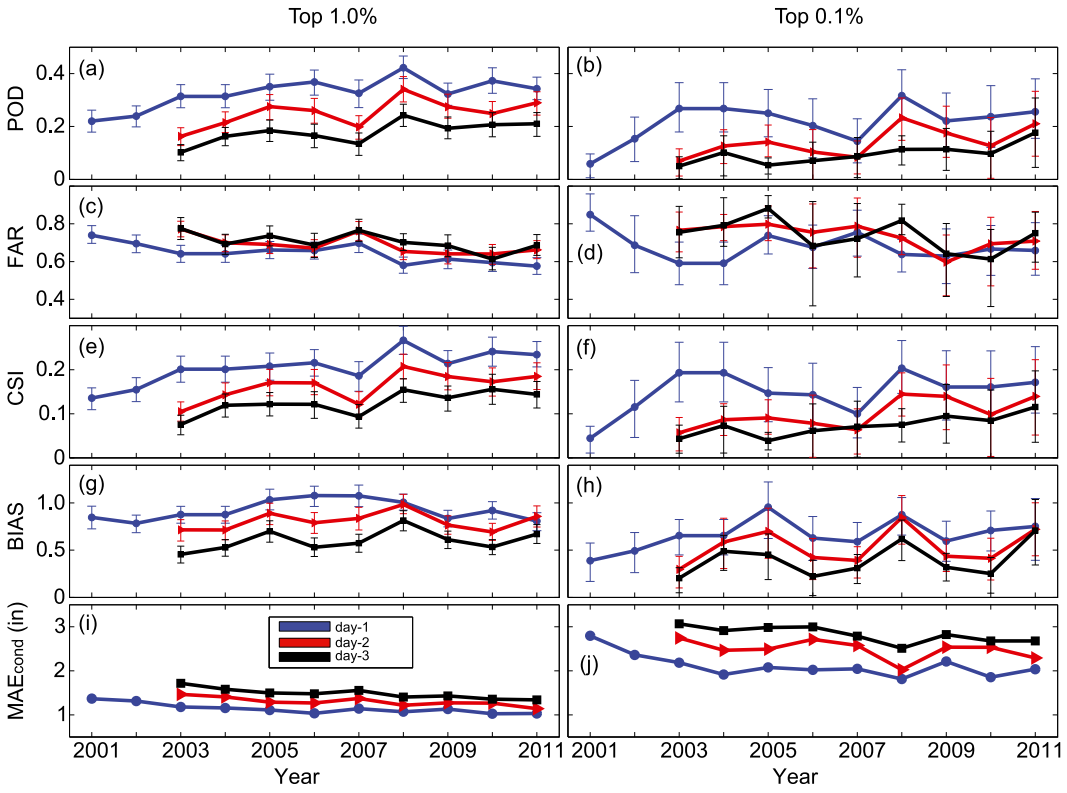

FIG. 6. Yearly (a) POD, (c) FAR, (e) CSI, (g) bias, and (i) MAE_{cond} values for the CONUS from 2001 to 2011 using the regional precipitation thresholds in Table 2 for the top 1.0% of precipitation events. Yearly (b) POD, (d) FAR, (f) CSI, (h) bias, and (j) MAE_{cond} values for the CONUS from 2001 to 2011 using the regional precipitation thresholds in Table 2 for the top 0.1% of precipitation events. Blue, red, and black lines indicate the day-1–3 forecast intervals, respectively. (Note that data for days 2 and 3 exist only from 2003 to 2011.)

applied to each RFC region and Fig. 7 shows the aggregated monthly frequency of these events over the 11-yr period. Over the calendar year, the number of extreme events peaks (for both the top 1.0% and top 0.1% thresholds) in September and is at a minimum in February.

As expected, the CSIs were higher during the cool season and lower during the warm season (Figs. 8e,f), even though more extreme events tend to occur during the warm season months of June–August (Fig. 7). This seasonal difference in CSIs can be attributed to the scale and type of precipitation events during each season, where summer events tend to be dominated by small-scale processes (i.e., convection) and winter events tend to be dominated by more synoptic-scale events (e.g., atmospheric rivers). The slight peak in the CSI values for days 1 and 2 during September is probably related to tropical cyclones that make landfall in the southeastern United States. These events are synoptically driven and tend to be better forecast than the smaller-scale convection.

The monthly POD and FAR scores by lead time are shown in Figs. 8a–d. For the top 1.0% events, the PODs are lower during the warm season and higher during the cool season, with an increase in the September POD scores for all three lead times (Fig. 8a). The FARs are at a maximum during the summer months and are lower during the winter months (Fig. 8b). Although it appears that there is a FAR local maxima in the 0.1% events during February (Fig. 8g), the 95% confidence interval plus the sample distribution in Fig. 7 indicate that this maxima is likely the result of a small sample size. For the 1.0% events, while the day-1 FAR is better for all months, the day-2 FAR is not always better than the day-3 FAR (e.g., July and November). For the top 0.1% events, the PODs have a similar pattern to the top 1.0% events; however, for the month of June there is essentially no POD and the FAR is almost one, indicating that there is essentially no skill for the approximately 1000 events that occurred during this month (Figs. 8b,d).

