# An Adaptive Approach for the Calculation of Ensemble Gridpoint Probabilities

BENJAMIN T. BLAKE,[a,b] JACOB R. CARLEY,[b] TREVOR I. ALCOTT,[c] ISIDORA JANKOV,[c,d] MATTHEW E. PYLE,[b] SARAH E. PERFATER,[a,e] AND BENJAMIN ALBRIGHT[e,f]

[a] *I.M. Systems Group, Inc., Rockville, Maryland*
[b] *NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*
[c] *NOAA/Earth System Research Laboratory, Boulder, Colorado*
[d] *CIRA, Colorado State University, Fort Collins, Colorado*
[e] *NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland*
[f] *Systems Research Group, Inc., Colorado Springs, Colorado*

(Manuscript received 28 February 2018, in final form 15 June 2018)

ABSTRACT

Traditional ensemble probabilities are computed based on the number of members that exceed a threshold at a given point divided by the total number of members. This approach has been employed for many years in coarse-resolution models. However, convection-permitting ensembles of less than ~20 members are generally underdispersive, and spatial displacement at the gridpoint scale is often large. These issues have motivated the development of spatial filtering and neighborhood postprocessing methods, such as fractional coverage and neighborhood maximum value, which address this spatial uncertainty. Two different fractional coverage approaches for the generation of gridpoint probabilities were evaluated. The first method expands the traditional point probability calculation to cover a 100-km radius around a given point. The second method applies the idea that a uniform radius is not appropriate when there is strong agreement between members. In such cases, the traditional fractional coverage approach can reduce the probabilities for these potentially well-handled events. Therefore, a variable radius approach has been developed based upon ensemble agreement scale similarity criteria. In this method, the radius size ranges from 10 km for member forecasts that are in good agreement (e.g., lake-effect snow, orographic precipitation, very short-term forecasts, etc.) to 100 km when the members are more dissimilar. Results from the application of this adaptive technique for the calculation of point probabilities for precipitation forecasts are presented based upon several months of objective verification and subjective feedback from the 2017 Flash Flood and Intense Rainfall Experiment.

## 1. Introduction

The value and skill offered by convection-permitting models (CPMs) has been recognized in many operational numerical weather prediction (NWP) centers in the past decade through the general widespread implementation of deterministic CPMs (e.g., Saito et al. 2006; Smith et al. 2008; Rogers et al. 2009; Baldauf et al. 2011; Seity et al. 2011; Tang et al. 2013). CPMs have been shown to develop storms with more realistic attributes that are not present at comparatively coarser spatial resolutions where convection is parameterized and, consequently, they produce better forecasts (e.g., Done et al. 2004; Kain et al. 2006; Lean et al. 2008; Roberts and Lean 2008; Weisman et al. 2008; Schwartz et al. 2009; Clark et al. 2010); however, errors grow quickly at such a fine scale (Lorenz 1969; Hohenegger and Schär 2007; Melhauser and Zhang 2012; Radhakrishna et al. 2012). Many NWP centers, including the National Centers for Environmental Prediction (NCEP), have begun to implement a convection-permitting ensemble (CPE) prediction system or have strategies in place to do so (Gebhardt et al. 2011; Peralta et al. 2012; Tennant 2015; Rogers et al. 2017). A CPE has the capability to provide information about a wide range of solutions that are related to the timing, location, and structure of convection and are sensitive to small environmental changes, which helps to quantify forecast uncertainty.

Some obstacles still remain in the extraction of useful information from CPEs, such as effective calibration techniques and the generation of ensemble probabilistic output. Probabilistic guidance allows forecasters to

*Corresponding author*: Benjamin Blake, benjamin.blake@noaa.gov

DOI: 10.1175/WAF-D-18-0035.1

quantify uncertainty and allows users to make better decisions compared to those made with yes–no forecasts (Murphy 1993). As has been recognized in the forecast verification community (e.g., Gallus 2002; Baldwin and Kain 2006; Roberts and Lean 2008), the skill at the gridpoint level in CPMs tends to be relatively poor, owing to well-known challenges associated with the double penalty (e.g., Gilleland et al. 2009), when the spatial scale of a simulated phenomenon is less than or equal to the scale of the mean spatial forecast errors. Objective point verification methods require a nearly perfect match for a forecast to be considered skillful; hence, traditional verification methods tend to favor comparatively smoother fields of lower-resolution models over the more realistic fields present at higher spatial resolutions (e.g., Wolff et al. 2014). In addition, CPEs are generally underdispersive (Hohenegger et al. 2008; Novak et al. 2008; Gebhardt et al. 2011; Vié et al. 2011; Romine et al. 2014; Schwartz et al. 2014), yielding low spread that can be attributed to model biases or not accounting for all potential sources of forecast error. Addressing the double-penalty problem in CPMs, along with the need to increase spread in CPEs, extends to the generation of useful probabilistic output and has resulted in the development of several different ensemble postprocessing approaches, including fractional coverage (Theis et al. 2005; Roberts and Lean 2008; Schwartz et al. 2010) and neighborhood maximum value methods (Harless et al. 2010; Jirak et al. 2012; Hitchens et al. 2013).

