Verifying and Redefining the Weather Prediction Center's Excessive Rainfall Outlook Forecast Product

MICHAEL J. ERICKSON,[a,b] BENJAMIN ALBRIGHT,[b,c] AND JAMES A. NELSON[b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland*
[c] *Systems Research Group, Inc., College Park, Maryland*

(Manuscript received 3 February 2020, in final form 17 November 2020)

ABSTRACT: The Weather Prediction Center's Excessive Rainfall Outlook (ERO) forecasts the probability of rainfall exceeding flash flood guidance within 40 km of a point. This study presents a comprehensive ERO verification between 2015 and 2019 using a combination of flooding observations and proxies. ERO spatial issuance frequency plots are developed to provide situational awareness for forecasters. Reliability of the ERO is assessed by computing fractional coverage of the verification within each probabilistic category. Probabilistic forecast skill is evaluated using the Brier skill score (BSS) and area under the relative operating characteristic (AUC). A "probabilistic observation" called practically perfect (PP) is developed and compared to the ERO as an additional measure of skill. The areal issuance frequency of the ERO varies spatially with the most abundant issuances spanning from the Gulf Coast to the Midwest and the Appalachians. ERO issuances occur most often in the summer and are associated with the Southwestern monsoon, mesoscale convective systems, and tropical cyclones. The ERO exhibits good reliability on average, although more recent trends suggest some ERO-defined probabilistic categories should be issued more frequently. AUC and BSS are useful bulk skill metrics, while verification against PP is useful in bulk and for shorter-term ERO evaluation. ERO forecasts are generally more skillful at shorter lead times in terms of AUC and BSS. There is no trend in ERO area size over 5 years, although ERO forecasts may be getting slightly more skillful in terms of critical success index when verified against the PP.

KEYWORDS: Flood events; Rainfall; Hydrometeorology; Bias; Forecast verification/skill; Flood events

1. Introduction

a. Background

Accurate quantitative precipitation forecasts (QPF) are of critical importance to improving flash flood predictability (Cosgrove and Klymmer 2016; Gourley et al. 2017). Between 2015 and 2017, flash flooding has resulted in more fatalities than lightning, hail, tornadoes, or straight-line wind damage combined (National Weather Service 2017). Direct flood related damages between 2015 and 2017 exceeded $74 billion (U.S. dollars), with $61 billion in damage (mostly from Hurricane Harvey) for water year 2017 alone (Water Resources Services 2017).

Despite recent advances, QPF remains challenging (Cuo et al. 2011; Sharma et al. 2017), particularly during the warm season when convection associated with localized features (e.g., outflow boundaries) in weak forcing regimes are inherently less predictable than other seasons (Fritsch and Carbone 2004; Sukovich et al. 2014; Sharma et al. 2017). In addition, flooding rains during landfalling tropical cyclones carries its own challenges associated with track position, intensity, interaction with topography, and extratropical transition (Marchok et al. 2007; Brennan et al. 2008; Luitel et al. 2018).

On average, precipitation forecasts have been gradually improving over the past few decades. Equitable threat score of 24-h QPF at the 1-in. threshold between 1993 and 2018 has almost doubled for the National Oceanic and Atmospheric Administration Global Forecast System and North American Mesoscale models [Weather Prediction Center (WPC) 2017]. In addition, the feasibility of running operational convection-allowing models (CAMs) has resulted in additional improvements to warm season QPF in the short term (e.g., lead times less than 48 h; Cookson-Hills et al. 2017; Iyer et al. 2016; Ma et al. 2018). However, CAMs exhibit sharper gradients and larger magnitudes, which can result in "double penalty errors" with QPF (Newman et al. 2019) compared to lower-resolution models using traditional verification techniques.

Verification of QPF from CAMs requires more novel neighborhood or object-based methods (Clark et al. 2016; Ma et al. 2018) that consider the structure and displacement of precipitation objects. Furthermore, QPF error and bias generally increases with heavy precipitation events (Scheuerer and Hamill 2015; Sharma et al. 2017). Hence, raw model QPF is typically less accurate for high-impact events that are responsible for flash flooding, although bias correction can alleviate this problem somewhat.

An integral and somewhat underrepresented component in creating accurate flash flood forecasts is the proper consideration of the hydrological response. The relationship between precipitation and flooding is not linear and depends on antecedent soil moisture, streamflow conditions, land use, and terrain slope, to name a few (Barthold et al. 2015; Gourley et al. 2017; Erickson et al. 2019). Starting in 1978, the Quantitative Precipitation Branch (now folded into the modern-day WPC Forecast Operations Branch) started issuing the Excessive Rainfall Potential Outlook to explicitly highlight regions that could experience flash flooding (Cooley 1978). The goal was to

Corresponding author: Michael Erickson, mjaerickson@gmail.com

DOI: 10.1175/WAF-D-20-0020.1

provide forecasters in the field with a product that went beyond QPF and focus on potential impacts. Eventually this product was renamed the Excessive Rainfall Outlook (ERO) with three different risk categories (e.g., slight, moderate, and high) introduced on 5 October 2004 (WPC 2013), and an additional marginal category added in October of 2016 (WPC 2016).

The modern-day version of the ERO is a probabilistic forecast of QPF exceeding 1-, 3-, and 6-h flash flood guidance (FFG; Barthold et al. 2015; Schmidt et al. 2007) within 40 km of a point (WPC 2019; Erickson et al. 2019). FFG is created by the River Forecast Centers and is an estimate of rainfall over a given time duration that may cause small streams to flood when the stream is at bankfull. The ERO consists of four risk categories: marginal (MRGL) ranging between 5%–10% probability of QPF exceeding FFG, slight (SLGT) between 10% and 20%, moderate (MDT) between 20% and 50%, and high (HIGH) exceeding 50%. The purpose of the ERO is to provide advance notice (e.g., 1–3 days) regarding the potential for QPF to exceed FFG over the contiguous United States (CONUS). Hence, the ERO represents WPC's best attempt to provide an indicator for flash flooding, herein defined as flooding that begins within 6 h of the causative rain event.

b. *Motivation*

WPC forecasters use a variety of sources to create the ERO probabilistic forecast; including deterministic models, ensembles, statistical tools, current environmental conditions, and forecaster experience. Statistical tools include the Colorado State University Machine-Learning Probabilities first-guess field for the ERO (Herman and Schumacher 2018a,b), the Automated Atmospheric River Detection method (Wick et al. 2013), the Ensemble Situation Awareness Table [National Centers for Environmental Prediction (NCEP) 2019] to recognize extreme events, and ensemble clustering methods (Zheng et al. 2017). Dynamical tools include a variety of operational and experimental atmospheric models and blends from the United States [e.g., the Global Forecasting System (Yang and Tallapragada 2018), the High-Resolution Ensemble Forecast system (Pyle and Manikin 2018), and the National Blend of Models (Hamill et al. 2017), to name a few] and internationally (including the government of Canada, the European Centre for Medium-Range Weather Forecasts, and the Met Office).

As mentioned in section 1a, QPF is not the only important factor to consider when creating a probabilistic flash flood forecast. To explicitly simulate the streamflow/run-off component of the hydrological cycle, several operational hydrological models have been developed including the Flooded Locations and Simulated Hydrographs project that uses the Ensemble Framework for Flash Flood Forecasting hydrological system (Gourley et al. 2017) and the National Water Model (Cosgrove and Klymmer 2016). WPC has begun to utilize these products when issuing experimental EROs in the Flash Flood and Intense Rainfall summer experiments (Barthold et al. 2015) starting in 2016 (Erickson et al. 2019).

As forecasting tools have evolved over the years, so too has the ERO product. However, an extensive verification of the ERO has not been performed for several years. A complicating factor associated with verifying the ERO is related to the difficulty of finding a suitable observation dataset. Single source observations, such as Local Storm Reports (LSRs), can suffer from spatial reporting biases caused by missed events and inaccurate reporting (Gourley et al. 2013). Flooding proxies such as exceedances of FFG do not suffer from missed events but rely on approximations and assumptions that may not always accurately reflect flooding occurrence (Clark et al. 2014). This study combines several observations and proxies, as detailed in Erickson et al. (2019) and section 2a, to better capture flooding instances that may be missed with traditional observations. In addition, this study creates a new WPC-specific practically perfect (PP) method, which uses observations and proxies to develop a best-case forecast assuming perfect knowledge of the prior events.

To ensure that the current ERO forecast probabilities are reliable, this study verifies the ERO over a representative training period between 2015 and 2019. Five years is long enough to produce meaningful results while ensuring that forecasting philosophies have not significantly changed throughout the period. This study utilizes several ways to verify ERO bias and skill with the following primary goals:

- Verify the ERO using a variety of flooding observations and proxies.
- Determine if ERO probabilities are reliable.
- Determine the spatial issuance frequency of ERO forecasts.
- Analyze ERO bulk skill, including any potential trends in ERO performance over time.
- Create an informative skill metric for forecasters to assess individual events.
- Increase public awareness, understanding, evaluation, and usage of ERO forecast products.

Section 2 describes the methods of the paper, including the datasets used in the verification and metrics used to evaluate ERO bias and skill. Section 3 details the development of a new PP-method to be used as an additional ERO skill metric. Section 4 presents the results, such as the ERO issuance frequency, reliability, skill, and bias. Section 5 discusses the pros, cons, and ways to interpret the verification results of sections 3 and 4, while section 6 concludes.

2. Methodology

a. *Data*

As mentioned in section 1a, the ERO is defined as the probability of precipitation exceeding FFG within 40 km of a point. The ERO product is issued for Day 1 (from the current day valid to 1200 UTC the next day), Day 2 (valid from 1200 UTC the next day to 1200 UTC two days into the future), and Day 3 (valid from 1200 UTC two days to 1200 UTC three days into the future). On 13 October 2017, the ERO probabilistic definition was redefined from a point-based probability to the probability of flooding within 40 km of a point based on a 1-yr verification (Erickson and Nelson 2018). This change was made for two reasons; to be more consistent between National Weather Service (NWS) national center outlook products (e.g., Storm Prediction Center's Convective Outlook) and

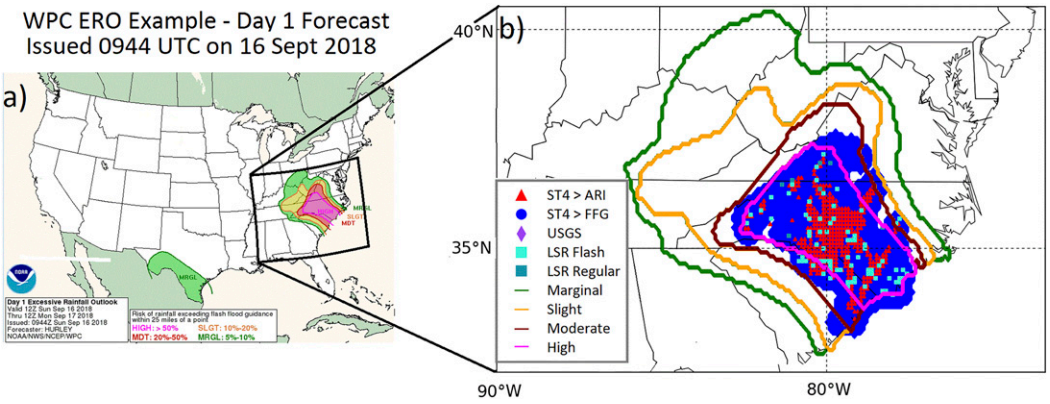

FIG. 1. Example of a Day 1 WPC ERO forecast issued at 0944 UTC 16 Sep 2018 (a) showing the marginal (green), slight (yellow), moderate (red), and high (magenta) and (b) the corresponding verification over the Southeast United States. Observations and proxies in (b) are Stage IV exceeding ARI (red triangles), Stage IV exceeding FFG (blue circles), USGS instances (purple diamonds), LSR flash floods (cyan squares), and LSR regular floods (blue squares).

because ''probability at a point'' values are sensitive to the grid resolution. Despite this redefinition, WPC forecasters did not change their forecasting philosophies. Verification does not exhibit any significant differences before or after the ERO neighborhood-based definition was implemented (not shown), provided all forecasts have the same 40 km definition applied before and after 13 October 2017. This paper represents an updated and more elaborate version of the prior 1-yr verification (Erickson and Nelson 2018) by extending the verification period to 5 years (e.g., from 2015 to 2019) and including additional ERO skill metrics.