The monthly MAE_{cond} values have larger values (more error) during the summer months and lower error during the winter months (Figs. 8i,j). The MAE_{cond} values increase with increasing lead time for both the top 1.0% and top 0.1% events. Monthly biases decrease (e.g.,

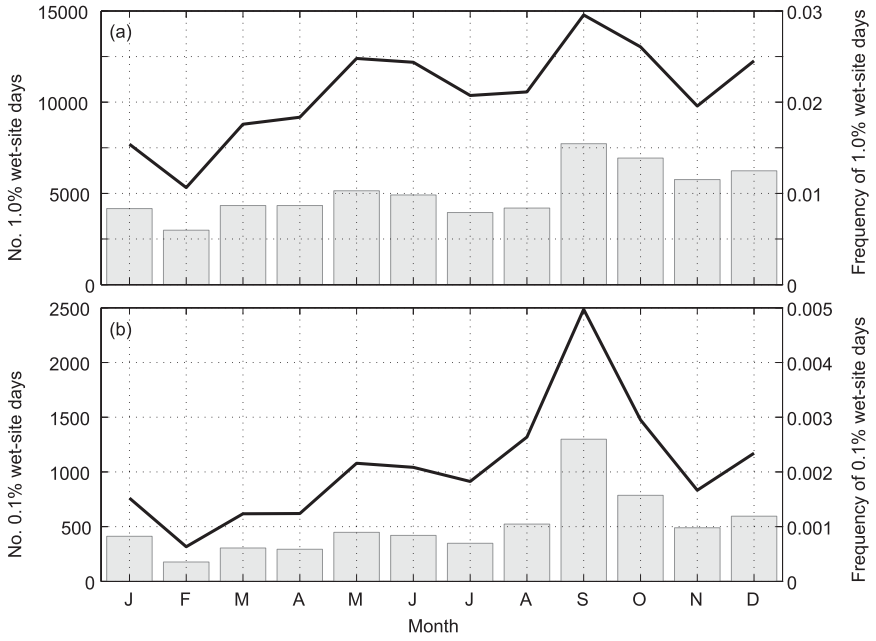

FIG. 7. Monthly number of wet-site days (line) and frequency of wet-site days (bars) greater than or equal to the top (a) 1.0% and (b) 0.1% precipitation events aggregated over 2001–11.

underforecast) with increasing lead time (Figs. 8g,h). Seasonally, more underforecasting occurs during the warm season (June–August) than the cool season (December–February). The underforecasting that occurs in the warm season is due in part to the small-scale convective nature of many of these extreme events. Numerical model guidance often struggles to accurately predict both the location and the magnitude of convectively driven warm season precipitation (Fritsch and Carbone 2004). Location errors alone for heavy precipitation associated with MCSs average around 250 km in NCEP models, even within 36 h of an event (http://www.comet.ucar.edu/outreach/abstract_final/1083387_TAMU.PDF). In scattered thunderstorms, short distances (e.g., <40 km) can mean the difference between significant rainfall and sunny skies. The small spatial scales and high uncertainty make these events particularly challenging for both numerical model guidance and WPC forecasts.

6. Regional results

The extreme precipitation verification metrics presented thus far have been summary values for the 12 CONUS RFC areas combined; thus, the differences between geographical regions are lost, being lumped into one national dataset of annual POD, FAR, CSI, bias, and MAE_{cond}. As stated previously, different regions of the United States experience different meteorological conditions and phenomena, and these phenomena occur at different times throughout the year. It is useful to examine the verification scores of each RFC region separately, both annually and by season, to better identify geographic regions, and thus meteorological conditions, over which the WPC forecasts are better or worse over time. In addition, it is important to recognize that improvements to the overall national verification metrics, which may be monitored and evaluated similar to the current GPRA threat score, can be made through smaller improvements by RFC region; however, methods for improvement in one region may not work for another.

The following sections examine the top 1.0% and 0.1% events for each CONUS RFC domain over a 5-yr period (2007–11), color-coded into four broad U.S. geographical regions: West (green), upper Midwest (red), South/Southeast (yellow), and East/Northeast (blue). The 5-yr period was chosen not only to maximize the sample size of the extreme events within an RFC region, but also to ensure that the analysis period was not too long such that significant changes in RFC QPF performance occurred.

a. Overall regional results

The WPC's QPF performance for the top 1.0% precipitation events shows that the highest POD and CSI scores plus the lowest FAR values are over the western United States (i.e., the CNRFC, NWRFC, and CBRFC regions), followed closely by the East/Northeast U.S. [i.e., the Middle Atlantic RFC (MARFC) and the Northeast RFC (NERFC)] RFC areas (Figs. 9a,c,e). These better performance values are likely the result of the dominance of synoptically driven systems in these