This paper verifies the ERO using the current definition by comparing the NWS NCEP Stage IV quantitative precipitation estimate analysis exceeding FFG. The Stage IV analysis is a near real-time product generated at River Forecast Centers by utilizing radar precipitation estimates and rain gauges with some bias correction and manual quality control of data (Nelson et al. 2016).

Since FFG exceedances are a proxy for flash flooding, there is a strong and vested interest at WPC to expand the current definition of the ERO by including a more comprehensive flooding database. WPC forecasters utilize a variety of flash flooding tools (see section 1b) in addition to FFG, and the consideration of additional flooding observations and proxies would make the verification effort more consistent with how forecasters issue the ERO. As mentioned in section 1b, this extension is not intuitive since there is no single observation or flooding proxy that properly captures all instances of flooding across CONUS due to errors in reporting (e.g., missed observations) and assumptions associated with proxies (Gourley et al. 2013; Clark et al. 2014; Erickson et al. 2019). Flash flooding is not treated consistently within NWS starting with the definition and carrying through to the forecasting, reporting, and verification of these events (Barthold et al. 2015). As a result, in addition to comparing to Stage IV analysis exceeding FFG, the ERO is also compared to a suite of flooding observations and proxies including Stage IV analysis exceeding 5-yr average recurrence interval (ARI; Perica et al. 2013), U.S. Geological Survey (USGS) river gauge observations, and NWS local storm reports (LSR) observations. The 5-yr ARI is used in this study since it qualitatively matches the flooding observations well, with 3-h rainfall exceeding 5-yr ARI capturing 80% of all floods (Lincoln and Thomason 2018). This combination of all flooding observations and observation proxies is called the Unified Flooding Verification System (UFVS) within WPC and is also discussed in Erickson et al. (2019).

An example of the Day 1 ERO product for Hurricane Florence valid 1200 UTC 16 September–1200 UTC 17 September 2018 with corresponding UFVS verification is shown in Fig. 1. In this case, all observations and proxies highlight similar regions, with Stage IV exceeding FFG instances covering a larger region than Stage IV exceeding ARI instances. In general, FFG exceedances are more common over the eastern half of CONUS, with 5-yr ARI exceedances being more common over portions of the High Plains and Intermountain West (not shown). Flooding proxy instances occur much more frequently than LSRs and USGS observations. Spatial consistency among the verification datasets in the UFVS is typical with severe flooding events, but borderline flooding events typically exhibit weaker consistency with one or two observations/proxies highlighting a similar region.

b. *Verification*

The Model Evaluation Tools (MET) Version 8.0 (Halley Gotway et al. 2018) software are used in combination with Python wrappers to verify the ERO from 1 January 2015 to 31 December 2019. MET is a set of verification software developed by the Developmental Testbed Center (DTC) where the numerical weather prediction community can evaluate numerical weather prediction output in a variety of ways. MET features include regridding capability, evaluation to point-based observations, spectral decomposition, evaluation to grid-based analysis, options to perform tropical cyclone verification, and object-based identification and tracking, to name a few (Halley Gotway et al. 2018).

Since the MRGL category was added on 1 August 2016, most verification excludes this category for consistency and to

evaluate 5 years of data spanning back to 1 January 2015. The rarity of ERO issuances is evaluated spatially using heat maps (i.e., spatial frequencies of each ERO category) throughout CONUS. ERO reliability is assessed by computing the average fractional coverage of the verifying FFG exceedances and the entire UFVS within 40 km of a point for each risk category. Fractional coverage for an individual event is defined as the area of the verifying observation/proxy inflated to a 40-km radius divided by the total area of the WPC ERO contour drawn. Verification is performed on a 10-km spaced grid. An example of nearly 100% fractional coverage of Stage IV exceeding FFG instances at the HIGH category is shown in Fig. 1b (e.g., blue circles encompassing the magenta HIGH contour).

Considerable attention has been devoted to evaluating the probabilistic skill of the ERO. Evaluating ERO skill can be beneficial to forecasters when assessed in bulk over a verification period or when evaluating a single event. For this study, probabilistic skill is analyzed using the Brier skill score (BSS) and area under relative operating characteristic (AUC; Wilks 2011) using Stage IV exceeding FFG and the entire UFVS. BSS is computed by referencing the Brier score of the WPC probabilities against the Brier score computed using the entire UFVS climatology between 2015 and 2019. For consistency, only events with a SLGT area covering greater than 300 km^2 are considered in BSS and AUC computations. This 300-km^2 size limitation in the BSS and AUC computations is chosen to focus on larger flooding events.

To analyze the skill of individual events, this study uses a modified form of the PP forecast technique originally developed at the Storm Prediction Center (Hitchens et al. 2013). PP forecasts are designed to resemble the best-case forecast given perfect knowledge of the events beforehand. While Hitchens et al. (2013) focus on convective outlooks at one threshold (e.g., slight outlooks) and compare daily forecasts to practical minimum/maximum skill scores, this study seeks to develop reliable PP probabilities for multiple ERO thresholds. This PP methodology is designed to answer questions such as ''Would this storm be considered a moderate ERO event?''

The PP method is accomplished by applying a neighborhood radius of influence (ROI) around the verifying point observations (i.e., inflating the point observation), and then smoothing the binomial observation field (e.g., 100% when flood, 0% when no flood) with a Gaussian filter to produce PP probabilities that are most similar to the ERO probabilities. Since two probabilistic fields are being compared, the PP approach is suitable for assessing forecasters performance for an individual storm (e.g., did this event reach the moderate threshold?) and can be used to develop a bulk skill metric (e.g., mean absolute error). Note that the PP method must first be tuned to the forecast product with which it will eventually be compared.

WPC has traditionally implemented PP in WPC's flash flood and intense rainfall experiments using LSRs with a ROI of 40 km and a Gaussian kernel smoother standard deviation of 100 km (Perfater and Albright 2017). However, grid-based flooding proxies (e.g., Stage IV exceeding FFG or ARI) occur more frequently than LSRs (e.g., Fig. 1), which can result in undesirably high PP probabilities compared to the ERO when using the default PP configuration. This issue is even more egregious when aggregating all flooding observations and proxies in the UFVS into one PP field.

An example of the high PP bias is shown for the 24-h period valid between 1200 UTC 17 August and 1200 UTC 18 August 2018 (Fig. 2). For this example, the PP is generated separately using LSRs and USGS observations (Fig. 2a), instances of Stage IV exceeding FFG (Fig. 2b), instances of Stage IV exceeding ARI (Fig. 2c), and all combined observations and proxies applied to one grid (Fig. 2d). PP using LSRs and USGS observations only highlights the approximate region over the Northeast but fails to identify many flooding instances within the SLGT risk area over the High Plains (Fig. 2a). While the High Plains forecast may look like a forecast bust (e.g., an event is forecasted to occur but does not) using just observations, FFG and ARI instances identify several regions in this area and elsewhere over CONUS (Figs. 2b,c). These results suggest that proxies may be important in data sparse regions (in this case portions of Colorado and New Mexico) for identifying flooding instances. However, aggregating all observations and proxies' results in four separate MDT regions and one HIGH region when the actual ERO predicted two separate SLGT areas (Fig. 2d). For reference, instances of ERO HIGH are very rare, averaging 4.36% of all days between 2015 and 2018 (not shown). Hence, the PP approach utilizing all data from the UFVS must be refined to reduce the high bias and better align with the operational ERO before it can be used as a verification metric.

To determine the optimal PP configuration for the UFVS, sensitivity studies are performed from 1 January 2017 to 31 December 2017 while varying the value inside the ROI [hereafter referred to as the proxy fractional value (PFV) from 0.4 to 1] and Gaussian smoother (kernel standard deviation between 90 and 120 km) for all grid-based flooding proxies. Varying the PFV is related to the fraction of area inside the ROI of the UFVS proxies that is expected to experience flooding, while varying the Gaussian filter is associated with the spatial uncertainty of the forecast. The goal of this optimal PP is to create a product with similar average magnitudes and properties to that of the ERO. To be consistent with previous implementations of PP, LSR and USGS observations retain a PFV of 1.

Similar to Figs. 2a–c, three unique PP fields are created separately and then averaged consisting of 1) combined LSR and USGS observations, 2) instances of Stage IV exceeding FFG and 3) instances of Stage IV exceeding ARI. The averaging of three unique PP fields rather than aggregating all data from the UFVS into one field (e.g., as is done in Fig. 2d) serves as a spatial consistency check (i.e., it will be difficult for PP to create a HIGH if all three data types are not in close proximity to each other) between the different observations and proxies of the UFVS. The optimal PP from the sensitivity studies is selected as the default WPC configuration and rerun from 2015 to 2019 so it can be compared to the ERO. While the training and verification period for PP are not independent, similar results can be found by validating the ERO for 2015, 2016, 2018, and 2019 only.