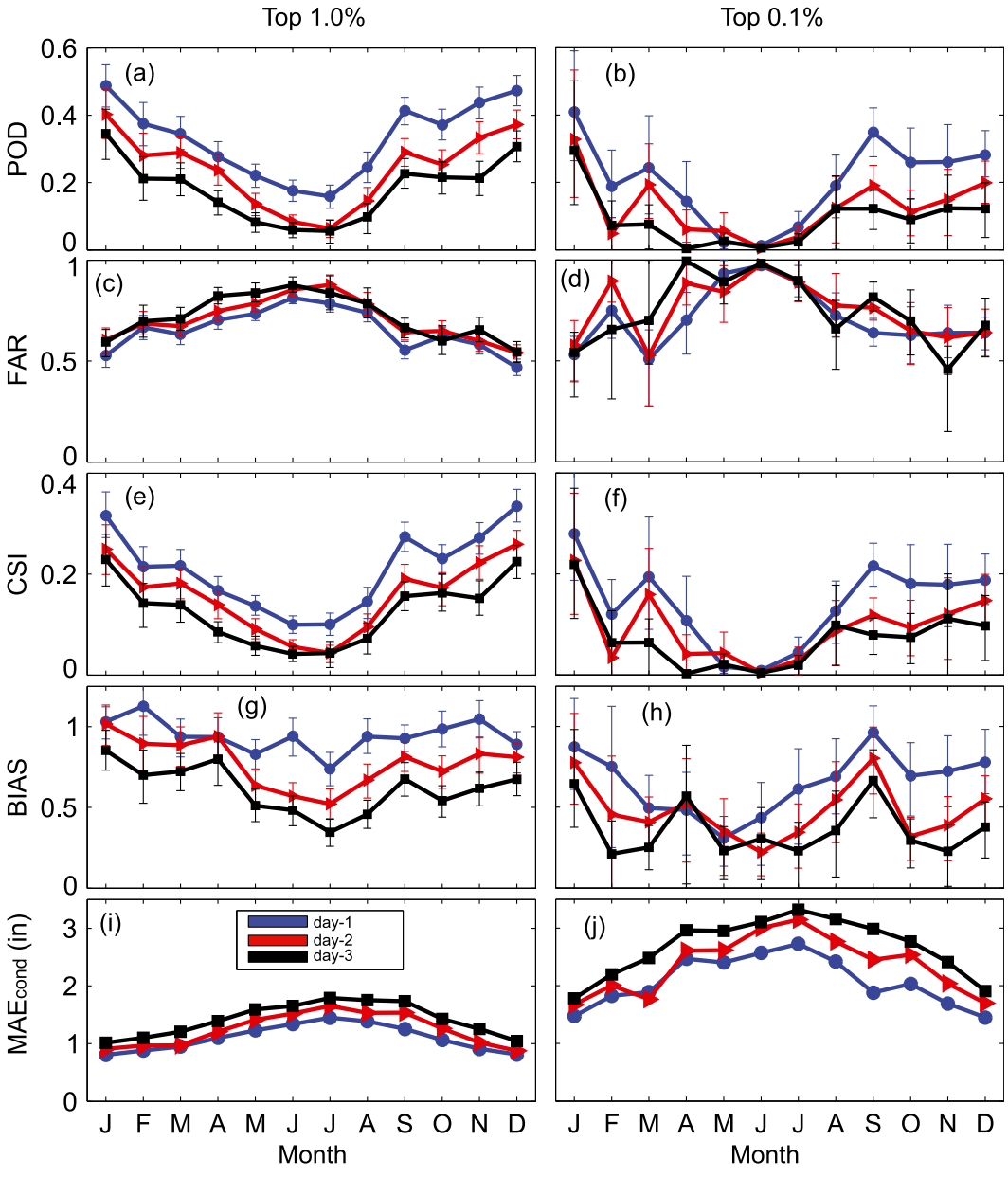

FIG. 8. Monthly (a) POD, (c) FAR, (e) CSI, (g) bias, and (i) MAE_{cond} values for the CONUS from 2001 to 2011 using the regional precipitation thresholds in Table 2 for the top 1.0% of precipitation events. Monthly (b) POD, (d) FAR, (f) CSI, (h) bias, and (j) MAE_{cond} values for the CONUS from 2001 to 2011 using the regional precipitation thresholds in Table 2 for the top 0.1% of precipitation events. Blue, red, and black lines indicate the day-1–3 forecast intervals, respectively. Brackets indicate 95% confidence intervals for the skill scores and bias.

areas (particularly atmospheric rivers) plus the orography in the western United States. In contrast, the upper Midwest and South/Southeast regions, which are prone to smaller, convective-scale extreme events, have lower POD and CSI values but higher FAR values. A similar QPF performance pattern emerges for the 0.1% events; however, the East/Northeast RFCs (MARFC and NERFC) have the highest POD and CSI values and lowest FAR values followed by the West Coast RFCs and CBRFC (Figs. 9b,d,f). Again, the lower-performing RFC regions are in the central and southern parts of the United States.

The bias and MAE_{cond} values associated with the skill scores previously discussed are shown in Figs. 9g–j. The relative performance patterns of the RFC MAE_{cond} values are very similar for both the top 1.0% and top 0.1% events with the smallest errors occurring along the