When appropriate, uncertainty in the ERO verification is assessed using bootstrapping (Wilks 2011), by resampling the

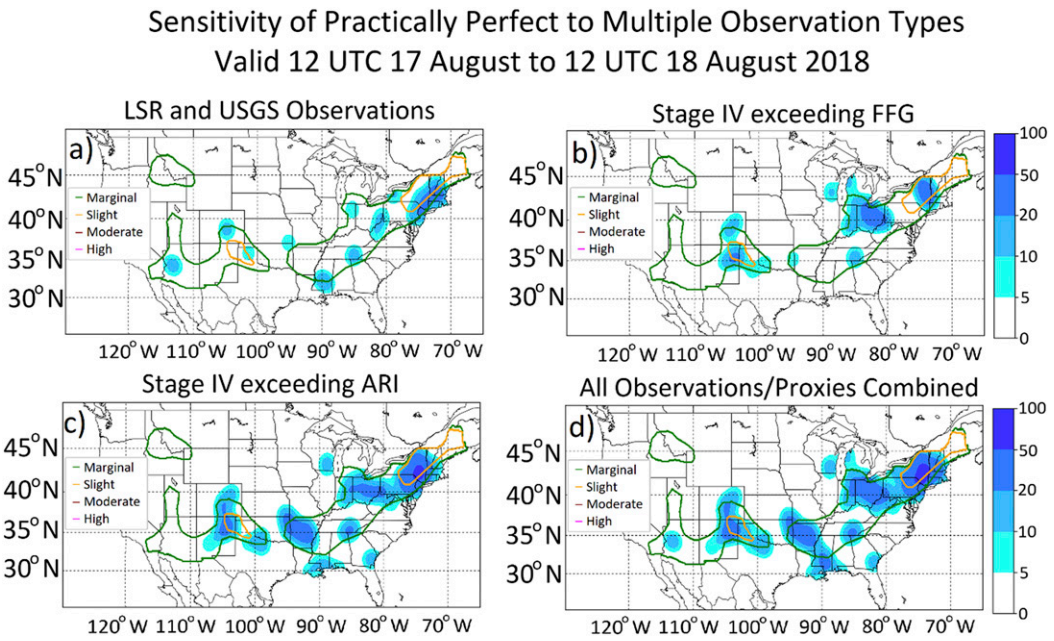

FIG. 2. Practically perfect utilizing different types of flash flooding observations and proxies compared to the Day 1 ERO issued at 0900 UTC and valid between 1200 UTC 17 Aug and 1200 UTC 18 Aug 2018. Practically perfect uses a 40-km radius of influence and a 100-km Gaussian filter with (a) just LSRs and USGS gauge observations, (b) just instances of Stage IV exceeding FFG, (c) just instances of Stage IV exceeding the 5-yr ARI, and (d) all observations and proxies to compute one practically perfect field.

original dataset with replacement 10 000 times. In cases where skill metrics (e.g., BSS and AuROC) are subset by month, year, or some other condition, the raw data are conditionally subset and resampled before the skill metric is calculated. In cases of bootstrapping, error bars represent the 2.5th and 97.5th percentile.

3. Practically perfect results

a. 1-yr practically perfect sensitivity studies

Frequency bias (FB) and critical success index (CSI; Wilks 2011) are computed by verifying the PP sensitivity studies against the ERO throughout 2017 (Fig. 3) for the same PP and ERO thresholds (e.g., 50% ERO is compared to 50% PP). Further details on how these metrics are computed are shown in the appendix. The thresholds reaching 10%, 20%, and 50% are analogous to the ERO's SLGT, MDT, and HIGH risk categories, respectively. MRGL is excluded from this verification due to the low-impact nature of this category and the complex shape of MRGL ERO contours, as explained in section 5. The sensitivity studies properly identify the zero-bias in the parameter space (i.e., the region where the ERO threshold is unbiased for the corresponding PP threshold) for all ERO categories, although they are in slightly different locations depending on the category. The zero bias in the parameter space associated with a smoothing radius of 105 km is near a PFV of 0.75, 0.45, and 0.8 for SLGT, MDT, and HIGH, respectively. There is no reason why the PP should be simultaneously unbiased within the same parameter space for all three ERO risk categories, since they are based on arbitrary probabilistic thresholds. The parameter space of highest CSI (Figs. 3d–f) is at PFV = 0.8; smoother = 120 km, PFV = 0.5; smoother = 120 km, and PFV = 0.85 smoother = 90 km for the SLGT, MDT, and HIGH categories, respectively.

The optimal parameter space of the PP parameters (e.g., PFV and Gaussian filter) was analyzed by comparing the PP probabilities to the Days 1–3 ERO probabilities. In general, the optimal parameter space was identifiable by minimizing bias and error for most metrics analyzed but varied slightly depending on the bias metric (e.g., FB or mean error), skill metric (e.g., mean absolute error, CSI, or equitable threat score) and category analyzed (not shown). Since there are less issuances of the ERO SLGT, MDT, and HIGH on Day 3 compared to Day 1, the positive bias in PP grew with increasing lead-time (not shown). However, the goal of this study is to construct a PP methodology based on the Day 1 ERO when forecast confidence is highest. The PP sensitivity study with the probabilities that best matched the Day 1 ERO during a 1-yr validation study had a PFV of 0.8 (1) for all grid-based proxies (observations), and a Gaussian smoother of 105 km for all observations and proxies. This PP methodology is selected as the default configuration for the remainder of this study.

b. 5-yr verification of the optimal practically perfect configuration

To assess how similar the new PP configuration is to the WPC ERO, a 5-yr verification (2015–19) is performed (Fig. 4). The 5-yr PP verification can be compared to the ERO to assess bulk skill. For this comparison to be consistent, the continuous PP field is converted to the discrete values of the ERO (e.g., MRGL through HIGH). The larger sample size of the 5-yr PP

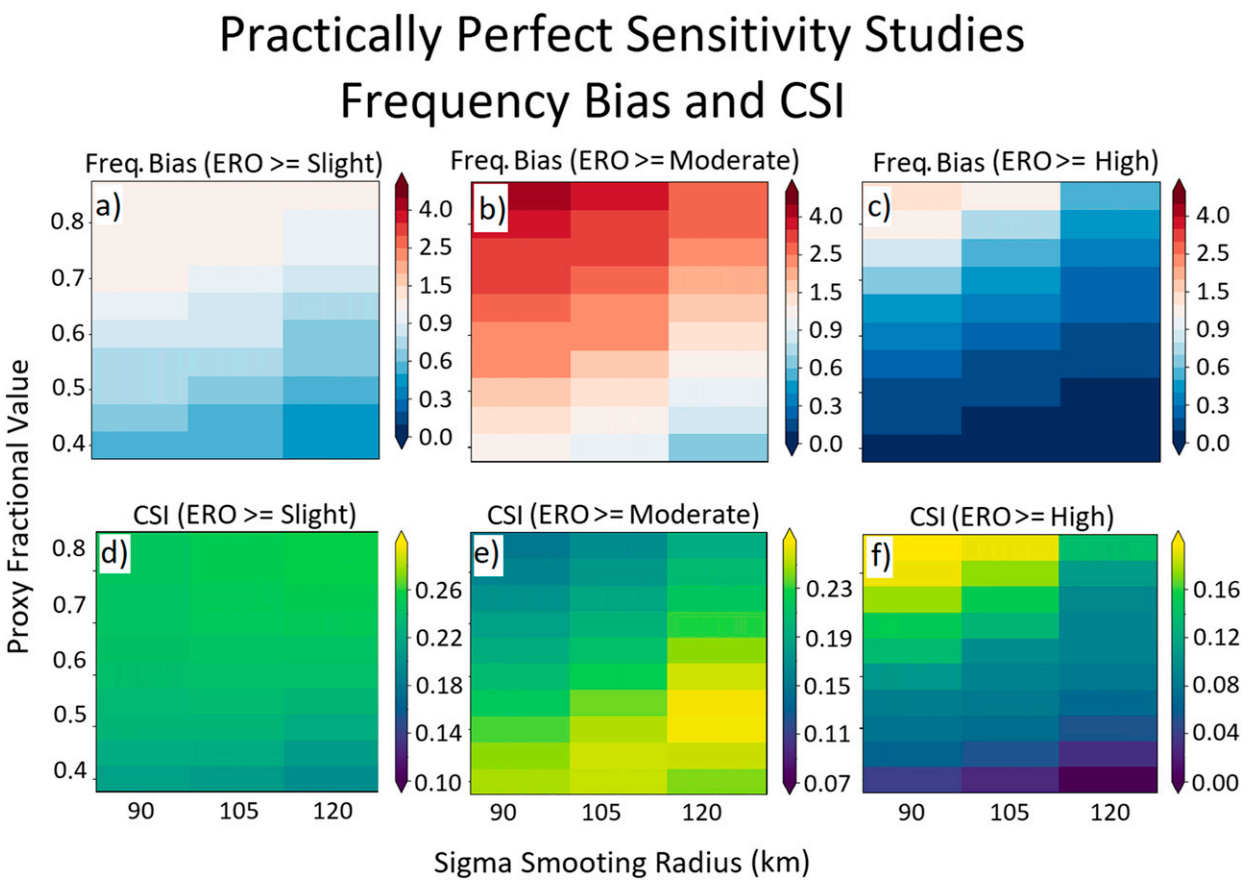

FIG. 3. Day 1 sensitivity studies spanning all of 2017 from different practically perfect configurations varying the proxy fractional value and smoothing radius for (top) frequency bias and (bottom) critical success index reaching the ERO (a),(d) slight; (b),(e) moderate; and (c),(f) high categories.

can be used to gather seasonal statistics and infer potential displacement issues when analyzed spatially. In addition, PP can be compared to the ERO for individual days or events to evaluate WPC forecast performance.

Mean error is computed by averaging the event-specific total instances of PP exceedances minus total instances of ERO exceedances, normalized by instances of PP exceedances. Mean absolute error is computed similarly using the absolute difference. In general, there is very little bias for SLGT and HIGH (0.0 and −0.1, respectively), and a positive bias for the MDT (mean error = 0.6; Fig. 4a). These 5-yr bias results are consistent with the 1-yr sensitivity studies for this configuration (Figs. 3a–c). As discussed in section 3a, the sensitivity studies have difficulty in simultaneously targeting the zero-bias in the parameter space for all categories, and this study prioritized PP SLGT and HIGH exhibiting little bias over that of the MDT threshold. However, 1-yr verification studies (not shown) have shown that forecasters should issue more frequent or larger sized MDTs, which may partially explain the PP positive bias.

The false alarm ratio, hit rate, and CSI (Wilks 2011; appendix) are shown in Fig. 4b. Not surprisingly, the false alarm ratio is relatively high for all categories (exceeding 0.59), with hit rates exceeding 0.22 for all categories. The relatively high false alarm ratio and lower hit rates and CSI are the result of inherent predictability limitations associated with flash flood forecasting and displacement biases related to grid-based verification. To reduce the issue of displacement, the probability of an ERO category being issued anywhere in CONUS given the PP predicts said category within CONUS is presented in Fig. 4c. When the PP method predicts a SLGT event, there is an 79% probability that WPC forecasters will also issue a SLGT. This is encouraging and suggests that PP instances of SLGT coincide well with WPC forecasting SLGT. The conditional probabilities are lower for the MDT and HIGH categories at 30% and 31%, respectively.

The spatial frequency difference of the PP method is shown for SLGT, MDT, and HIGH categories in Fig. 5. Spatial frequency difference is the number of instances of a PP threshold being reached minus the number of instances of an ERO threshold being reached, divided by the total number of events and multiplied by 100. The overall zero bias for SLGT noted in Fig. 4a is seen in Fig. 5a. Users and interested parties of WPC forecast products should not interpret Fig. 5a to signify that WPC forecasters should issue less SLGTs to remove the negative bias, since it is difficult to remove PP bias for all ERO categories when tuning the PP method. Instead, strong

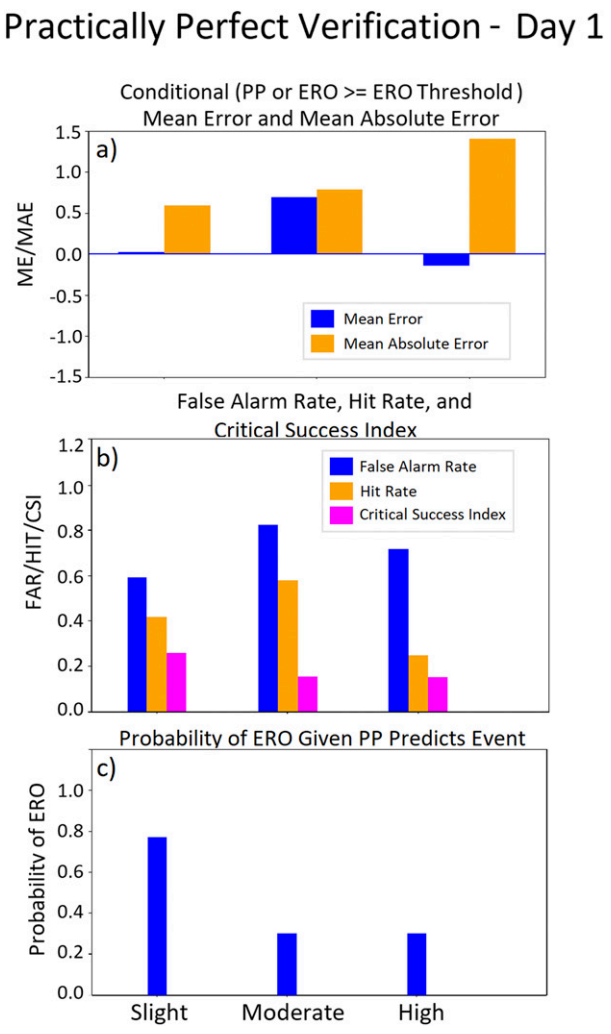

FIG. 4. Day 1 verification of the optimal practically perfect sensitivity study against EROs between 2015 and 2019 showing (a) conditional (PP or ERO achieving risk category) mean error and mean absolute error; (b) bulk false alarm ratio, hit rate, and critical success index; and (c) the probability of an ERO risk category being issued given it is predicted by practically perfect.

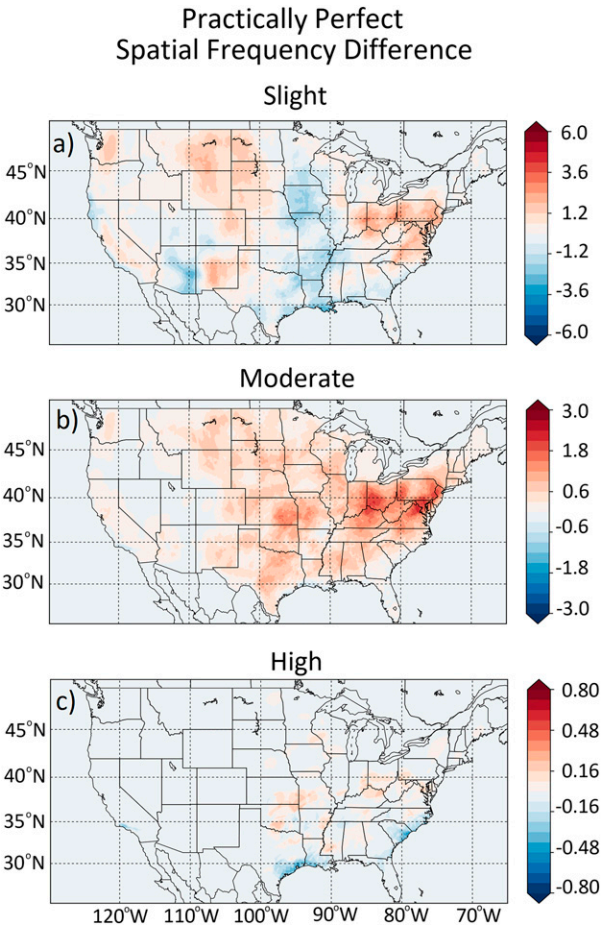

FIG. 5. Day 1 spatial frequency difference of the optimal practically perfect sensitivity study spanning 2015–19 compared to the ERO at the (a) slight, (b) moderate, and (c) high categories.

gradients in spatial frequency difference can be used to deduce where forecasters may be overforecasting or underforecasting ERO categories. For instance, there is a negative bias in the high terrain of Arizona following the Mogollon Rim, with a small positive bias in the lower terrain to the southwest (Fig. 5a). The gradient of bias, rather than the specific bias values, suggest that forecasters issue considerably more SLGTs in the higher elevation regions (i.e., where the PP bias is negative) compared to the lower elevations (i.e., where the bias is zero or slightly positive). WPC forecasters in an internal verification study have noted that issuances of SLGT are overforecast (underforecast) in the higher (lower) terrain of Arizona (Lamers 2019), and this study supports that forecasters may want to shift their ERO SLGT contours toward the lower terrain of Arizona. Similar frequency difference gradients are also noted for SLGT in the central Appalachians, High Plains, Central Plains/Midwest, and in California (Fig. 5a).

For the MDT category, there is a large gradient of frequency difference between the positively biased Washington, DC area and the near zero bias in southern North Carolina (Fig. 5b). This gradient is also apparent for HIGH, albeit with a smaller sample size (Fig. 5c). These results suggest that less flooding events verify along most of the Gulf Coast compared to the central mid-Atlantic. During an event where both the Gulf Coast and mid-Atlantic are equally threatened by a MDT or HIGH event, these results suggest that historically the mid-Atlantic region is more likely to verify. This increased mid-Atlantic verification is likely caused by a greater population density resulting in more LSRs and USGS observations.

4. ERO verification

a. ERO issuance frequency

The issuance frequency of the Days 1–3 operational ERO is presented spatially for the SLGT, MDT, and HIGH categories in Fig. 6. Focusing on the Day 1 ERO (Figs. 6a,d,g), the most

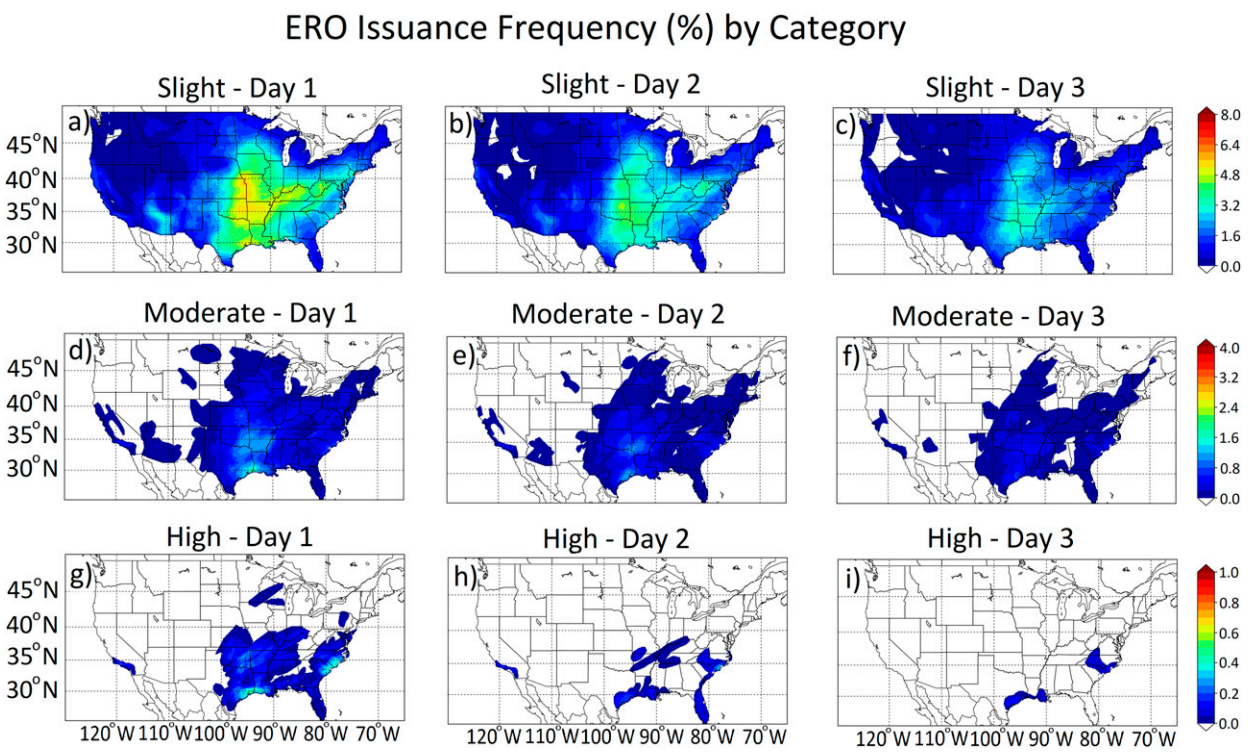

FIG. 6. Issuance frequency of ERO forecasts spanning 2015–19 by risk category for (a)–(c) slight, (d)–(f) moderate, and (g)–(i) high over (left) Day 1, (center) Day 2, and (right) Day 3.

common issuance regions span from the western and central Gulf Coast region northward to the southern Midwest and eastward to the central Appalachians. MDT and HIGH instances are considerably rarer than SLGT, with the maximum issuance frequency for SLGT, MDT, and HIGH being 5.9%, 1.7%, and 0.6%, respectively. SLGT and MDT events are categorized by a variety of atmospheric phenomena, including tropical, monsoon, mesoscale convective systems, and synoptic storms. HIGH events typically occur with tropical or tropical-transitioning cyclones, although there are rare exceptions in Southern California and the Midwest.

Issuances of SLGT, MDT, and HIGH are considerably rarer at Day 3 (Figs. 6,c,f,i) compared to Day 1, due to increased forecast confidence at shorter lead times. Some regions experience a large increase in issuance frequency at shorter lead times compared to longer lead times, such as the desert Southwest, portions of the Midwest, and Ohio/Tennessee River Valley region. Between 2015 and 2019, WPC did not operationally issue HIGH risks in their ERO on Day 3 until the latter half of 2019, with Hurricane Harvey (2017), Hurricane Florence (2018), and Hurricane Barry (2019) being three exceptions.

The Day 1 SLGT ERO issuance frequency is separated by season [e.g., December–January–February (DJF), March–April–May (MAM), June–July–August (JJA), and September–October–November (SON)] in Fig. 7. The ERO exhibits a strong seasonal dependence, with DJF SLGT issuances dominated by synoptic events (synoptic events and atmospheric rivers) in the Southeast (West Coast; Fig. 7a). Many of the West Coast SLGT issuances occurred during the anomalously active 2016/17 winter season (Guirguis et al. 2019). MAM (Fig. 7b) represents a transition period where synoptic storms are common, but with increased mesoscale convective systems in the Plains and Midwest. The JJA period is most active and dominated by mesoscale convective systems, tropical activity, the Southwest monsoon, and occasionally synoptic low pressure systems (Fig. 7c). SON (Fig. 7d) is also a period of transition and features the decaying Southwest monsoon and tropical activity with an increase in synoptically forced events.

b. *Reliability of the ERO*

As mentioned in section 2b, the Days 1–3 average fractional coverage of the FFG exceedances (blue) and the entire UFVS (orange) within 40 km of a point is computed for the SLGT, MDT, and HIGH categories between 2015 and 2019 (Fig. 8). The fractional coverage by category is assessing reliability, and Fig. 8 is equivalent to a reliability plot tuned to the ERO-specific probability thresholds. Since the operational ERO is defined as the probability of Stage IV exceeding FFG within 40 km of a point, the blue bars are used to assess the ERO's reliability. Hence, the ERO is considered reliable if the blue bar (FFG exceedances) falls between the green line (lower ERO probabilistic definition) and the red line (higher ERO probabilistic definition).