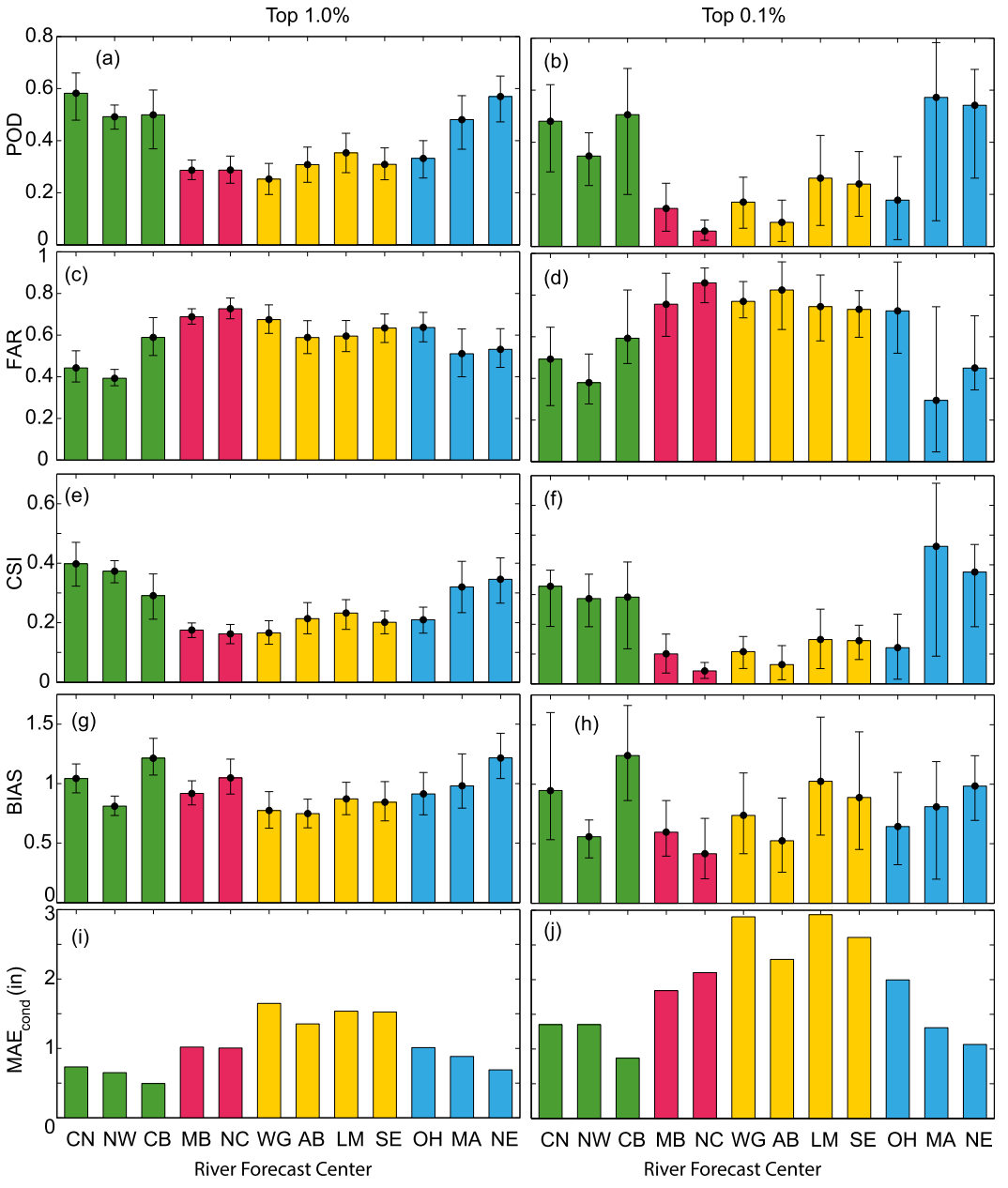

FIG. 9. Regional (a) POD, (c) FAR, (e) CSI, (g) bias, and (i) MAE_{cond} values by RFC aggregated over 2007–11 for the top 1.0% of precipitation events using the regional precipitation thresholds in Table 2. Regional (b) POD, (d) FAR, (f) CSI, (h) bias, and (j) MAE_{cond} values by RFC aggregated over 2007–11 for the top 0.1% of precipitation events using the regional precipitation thresholds in Table 2. Bar graphs are color coded into four broad U.S. geographical regions: West (green), upper Midwest (red), South/Southeast (yellow), and East/Northeast (blue). Brackets indicate 95% confidence intervals for the skill scores and bias.

West and East Coasts, as well as in the Northeast, and the largest errors occurring in the central and southern United States. In fact, the errors are about twice as large in the southern and central RFCs [i.e., the Lower Mississippi RFC (LMRFC), Arkansas-Red basin RFC (ABRFC), West Gulf RFC (WGRFC), and Southeast RFC (SERFC)] as in the coastal and Intermountain West RFCs (i.e., NWRFC, CNRFC, CBRFC, NERFC, and MARFC). The bias values for the 1.0% events show that the RFCs with the lowest MAE_{cond} values (i.e., NERFC and CBRFC) have bias values greater than 1, indicating overforecasting (Fig. 9g). In addition, the CNRFC and MARFC regions are unbiased (bias = 1). For the 0.1% events, the NERFC and CBRFC regions