For all days and categories in Fig. 8, the ERO can be considered reliable. There is very little difference between fractional coverage values at Days 1, 2, or 3. When considering the entire UFVS database (orange bar), the fractional coverage is

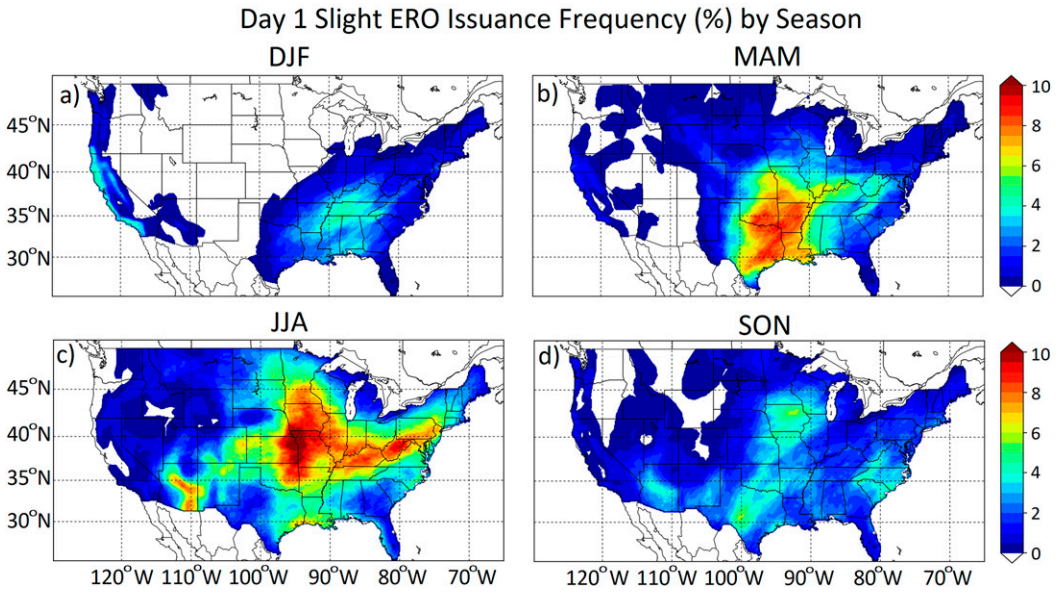

FIG. 7. Issuance frequency of Day 1 ERO forecasts spanning 2015–19 at the slight risk category for (a) DJF, (b) MAM, (c) JJA, and (d) SON.

on the high end or slightly higher than the ERO probabilistic definition. This result suggests that the ERO probabilistic categories would need to be adjusted upward if the ERO product were to be redefined to consider all the UFVS. Otherwise forecasters should draw larger areas or make more frequent issuances.

As mentioned in section 2b, the MRGL category is excluded from most of the verification in this study since it was introduced within the study's verification window. To assess the reliability of the MRGL category, fractional coverage is computed from 1 August 2016 to 31 December 2019 for all categories (Fig. 9). All FFG-based ERO categories are reliable during this period, except for Days 1–3 SLGT and Day 1 MDT, which falls just above the probabilistic range. Although the MRGL category is reliable on the average, differences between Figs. 8 and 9 emphasize that ERO probabilities may not be as well calibrated over smaller intervals. For instance, Day 1 preliminary plots from 1 January 2019 to 31 August 2019 suggest that the SLGT and MDT fractional coverage continues to fall just above the probabilistic definition (not shown). This could be caused by forecasters drawing smaller, more accurate, or less frequent SLGT and MDT contours starting in the latter half of 2017.

c. *ERO skill metrics*

BSS (cool colors) and AUC (warm colors) are shown by ERO issuance time for Days 1–3 (Fig. 10). Results with the MRGL category excluded (cyan and magenta) and included (blue and red) result in higher BSS and higher AUC when MRGL is included due to additional probability categories. For Day 1 (Fig. 10a), the BSS and AUC is slightly higher with later issuance times, suggesting improvement in the forecast closer to the event. In addition, there is a small improvement in BSS and AUC throughout the day for Day 2 (Fig. 10b) and Day 3 (Fig. 10c), although these changes are minor. Of note is the improvement in these metrics (averaging 0.03 for BSS and by 0.05 for AUC) from Day 2 to Day 1 with the MRGL included.

Figure 11 shows BSS averaged and grouped by month for 2015 (green), 2016 (blue), 2017 (cyan), 2018 (red), 2019 (magenta), and the arithmetic mean (black line). As expected, there is a strong seasonality to BSS with higher (lower) values during the cool (warm) season. This is likely the result of BSS being sensitive to the observed frequencies (e.g., lower observed frequencies have lower BSS) and modeled precipitation exhibiting lower error during the cool season, due to more predictable larger-scale forcing (Sukovich et al. 2014). However, there is considerable variability in BSS from year to year, with BSS varying by a factor of 2 for many of the months analyzed.

Figure 12 displays annual averages of BSS, AUC, and ERO size (in grid points) between 2015 and 2019. In general, BSS (AUC) values have decreased (increased) slightly between 2015 and 2019 (Figs. 12a–c), but these results do not lend themselves well to statistical significance tests. Although there was a decrease in ERO SLGT size in 2017, this appears to be an outlier compared to the other years analyzed (Figs. 12d–f). Hence, the increase in SLGT and MDT fractional coverage shown in Fig. 9 compared to Fig. 8 is likely not caused by forecasters systematically drawing smaller contours. However, it is possible that the higher fractional coverage is caused by a combination of higher forecaster confidence with an increased proclivity to not issue a SLGT and MDT when the forecaster is on the fence.

As mentioned in section 2b, BSS and AUC exhibited large variability from day-to-day, even when less significant events were filtered out. This is not apparent when verifying the ERO against the PP probabilities using monthly averaged FB and

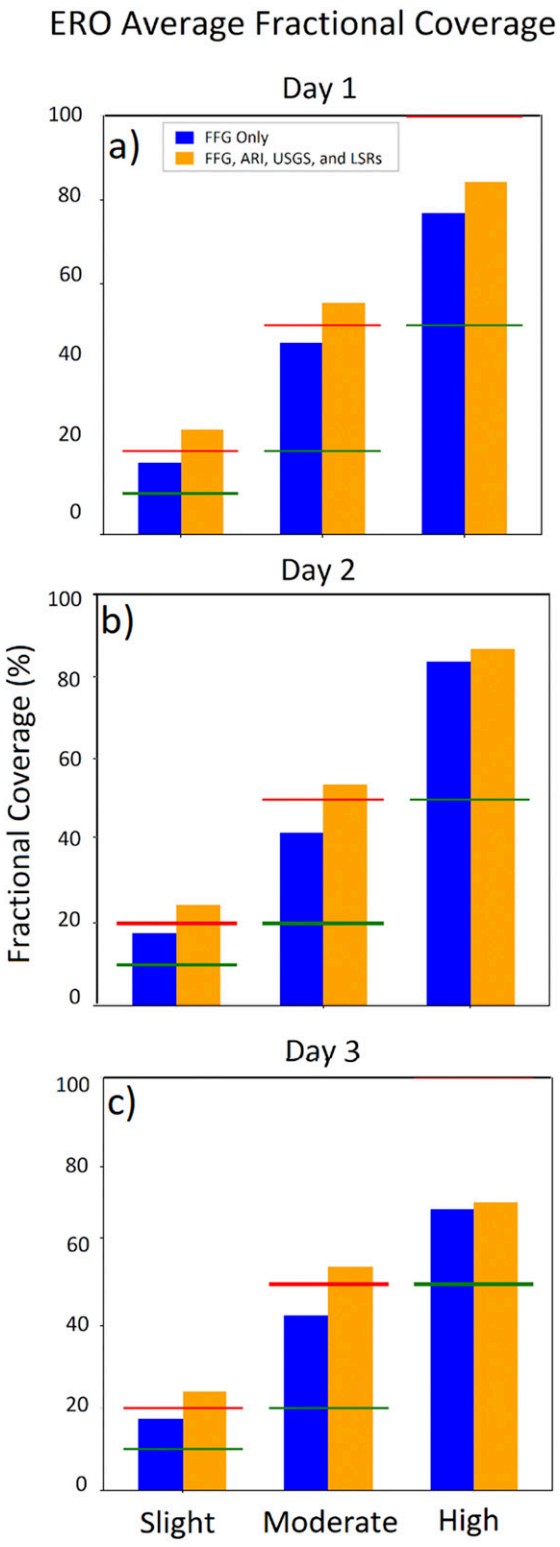

FIG. 8. ERO average fractional coverage between 1 Jan 2015 and 31 Dec 2019 by risk category verified against Stage IV exceeding FFG (blue) and Stage IV exceeding all flooding proxies/observations (orange) for (a) Day 1, (b) Day 2, and (c) Day 3.

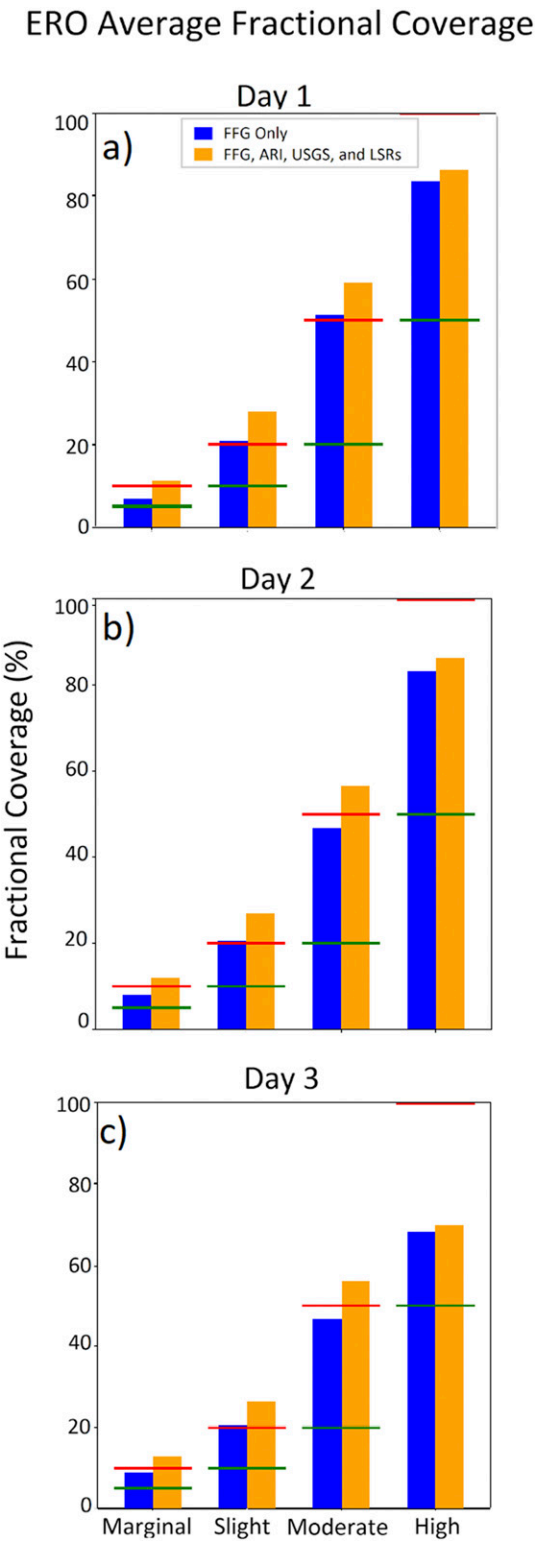

FIG. 9. As in Fig. 8, but including the newer marginal risk category introduced to the ERO on 16 Aug 2016. Note all categories are verified between 16 Aug 2016 and 31 Dec 2019.