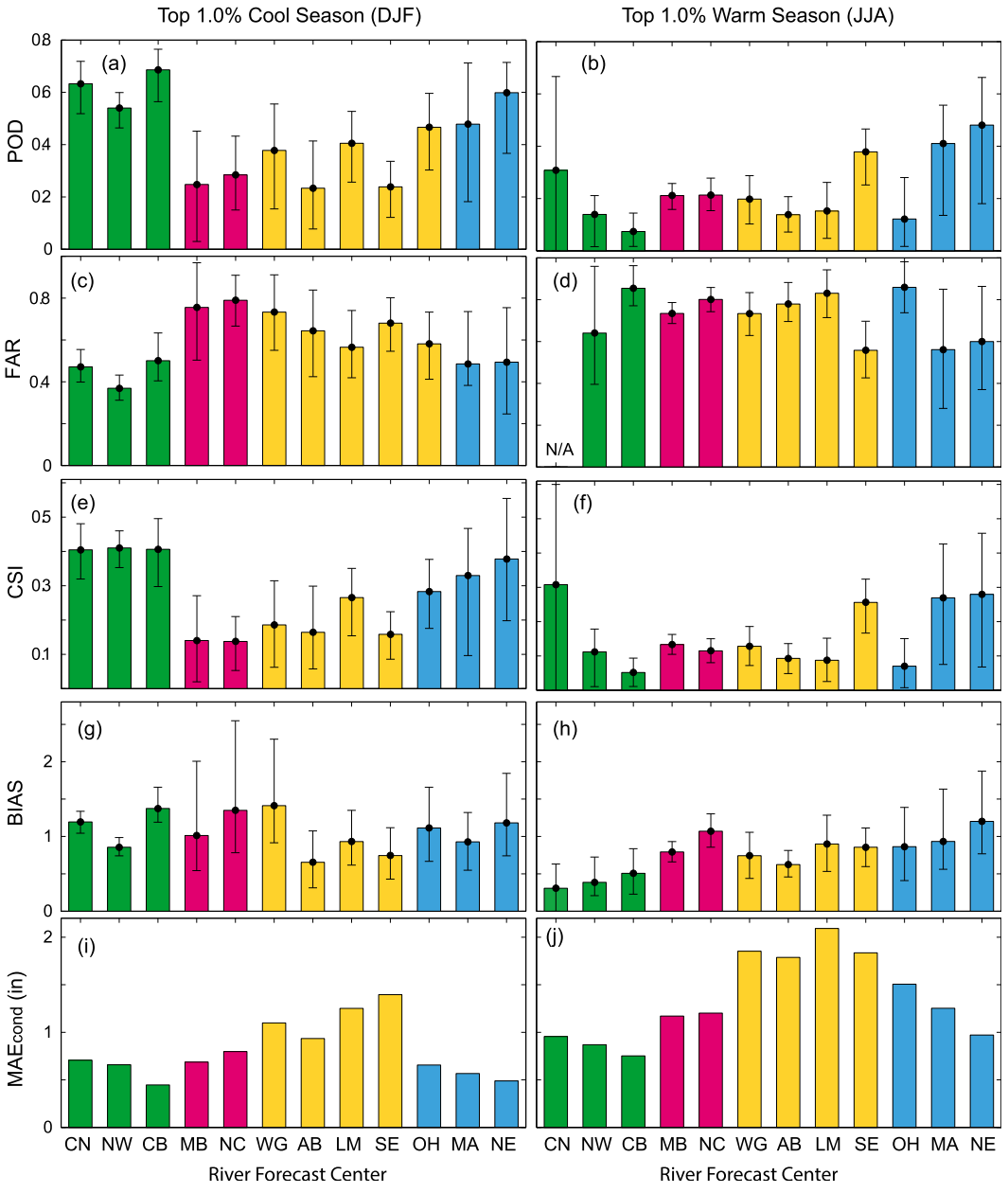

FIG. 10. Regional (a) POD, (c) FAR, (e) CSI, (g) bias, and (i) MAE_{cond} values by RFC aggregated over 2007–11 for the top 1.0% of precipitation events using the regional precipitation thresholds in Table 2 for the cool season (DJF). Regional (b) POD, (d) FAR, (f) CSI, (h) bias, and (j) MAE_{cond} values by RFC aggregated over 2007–11 for the top 1.0% of precipitation events using the regional precipitation thresholds in Table 2 for the warm season (JJA). Bar graphs are color coded into four broad U.S. geographical regions: West (green), upper Midwest (red), South/Southeast (yellow), and East/Northeast (blue). Brackets indicate 95% confidence intervals for the skill scores and bias.

are again overforecast (bias > 1); however, the rest of the RFC regions are underforecast (bias < 1) (Fig. 9h).

b. *Seasonal regional extreme QPF performance*

Relative seasonal variations of extreme QPF performance within each CONUS RFC region are illustrated in color-coded bar graphs of POD, FAR, CSI, bias, and MAE_{cond} in Fig. 10. To maximize the sample size for this seasonal analysis, only the top 1.0% events are analyzed and these events are divided into cool season (December–February; DJF) events and warm season (June–August; JJA) events rather than by monthly events.

During the cool season, WPC forecasts had the highest POD and CSI values with the lowest FARs over the West regions (i.e., NWRFC, CNRFC, and CBRFC) and East/Northeast regions (i.e., MARFC and NERFC), while the scores over the upper Midwest RFCs [i.e., Missouri basin RFC (MBRFC) and North Central RFC (NCRFC)] and some of the South/Southeast RFCs (i.e., WGRFC and SERFC) were poorer (Figs. 10a,c,e). During the warm season months, as expected, the overall accuracy (i.e., CSI values) is lower for most of the RFC regions (Fig. 10f)—with two exceptions. First, although the CNRFC CSI value is higher than the other RFC warm season CSI values, the CNRFC does not have enough 1.0% events during the warm season to be statistically significant and/or included in the warm season analysis. Second, the QPF performance in the SERFC region is actually better during the warm season versus the cool season. This again may be the result of more synoptic-scale precipitation (i.e., tropical cyclones) during this season, especially August, versus smaller-scale, more convection extreme precipitation.

Figures 10i and 10j show that MAE_{cond} values tend to be lower during the cool season and higher during the warm season. Relative to the other RFCs, the West and East/Northeast RFCs have smaller errors than do the South/Southeast RFCs for both the warm and the cool seasons. It is also interesting to note that for the cool season, more overforecasting appears to occur than during the warm season (Figs. 10g,h).

7. Summary and future work

This study benchmarks the performance of the WPC 32-km gridded QPFs for extreme precipitation events over an 11-yr period (2001–11). Per the results of Ralph et al. (2010), extreme precipitation thresholds were quantitatively defined to be the 99th percentile and the 99.9th percentile of all wet-site days [i.e., observed precipitation >0.0 in. (mm) $(24\,h)^{-1}$] within an RFC region. These extreme precipitation events are referred to as the top 1.0% and the top 0.1% events, respectively.

Extreme WPC 32-km gridded QPF performance was evaluated based on five verification measures: POD, FAR, CSI, bias, and MAE_{cond}. These scores were chosen because they are easily understood by operational forecasters. DTC MET software was used to calculate the verification output of hits H, misses M, and false alarms F for each RFC region during extreme precipitation events and to identify observed–predicted gridpoint pairs exceeding a specified threshold.

National extreme QPF scores were constructed by aggregating all of the RFC wet-site days exceeding the respective regional thresholds, and results were compared to the current NOAA Government Performance and Results Act (GPRA) precipitation threshold [≥ 1.0 in. $(24\,h)^{-1}$]. Not surprisingly, results show that extreme events have lower skill in all five verification metrics than the GPRA precipitation threshold (Fig. 5) and that skill tends to be lower with longer lead time (Fig. 6). However, the verification measures indicate that the 32-km gridded extreme QPFs have incrementally improved in forecast accuracy over the last 11 yr, and that the yearly rates of improvement for the extreme events are higher than the improvement rates of the GPRA threshold events.

National extreme QPFs by season were found to have higher skill and less error during the cool season months (December–February) and lower skill and more error during the warm season months (June–August) for all three lead times. These seasonal differences can be attributed to the scale and type of precipitation events that occur during each season. The cool season is dominated by synoptic-scale events, while the warm season tends to be dominated by convection and small-scale processes. For all the verification metrics there is a slight increase in performance for the month of September, which is most likely related to landfalling tropical cyclones, as these events are well-defined rain producers and track errors have improved over the last 11 yr (Brown et al. 2010).

To analyze regional forecast performance, verification metrics for the extreme precipitation events for each CONUS RFC domain over a 5-yr period (2007–11) were analyzed. Overall, the WPC QPFs tend to verify better in the western and eastern/northeastern United States (i.e., CNRFC, NWRFC, CBRFC, MARFC, and NERFC). WPC forecast errors tend to be higher and forecast amounts to be lower in the upper Midwest and South/Southeast United States.

Analysis of the WPC extreme QPFs by region and season indicates that for the cool season (December–February) the same western and eastern/northeastern regions (CNRFC, CBRFC, NWRFC, MARFC, and NERFC) tend to have better skill than the rest of the United States. Also, similar to the national QPF verification by season, the WPC QPFs have overall better skill during the cool season and less skill during the warm season.

The verification of extreme QPFs is highly important to improving the prediction of these extreme events. As previously stated, QPF verification is critical to identifying and evaluating forecast trends, biases, and errors and to monitoring forecast progress and improvement. Although this study focused on traditional verification metrics (i.e., POD, FAR, CSI, bias, and MAE_{cond}), future extreme QPF verification work should explore other, more recently developed, verification metrics

specifically designed for extreme events, such as the EDS, SEDS, EDI, and SEDI. Computing these newer extreme verification metrics and comparing them with the more traditional verification measures should be conducted to determine whether better verification feedback can be gained with the addition of these metrics for improving extreme QPFs.

Overall, the improvement of extreme QPFs has the potential to yield major benefits. For example, reservoir operation control rules developed decades ago specifically did not allow for the formal use of precipitation forecasts in making reservoir operations decisions (i.e., water releases), presumably because precipitation forecasts were too inaccurate. The water had to already be in the streams or in the reservoirs. In the decades since then, however, improvements have been made in QPF and reservoir operations rules are on the verge of being revised through a rigorous engineering and political process. During these revisions, it is expected that criteria could be established for QPF accuracies that would enable forecast-based reservoir operations decisions (Demargne et al. 2014). These criteria could be expressed in terms of POD, FAR, CSI, bias, and MAE_{cond}. Many of these reservoirs are in the western coastal states, where the results in this paper indicate extreme QPF performance is generally highest, implying that use of extreme QPFs for major reservoir operations decisions is most within reach for this region, and that focused efforts to raise that performance even higher could allow for explicit use of such forecasts in water supply and flood control decisions that could yield significant benefits. Future work will help quantify such potential benefits.

Although this study focused on 24-h precipitation totals, it is clear that other precipitation time periods (both shorter and longer) need to be analyzed. First, improved extreme QPF performance over shorter time periods (e.g., 1, 3, and 6 h) can aid in hydrological model improvement, particularly since RFC models use short-term QPF to force their hydrological models. Second, recent research has shown that longer-lasting events are key to creating floods in major watersheds. Ralph and Dettinger (2012) found that 3-day totals were the most useful in defining such events from a national perspective, Moore et al. (2012) documented a devastating 2-day extreme event in Tennessee, and Ralph et al. (2013a) documented that the most extreme events in Northern California were associated with landfalling atmospheric rivers that stalled for over 40 h (some lasted over 48 h). Thus, future work will include evaluations of extreme QPFs for precipitation intervals ranging from 6 to 72 h at various lead times.

The extreme QPF performance presented in this study will help identify gaps in current forecast skill on a regional basis. A key step in understanding the causes of these errors is to understand the meteorological conditions that created the extreme precipitation. This is being pursued by HMT and others through case studies, such as the recent examples from Tennessee and California mentioned above, as well as in Washington–Oregon (Neiman et al. 2008), Arizona (Neiman et al. 2013), and the eastern and southeastern United States (Moore et al. 2012). Such case studies can then help guide evaluations of current forecast tools, such as mesoscale models and data assimilation methods (e.g., Ma et al. 2011).

Acknowledgments. This research was supported by NOAA's Hydrometeorology Testbed (HMT) program and the U.S. Weather Research Program (USWRP). The authors thank Leticia Solliard (NPVU) for her assistance in obtaining the NPVU QPF and QPE data; Tara Jensen, John Halley Gotway, and Tressa Fowler from the DTC for their assistance with the confidence intervals and verification software package; and both Eric Parrish and John Adams for their graphical assistance. The authors appreciate the suggestions and input provided by Keith Brill, Mike Bodner, and Tom Workoff of NCEP's WPC and by Tim Schneider and Lynn Johnson at NOAA's ESRL/PSD. Finally, the authors wish to thank the three anonymous reviewers for their comments and suggestions.