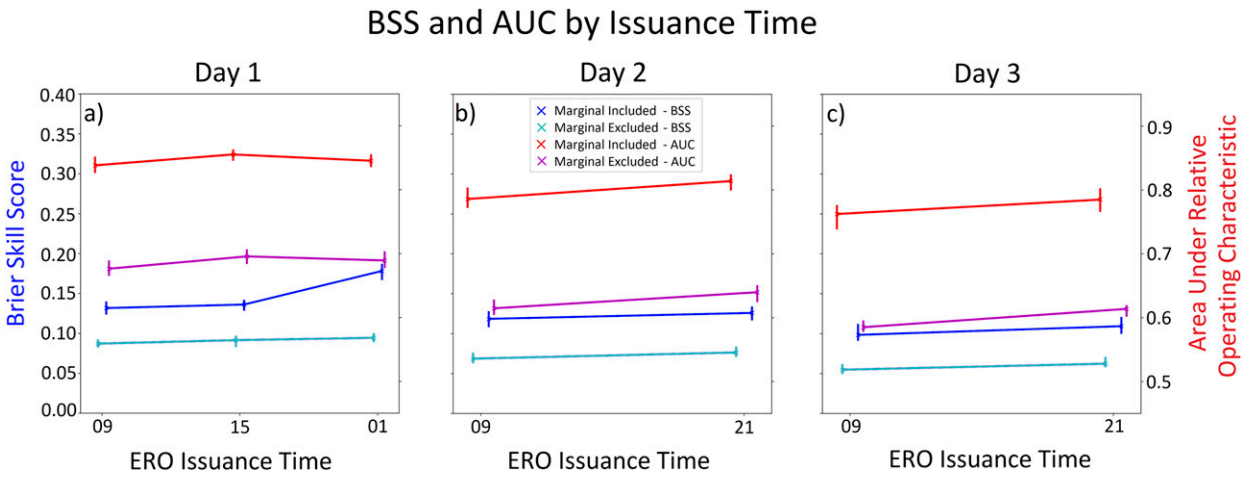

FIG. 10. Bulk BSS (cooler colors) and AUC (warmer colors) averaged by ERO issuance time for (a) Day 1, (b) Day 2, and (c) Day 3. The ERO is verified against Stage IV exceeding FFG for marginal included BSS (blue) marginal excluded BSS (cyan), marginal included AUC (red), and marginal excluded AUC (magenta). Error bars represent the 2.5th and 97.5th confidence intervals using a bootstrapping method.

CSI (Fig. 13). Most months exhibit a near zero bias with SLGT (Fig. 13a), which may be decreasing in the last 2 years of verification. Note that while this trend is statistically significant at 90% using a nonparametric Mann–Kendall (M–K; Mann 1945; Kendall 1975) test, this verification only spans 5 years and considers each data point (e.g., each month) as independent. Note that the verification in section 3b treats the ERO as the observation, while this verification treats the PP as the observation. There also may be a small positive trend in CSI for SLGT over the 5-yr verification (Fig. 13d), which is statistically significant at 90% using a M–K test. This result is consistent with previous model (Hamill et al. 2013; Baxter et al. 2014) and WPC QPF studies (Novak et al. 2014; Sukovich et al. 2014) that show a gradual improvement in skill over time. There is no seasonality to FB or CSI verified against the PP. This is encouraging when comparing the ERO to PP in day-to-day or bulk verification like Fig. 13.

5. Day-to-day utility of the practically perfect method

As mentioned earlier, evaluating the ''goodness'' of the ERO for a specific forecast or storm remains challenging. The usage of standard error verification metrics such as BSS, AUC, and equitable threat score exhibit significant variability from day-to-day due to changes in spatial coverage and are more useful when assessed in bulk. An ERO-specific PP is developed to remedy this issue, and the utility of this methodology is still being evaluated internally at WPC. However, this section will describe some of the pros and cons of the PP methodology thus far.

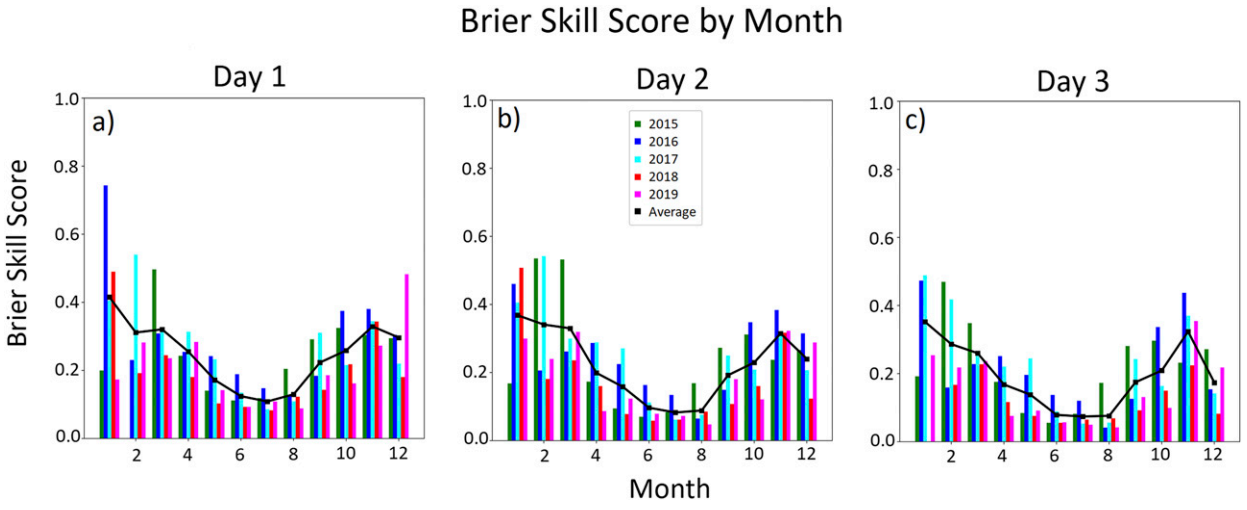

FIG. 11. Averaged monthly BSS grouped by year for 2015 (green), 2016 (blue), 2017 (cyan), 2018 (red), 2019 (magenta), and the 5-yr average (black).

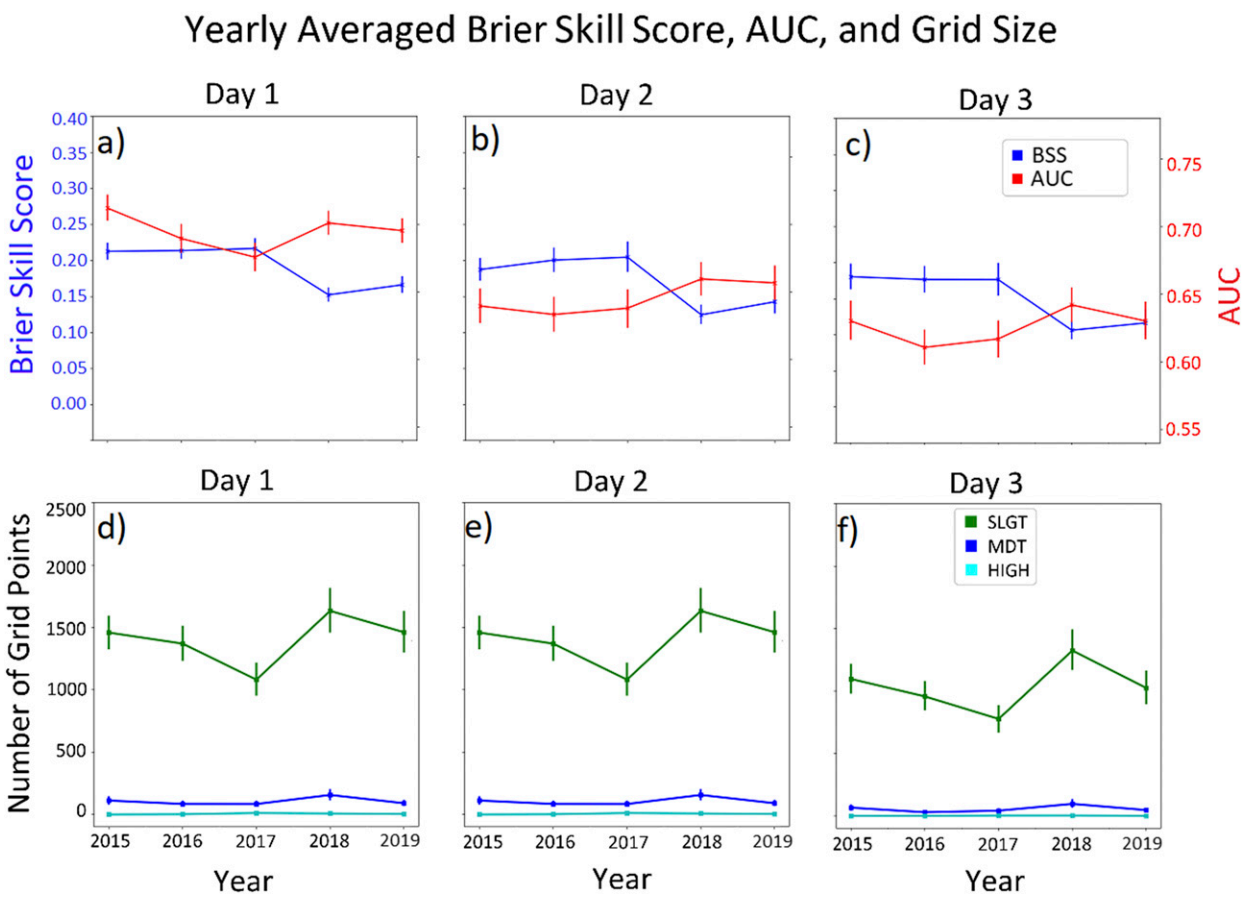

FIG. 12. Time series of yearly averaged BSS and AUC comparing (top) ERO to verification and (bottom) average number of grid points for each ERO object by category for (a),(d) Day 1; (b),(e) Day 2; and (c),(f) Day 3. Error bars represent the 2.5th and 97.5th confidence intervals using a bootstrapping method.