REFERENCES

Anthes, R. A., 1983: Regional models of the atmosphere in middle latitudes. *Mon. Wea. Rev.,* **111,** 1306–1335, doi:10.1175/1520-0493(1983)111<1306:RMOTAI>2.0.CO;2.

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.,* **239,** 306–337, doi:10.1016/S0022-1694(00)00361-9.

Beven, J. L., and Coauthors, 2008: Atlantic hurricane season of 2005. *Mon. Wea. Rev.,* **136,** 1109–1173, doi:10.1175/2007MWR2074.1.

Breidenbach, J. P., D.-J. Seo, P. Tilles, and K. Roy, 1999: Accounting for radar beam blockage patterns in radar-derived precipitation mosaics for River Forecast Centers. Preprints, *15th Int. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Dallas, TX, Amer. Meteor. Soc., 5.22 [Available online at https://ams.confex.com/ams/99annual/abstracts/1699.htm.]

Brennan, M. J., J. L. Clark, and M. Klein, 2008: Verification of quantitative precipitation forecast guidance from NWP models and the Hydrometeorological Prediction Center for 2005–2007 tropical cyclones with continental U.S. rainfall impacts. *28th Conf. on Hurricanes and Tropical Meteorology,* Orlando, FL, Amer. Meteor. Soc., P2H.9. [Available online at https://ams.confex.com/ams/pdfpapers/138022.pdf.]

Brown, D. P., J. L. Beven, J. L. Franklin, and E. S. Blake, 2010: Atlantic hurricane season of 2008. *Mon. Wea. Rev.,* **138,** 1975–2001, doi:10.1175/2009MWR3174.1.

Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.,* **15,** 3–18, doi:10.1002/met.52.

Charba, J. P., D. W. Reynolds, B. E. McDonald, and G. M. Carter, 2003: Comparative verification of recent quantitative precipitation forecasts in the National Weather Service: A simple approach for scoring forecast accuracy. *Wea. Forecasting,* **18,** 161–183, doi:10.1175/1520-0434(2003)018<0161:CVORQP>2.0.CO;2.

Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Wea. Forecasting,* **17,** 238–249, doi:10.1175/1520-0434(2002)017<0238:VOPFOT>2.0.CO;2.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical–topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.,* **33,** 140–158, doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.,* **95,** 79–98, doi:10.1175/BAMS-D-12-00081.1.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585, doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.,* **84,** 481–492, doi:10.1175/BAMS-84-4-481.

——, and Coauthors, 2013: Progress and challenges in forecast verification. *Meteor. Appl.,* **20,** 130–139, doi:10.1002/met.1392.

Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting,* **26,** 699–713, doi:10.1175/WAF-D-10-05030.1.

Fread, D., and Coauthors, 1995: Modernization in the National Weather Service River and Flood Program. *Wea. Forecasting,* **10,** 477–484, doi:10.1175/1520-0434(1995)010<0477:MITNWS>2.0.CO;2.

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,** 955–965, doi:10.1175/BAMS-85-7-955.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting,* **13,** 377–395, doi:10.1175/1520-0434(1998)013<0377:TWRA>2.0.CO;2.

Ghelli, A., and C. Primo, 2009: On the use of the extreme dependency score to investigate the performance of an NWP model for rare events. *Meteor. Appl.,* **16,** 537–544, doi:10.1002/met.153.

Henkel, A., and C. Peterson, 1996: Can deterministic quantitative precipitation forecasts in mountainous regions be specified in a rapid, climatologically-consistent manner with Mountain Mapper functioning as the tool for mechanical specification, quality control, and verification? *Extended Abstracts, Fifth National Heavy Precipitation Workshop,* State College, PA, NWS/NOAA, 31 pp. [Available from Office of Climate, Water, and Weather Services, W/OS, 1325 East–West Hwy., Silver Spring, MD 20910.]

Hogan, R., E. J. O'Connor, and A. J. Illingworth, 2009: Verification of cloud-fraction forecasts. *Quart. J. Roy. Meteor. Soc.,* **135,** 1494–1511, doi:10.1002/qj.481.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 254 pp.

Junker, N. W., J. E. Hoke, and R. H. Grumm, 1989: Performance of NMC's regional models. *Wea. Forecasting,* **4,** 368–390, doi:10.1175/1520-0434(1989)004<0368:PONRM>2.0.CO;2.

Lazo, J. K., R. E. Morss, and J. L. Demuth, 2009: 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bull. Amer. Meteor. Soc.,* **90,** 785–798, doi:10.1175/2008BAMS2604.1.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology,* San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/pdfpapers/83847.pdf.]

Lowrey, M. R. K., and Z. Yang, 2008: Assessing the capability of a regional-scale weather model to simulate extreme precipitation patterns and flooding in central Texas. *Wea. Forecasting,* **23,** 1102–1126, doi:10.1175/2008WAF2006082.1.

Ma, Z., Y.-H. Kuo, F. M. Ralph, P. J. Neiman, G. A. Wick, E. Sukovich, and B. Wang, 2011: Assimilation of GPS radio occultation data for an intense atmospheric river with the NCEP regional GSI system. *Mon. Wea. Rev.,* **139,** 2170–2183, doi:10.1175/2011MWR3342.1.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

McDonald, B. E., and T. M. Graziano, 2001: The National Precipitation Verification Unit (NPVU): Operational implementation. Preprints, *Precipitation Extremes: Prediction, Impacts, and Responses,* Albuquerque, NM, Amer. Meteor. Soc., P1.32. [Available online at https://ams.confex.com/ams/annual2001/webprogram/Paper17535.html.]