By design, the PP method employs both a neighborhood approach and a Gaussian filter, which tends to result in circular shapes. ERO forecasts are drawn by forecasters and do not have this constraint. Hence, when ERO contours take more elongated or complex shapes, the PP will likely appear as a combination of one or more circular regions. In addition, the MRGL category was largely not considered in this study since these contours can be large and irregularly shaped and do not typically have as high an impact as the MDT and HIGH categories. Furthermore, the PP method may not show mesoalpha or smaller risk regions (e.g., burn scars) that can work into the forecast process for the Day 1 ERO (e.g., when antecedent flash flooding conditions are in place), since they will be smoothed out by the 105-km Gaussian filter.

Instead the PP method is designed to show the general magnitude, placement, and size of the ''observed'' SLGT, MDT and HIGH regions. For example, the PP method can be used to infer if an ERO MDT forecast verified for a given day and if the aforementioned MDT forecast targeted the correct region. As an example, four case studies are shown highlighting both the PP (shaded) and the 0900 UTC operational ERO (contoured) issued one day prior valid ending 1200 UTC 1 July 2018, 24 July 2018, 18 August 2018, and 17 September 2018 (Fig. 14). The PP probabilities support the MDT issued in the ERO in Fig. 14a, although the ERO is too far north. On 24 July 2019 (Fig. 14b), the MDT region over the mid-Atlantic is strongly supported by the verifying PP, but the PP suggests a MDT occurred in Colorado, rather than a SLGT. The small SLGT over New Mexico and Texas valid 18 August 2018 (Fig. 14c) is not supported by PP, while the PP suggests a MDT occurred over a small portion of eastern NY. Finally, the HIGH risk issued during Hurricane Florence is strongly supported by the PP over a large area of North Carolina (Fig. 14d).

6. Conclusions

This study presents a comprehensive verification of the Weather Prediction Center (WPC) Excessive Rainfall Outlook (ERO; example in Fig. 1) valid from 1 January 2015 to 31 December 2019. The WPC ERO consists of four probabilistic categories; marginal (MRGL), slight (SLGT), moderate (MDT), and high (HIGH). Most results in this study omit the MRGL category since it was introduced on 1 August 2016. Verification is performed against a variety of flooding observations and proxies, including Stage IV analysis exceeding flash flood guidance (FFG), Stage IV analysis exceeding 5-yr

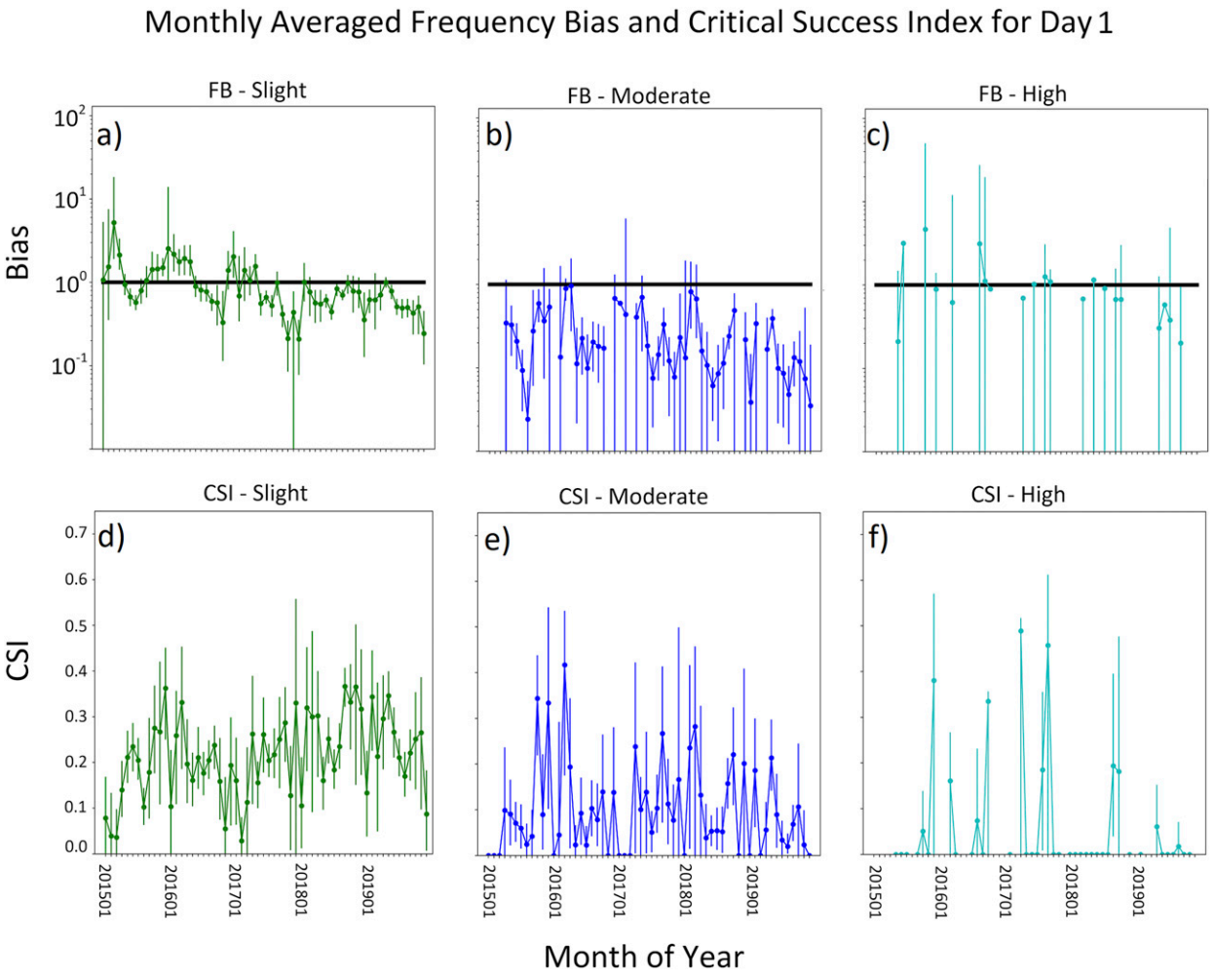

FIG. 13. Monthly averaged (top) frequency bias and (bottom) critical success index comparing the practically perfect to the ERO for (a),(d) slight; (b),(e) moderate; and (c),(f) high. Error bars represent the 2.5th and 97.5th confidence intervals using a bootstrapping method.

average recurrence interval, U.S. Geological Survey river gauge observations, and National Weather Service Local Storm Reports. The collection of all different verification sources is referred to as the Unified Flooding Verification System (UFVS). Two different ERO verifications are performed, the first comparing the ERO to FFG exceedances (consistent with the current operational definition of the ERO) and the second comparing the ERO to the entire UFVS.

Considerable attention is given to the appropriate skill metrics in evaluating the operational ERO both when assessing single events and bulk performance. This study utilizes Brier skill score (BSS) and area under the relative operating characteristic (AUC) to assess ERO skill in bulk and develops a practically perfect (PP; Fig. 2) method to evaluate both daily and bulk ERO skill. The PP method is developed by interpolating the UFVS binomial data (100% = flood, 0% = no flood) to a grid, applying a 40-km radius, setting a proxy fractional value (PFV), and applying a Gaussian smoother to create a "probabilistic observation." Sensitivity studies are performed by varying the PFV and Gaussian smoother and comparing the PP results to the ERO throughout 2017. The optimal parameter space, defined in terms of reduced bias and error (Fig. 3) for all ERO categories, has a 0.8 PFV for FFG and average recurrence interval exceedances, a 1 PFV for the flooding observations, and a 105-km Gaussian filter. A 5-yr verification of the optimal PP selection exhibits a near zero, positive, and near zero bias for SLGT, MDT, and HIGH, respectively (Fig. 4a). In addition, given the PP predicted a specific ERO category, there was a 79%, 30%, and 31% chance of an ERO SLGT, MDT, and HIGH, respectively, being realized (Fig. 4c). Overall, the PP method is consistent with the ERO on the average and can be used to evaluate ERO performance in bulk (Fig. 13), spatially (Fig. 5), and for individual events (Fig. 14).

ERO issuance frequency is evaluated for the SLGT, MDT, and HIGH categories from Days 1 to 3 (Figs. 6 and 7). The greatest frequency of issuances spans from the Texas Gulf Coast northward to the Midwest and eastward to the

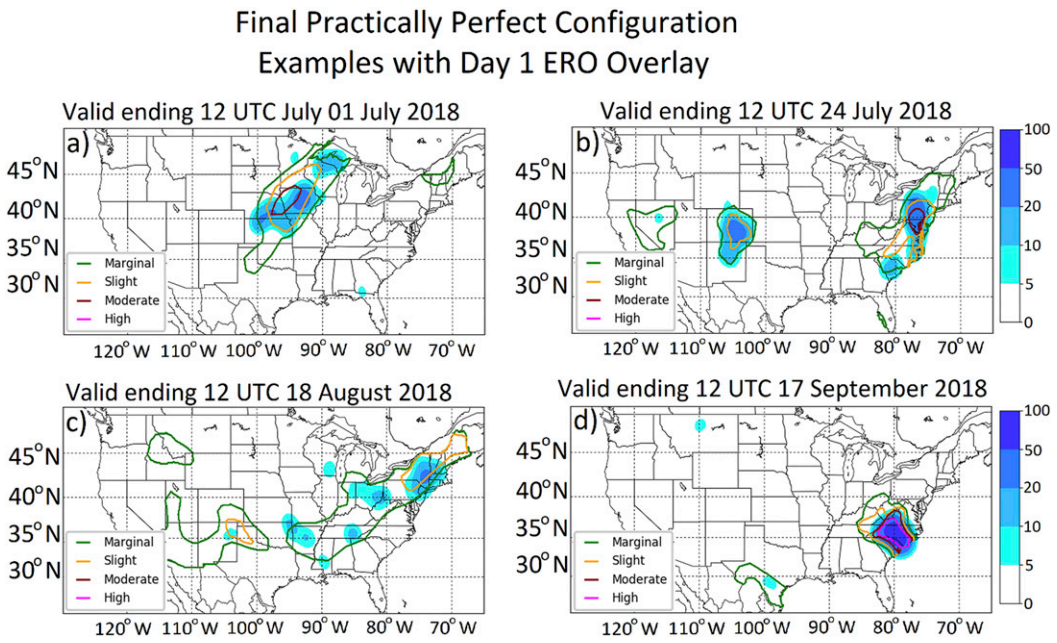

FIG. 14. Comparison of the optimal practically perfect configuration (blue filled contours) with the Day 1 ERO (open contours) for four examples valid ending at (a) 1200 UTC 1 Jul 2018, (b) 1200 UTC 24 Jul 2018, (c) 1200 UTC 18 Aug 2018, and (d) 1200 UTC 17 Sep 2018.