Moore, B. J., P. J. Neiman, F. M. Ralph, and F. E. Barthold, 2012: Physical processes associated with heavy flooding rainfall in Nashville, Tennessee, and vicinity during 1–2 May 2010: The role of an atmospheric river and mesoscale convective systems. *Mon. Wea. Rev.,* **140,** 358–378, doi:10.1175/MWR-D-11-00126.1.

Morss, R. E., and F. M. Ralph, 2007: Use of information by National Weather Service forecasters and emergency managers during CALJET and PACJET-2001. *Wea. Forecasting,* **22,** 539–555, doi:10.1175/WAF1001.1.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338, doi:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

National Weather Service, 1999: The modernized end-to-end forecast process for quantitative precipitation information: Hydrometeorological requirements, scientific issues, and service concepts. National Weather Service, 187 pp. [Available from Office of Climate, Water, and Weather Services, W/OS, 1325 East–West Hwy., Silver Spring, MD 20910.]

Neiman, P. J., F. M. Ralph, G. A. Wick, J. D. Lundquist, and M. D. Dettinger, 2008: Meteorological characteristics and overland precipitation impacts of atmospheric rivers affecting the west coast of North America based on eight years of SSM/I satellite observations. *J. Hydrometeor.,* **9,** 22–47, doi:10.1175/2007JHM855.1.

——, ——, B. J. Moore, M. Hughes, K. M. Mahoney, J. M. Cordeira, and M. D. Dettinger, 2013: The landfall and inland penetration of a flood-producing atmospheric river in Arizona. Part I: Observed synoptic-scale, orographic, and hydrometeorological characteristics. *J. Hydrometeor.,* **14,** 460–484, doi:10.1175/JHM-D-12-0101.1.

Nielsen-Gammon, J. W., F. Zhang, A. Odins, and B. Myoung, 2005: Extreme rainfall events in Texas: Patterns and predictability. *Phys. Geogr.,* **26,** 340–364, doi:10.2747/0272-3646.26.5.340.

Novak, D. R., F. E. Barthold, M. J. Bodner, K. F. Brill, and M. Eckert, 2011: Quantifying extreme rainfall threats at the

Hydrometeorological Prediction Center. *24th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., 14A.4. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper179714.html].

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting,* **10,** 498–511, doi:10.1175/1520-0434(1995)010<0498:EOYOQP>2.0.CO;2.

Pasch, R. J., M. B. Lawrence, L. A. Avila, J. L. Beven, J. L. Franklin, and S. R. Stewart, 2004: Atlantic hurricane season of 2002. *Mon. Wea. Rev.,* **132,** 1829–1859, doi:10.1175/1520-0493(2004)132<1829:AHSO>2.0.CO;2.

Ralph, F. M., and M. D. Dettinger, 2012: Historical and national perspectives on extreme West Coast precipitation associated with atmospheric rivers during December 2010. *Bull. Amer. Meteor. Soc.,* **93,** 783–790, doi:10.1175/BAMS-D-11-00188.1.

——, P. J. Neiman, D. E. Kingsmill, P. O. G. Persson, A. B. White, E. T. Strem, E. D. Andrews, and R. C. Antweiler, 2003: The impact of a prominent rain shadow on flooding in California's Santa Cruz Mountains: A CALJET case study and sensitivity to the ENSO cycle. *J. Hydrometeor.,* **4,** 1243–1264, doi:10.1175/1525-7541(2003)004<1243:TIOAPR>2.0.CO;2.

——, and Coauthors, 2005: Improving short-term (0–48 h) cool-season quantitative precipitation forecasting: Recommendations from a USWRP workshop. *Bull. Amer. Meteor. Soc.,* **86,** 1619–1632, doi:10.1175/BAMS-86-11-1619.

——, E. Sukovich, D. Reynolds, M. Dettinger, S. Weagle, W. Clark, and P. J. Neiman, 2010: Assessment of extreme quantitative precipitation forecasts and development of regional extreme event thresholds using data from HMT-2006 and COOP observers. *J. Hydrometeor.,* **11,** 1286–1304, doi:10.1175/2010JHM1232.1.

——, T. Coleman, P. J. Neiman, R. J. Zamora, and M. D. Dettinger, 2013a: Observed impacts of duration and seasonality of atmospheric-river landfalls on soil moisture and runoff in coastal northern California. *J. Hydrometeor.,* **14,** 443–459, doi:10.1175/JHM-D-12-076.1.

——, and Coauthors, 2013b: The emergence of weather-focused testbeds linking research and forecasting operations. *Bull. Amer. Meteor. Soc.,* **94,** 1187–1210, doi:10.1175/BAMS-D-12-00080.1.

Reynolds, D., 2003: Value-added quantitative precipitation forecasts: How valuable is the forecaster? *Bull. Amer. Meteor. Soc.,* **84,** 876–878, doi:10.1175/BAMS-84-7-876.

Schumacher, R. S., and C. A. Davis, 2010: Ensemble-based forecast uncertainty analysis of diverse heavy rainfall events. *Wea. Forecasting,* **25,** 1103–1122, doi:10.1175/2010WAF2222378.1.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Rep. 8, WMO/TD-358, 114 pp.

Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.,* **15,** 41–50, doi:10.1002/met.53.

Waple, A. M., and J. H. Lawrimore, 2003: State of the climate in 2002. *Bull. Amer. Meteor. Soc.,* **84** (Suppl.), doi:10.1175/BAMS-84-6-Waple.

White, A. B., and Coauthors, 2012: NOAA's rapid response to the Howard A. Hanson Dam flood risk management crisis. *Bull. Amer. Meteor. Soc.,* **93,** 189–207, doi:10.1175/BAMS-D-11-00103.1.

Zhang, F., A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting,* **21,** 149–166, doi:10.1175/WAF909.1.