Appalachians. Most HIGH issuances are associated with tropical or posttropical cyclones. There is a strong seasonality to ERO issuances with less activity during the winter primarily associated with synoptic storms and more activity in the summer caused by convection, tropical cyclones, and the Southwest monsoon.

The fractional coverage of UFVS occurrences within each ERO contour is assessed to evaluate the reliability of the ERO (Figs. 8 and 9). The ERO is reliable against FFG exceedances for each probabilistic definition when considering the entire 5-yr period (Fig. 8), but fractional coverage just exceeds the probabilistic definition after August 2016 for Days 1–3 SLGT and Day 1 MDT (Fig. 9). When fractional coverage exceeds the ERO probabilistic definition, this may suggest that forecasters should draw larger contours or issue the respective category more frequently.

The ERO exhibits greater skill at shorter lead times in terms of BSS and AUC (Fig. 10) when progressing closer to the event from Days 3 to 1. In addition, the ERO exhibits greater skill in the winter compared to the summer in terms of BSS, with large monthly variability between different years (Fig. 11). For the 5 years analyzed, BSS, AUC, and average contour size exhibit little change (Fig. 12). However, there may be a small decrease in the monthly average bias along with a small increase in CSI for SLGT when comparing the ERO against PP (Fig. 13).

The ERO is intended to focus on the potential for flash flooding in the Day 1–3 range, and hence can be considered a recommender to NWS forecast offices for the issuance of flash flood watches. This study is intended to inform the public of the ERO forecast product verification effort. A future WPC paper is currently being developed that will focus on the forecast-process that goes into the manual generation of the ERO.

The verification results presented in this study will be used to form a baseline agency goal for ERO skill in the future. Many of these results have also been adapted to a WPC internal ERO verification website, which is readily available to WPC forecasters. This WPC website updates daily and presents daily, monthly, and annual ERO verification statistics including AUC, BSS, fractional coverage, and PP and will gradually be transitioned to a public interface.

The PP methodology developed in this paper can be further utilized to enhance the ERO verification. This PP product is designed to closely match the operational ERO's SLGT, MDT and HIGH thresholds and can be used in the evaluation of individual events or to assess bulk performance in the long term. However, if one assumes that the ERO contours take the form of objects with low complexity (e.g., simple ovals), PP and ERO objects can be identified and compared to look for displacement and magnitude biases. Object-based identification software exists within the Model Evaluation Tools (Bullock et al. 2016) and is applicable to ERO objects. Future work will utilize object-based verification of the ERO compared to the PP contours, with a strong focus on forecaster displacement biases. Additional future work will explore regional variability in ERO performance, particularly associated with tropical cyclones and the Southwest monsoon. For interested users, an archive of the UFVS observations and ERO archive is available at: https://ftp.wpc.ncep.noaa.gov/ERO_verif/.

Acknowledgments. This work was funded by NOAA Cooperative Agreement Award NA17OAR4320101. We thank

three anonymous reviewers for their very helpful comments in improving this manuscript.

Data availability statement. The authors will provide all data used in this article upon request. Please contact Michael Erickson (mjaerickson@gmail.com) for more information.

APPENDIX

Verification Metrics

Frequency bias (FB), hit rate, false alarm ratio, and critical success index (CSI) are computed by considering a set of dichotomous forecast and observation pairs at various ERO thresholds using a 2 × 2 contingency table (Wilks 2011). The dichotomous events are defined as

a = observation yes, forecast yes,

b = observation no, forecast yes,

c = observation yes, forecast no.

Verification metrics are computed by

$$\text{FB} = \frac{a+b}{a+c}, \tag{A1}$$

$$\text{hit rate} = \frac{a}{a+c}, \tag{A2}$$

$$\text{false alarm ratio} = \frac{b}{a+b}, \tag{A3}$$

$$\text{CSI} = \frac{a}{a+b+c}. \tag{A4}$$

The Brier score (BS; Wilks 2011) is analogous to the mean square error of the forecast–observation pairs and is computed using the following equation:

$$\text{BS} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2,$$

where y is the forecast, o is the observation, and k denotes the number of n forecast–observation pairs. The relative operating characteristic (ROC) measures hit rate versus false alarm rate, while area under ROC assesses the ability of the forecast to discriminate between events and nonevents.

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.

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of quantitative precipitation reforecasts over the southeastern United States. *Wea. Forecasting*, **29**, 1199–1207, https://doi.org/10.1175/WAF-D-14-00055.1.

Brennan, M. J., J. 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, https://ams.confex.com/ams/pdfpapers/138022.pdf.

Bullock, R., B. Brown, and T. Fowler, 2016: Method for object-based diagnostic evaluation. NCAR Tech. Note NCAR/TN-532+STR, 84 pp., https://doi.org/10.5065/D61V5CBS.

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.

Cookson-Hills, P., D. J. Kirshbaum, M. Surcel, J. G. Doyle, L. Fillion, D. Jacques, and S.-J. Baek, 2017: Verification of 24-h quantitative precipitation forecasts over the Pacific Northwest from a high resolution ensemble Kalman filter system. *Wea. Forecasting*, **32**, 1185–1208, https://doi.org/10.1175/WAF-D-16-0180.1.

Cooley, D. S., 1978: The excessive rainfall potential outlook. NOAA/NWS Tech. Procedures Bull. 239, 7 pp.

Cosgrove, B., and C. Klymmer, 2016: The National Water Model. NOAA, accessed 29 July 2019, https://water.noaa.gov/about/nwm.

Cuo, L., T. C. Pagano, and Q. J. Wang, 2011: A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. *J. Hydrometeor.*, **12**, 713–728, https://doi.org/10.1175/2011JHM1347.1.

Erickson, M. J., and J. A. Nelson Jr., 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.

——, J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification results from the 2017 HMT-WPC flash flood and intense rainfall experiment. *J. Appl. Meteor. Climatol.*, **58**, 2591–2604, https://doi.org/10.1175/JAMC-D-19-0097.1.

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.

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.

Guirguis, K., A. Gershunov, T. Shulgina, R. E. S. Clemesha, and F. Martin Ralph, 2019: Atmospheric rivers impacting Northern California and their modulation by a variable climate. *Climate Dyn.*, **52**, 6569–6583, https://doi.org/10.1007/s00382-018-4532-5.

Halley Gotway, J., and Coauthors, 2018: Model Evaluation Tools version 8.0 (METv8.0): User's guide. Developmental Testbed Center Rep., 431 pp., https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v8.0.pdf.

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

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. national blend of models for statistical post-processing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

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.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36–60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **31**, 647–661, https://doi.org/10.1175/WAF-D-15-0143.1.

Kendall, M. G., 1975: *Rank Correlation Measures*. Charles Griffin, 202 pp.

Lamers, A., 2019: Pre-monsoon season summary of important information. WPC-WFO-RFC Southwest Monsoon Collaboration Call.

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**, 13–22, https://doi.org/10.15191/nwajom.2018.0602.

Luitel, B., G. Villarini, and G. A. Vecchi, 2018: Verification of the skill of numerical weather prediction models in forecasting rainfall from U.S. landfalling tropical cyclones. *J. Hydrol.*, **556**, 1026–1037, https://doi.org/10.1016/j.jhydrol.2016.09.019.

Ma, S., C. Chen, H. He, D. Wu, and C. Zhang, 2018: Assessing the skill of convection-allowing ensemble forecasts of precipitation by optimization of spatial-temporal neighborhoods. *Science*, **9**, 43, https://doi.org/10.3390/ATMOS9020043.

Mann, H. B., 1945: Nonparametric tests against trend. *Econometrica*, **13**, 245–259, https://doi.org/10.2307/1907187.

Marchok, T., R. Rogers, and R. Tuleya, 2007: Validation schemes for tropical cyclone quantitative precipitation forecasts: Evaluation of operational models for U.S. landfall cases. *Wea. Forecasting*, **22**, 726–746, https://doi.org/10.1175/WAF1024.1.

National Weather Service, 2017: Summary of natural hazard statistics for 2017 in the United States. NOAA, 3 pp., https://www.weather.gov/media/hazstat/sum17.pdf.

NCEP, 2019: Ensemble situational awareness table. Accessed 29 July 2019, https://satable.ncep.noaa.gov/naefs/.

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.

Newman, K., and Coauthors, 2019: Model Evaluation Tools version 8.1 (METv8.1): User's guide. Developmental Testbed Center Rep., 437 pp., https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v8.1.pdf.

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

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.

Perica, S., and Coauthors, 2013: *Precipitation-Frequency Atlas of the United States*. Vol. 9, NOAA Atlas NESDIS 14, 171 pp.

Pyle, M. E., and G. S. Manikin, 2018: The High-Resolution Ensemble Forecast (HREF) System and applications for aviation forecasting. *Sixth Aviation, Range, and Aerospace Meteorology Special Symp.*, Austin, TX, Amer. Meteor. Soc., 2.3, https://ams.confex.com/ams/98Annual/webprogram/Paper334406.html.

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

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.

Sharma, S., and Coauthors, 2017: Eastern U.S. verification of ensemble precipitation forecasts. *Wea. Forecasting*, **32**, 117–139, https://doi.org/10.1175/WAF-D-16-0094.1.

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

Water Resources Services, 2017: NWS annual flood loss summary reports to U.S. Army Corps of Engineers. 2 pp., https://www.weather.gov/media/water/WY17%20Flood%20Deaths%20and%20Direct%20Damagesv2.pdf.

Wick, G. A., P. J. Neiman, and F. M. Ralph, 2013: Description and validation of an automated objective technique for identification and characterization of integrated water vapor signature of atmospheric rivers *IEEE Trans. Geosci. Remote Sens.*, **51**, 2166–2176, https://doi.org/10.1109/TGRS.2012.2211024.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

WPC, 2013: About the excessive rainfall forecast. Accessed 29 July 2019, https://www.wpc.ncep.noaa.gov/qpf/about_excess.shtml.

——, 2016: What's new at WPC. Accessed 29 July 2019, https://www.wpc.ncep.noaa.gov/html/whats_new.shtml.

——, 2017: WPC verification page. Accessed 29 July 2019, https://www.wpc.ncep.noaa.gov/images/hpcvrf/WPCmdlsd110yrly.gif.

——, 2019: Product information. Accessed 29 July 2019, https://www.wpc.ncep.noaa.gov/html/fam2.shtml#excessrain.

Yang, F., and V. Tallapragada, 2018: Implementation and evaluation of the NOAA next generation global prediction system with FV3 dynamical core and advanced physics. *Eighth Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., 1.4, https://ams.confex.com/ams/98Annual/webprogram/Paper329963.html.

Zheng, M., E. K. Chang, B. A. Colle, Y. Luo, and Y. Zhu, 2017: Applying fuzzy clustering to a multimodel ensemble for U.S. East Coast winter storms: Scenario identification and forecast verification. *Wea. Forecasting*, **32**, 881–903, https://doi.org/10.1175/WAF-D-16-0112.1.