Fractional coverage probabilities, which are similar in approach to the verification metric known as the fractions skill score (FSS; Roberts and Lean 2008), are effectively point probabilities that acknowledge the existence of spatial uncertainty in a forecast. These probabilities represent the fraction of points from all members within a fixed radius of influence around each grid point that exceed a threshold. Fractional coverage is an example of a spatial filtering technique for increasing ensemble spread, where all grid points within the radius of influence are considered ensemble "members." Other approaches, like neighborhood maximum value, transform the forecast from a point probability to an areal probability (i.e., the probability of exceeding a threshold within some radius of a grid point). Schwartz and Sobash (2017) provide a thorough analysis of these two approaches, referring to fractional coverage as a neighborhood approach for deriving grid-scale probabilities and to neighborhood maximum value as a neighborhood approach for deriving non-grid-scale probabilities.

An alternative method of characterizing forecast ensemble spatial uncertainty was recently proposed by Dey et al. (2016). When applying filtering techniques by using only one scale over the whole domain, as with a fixed radius of influence for the fractional coverage approach, geographic and temporal variability in ensemble spread is ignored (Dey et al. 2014). These differences in spread arise because different phenomena (e.g., convective, frontal, or winter precipitation) may exhibit varying degrees of predictability that can evolve over time. Hence, it is useful to process forecasts in a manner that preserves some spread characteristics from the raw ensemble. Dey et al. (2016) proposed the ensemble agreement scale (EAS) technique to provide an estimate of spatial agreement among CPE members at each grid point by varying the radius of influence at each ensemble grid point according to member–member similarity criteria.

This paper proposes a refinement of the traditional fractional coverage method via the implementation of the EAS approach to provide locally adaptive radii of influence. If the forecasts at a grid point are in excellent agreement, then a small radius is utilized for the calculation of ensemble probabilities (and vice versa). This method would be applicable to a wider range of scenarios where ensemble spread is low as a result of inherently greater predictability (e.g., orographic precipitation, lake-effect snow, very short-term forecasts, etc.). When there is strong agreement among members, a large uniform radius is not appropriate; in such cases, the traditional fractional coverage approach can reduce the probabilities for these potentially well-handled events. While we know the actual magnitude of spatial spread in the ensemble is usually inadequate, we are hypothesizing that the spread–skill relationship is sufficiently adequate such that it is desirable to preserve some of that information in the postprocessed product, and doing so with the EAS technique results in better forecasts. These refined point probabilities will be compared to fractional coverage point probabilities that use a fixed radius of influence, as well as traditional ensemble point probabilities. Section 2 provides an overview of the experimental design, the different approaches for generating probabilities, and the verification methods. Section 3 presents results from an idealized experiment along with objective and subjective verification statistics, followed by discussion and conclusions in section 4.

## 2. Methods

### a. *Experimental design*

Version 2 of the High Resolution Ensemble Forecast system (HREFv2; Rogers et al. 2017) herein refers to

the collection of model runs comprising a multimodel ensemble of eight CPMs. There are four ARW-based members, two of which are time-lagged, and four NMMB-based members, two of which are time lagged. Here, the HREFv2 membership was combined into probabilistic products not directly related to the operational HREFv2 QPF probability products, which are based on the neighborhood maximum approach. The HREFv2 data in this study covered forecasts from the 0000 and 1200 UTC cycles beginning on 3 February 2017 and running through 30 September 2017. However, the system was not operational during the period of the study and was, therefore, subject to outages.

The primary focus was on 6-hourly quantitative precipitation forecasts (QPFs). Thus, the probabilities presented herein represent the chance of 0.5 or 1 in. of precipitation falling over a 6-h period. The 6-hourly Stage IV precipitation product was used as the verifying dataset for quantitative precipitation estimation (QPE; Lin and Mitchell 2005). Prior to the generation of probabilities, all accumulated precipitation values were budget interpolated (Accadia et al. 2003) to NCEP grid 227, a 5-km grid covering the contiguous United States, and the QPF from each ensemble member was bias corrected via a quantile-mapping technique (Scheuerer and Hamill 2015; Alcott et al. 2017). Bias correction coefficients were determined by calculating a second-order best fit between precipitation quantiles in the 50 most recent QPF–QPE pairs.

### b. Probability methods

Three different methods for generating ensemble probabilities were evaluated. The first approach was the ''Point'' method, which calculated the average of the binary probabilities from each of the ensemble members at a grid point. For each member, the binary probability at a grid point was 1 if the specified threshold was exceeded and 0 otherwise. These probabilities represented traditional ensemble point probabilities.

The second method was the fractional coverage (Frac) approach (Theis et al. 2005; Schwartz et al. 2010). The point probability field was calculated in a way that accounted for the spatial uncertainty in CPE probabilistic forecasts, where all grid points whose centers fell within a radius of influence around a given point were considered part of the spatial filter. A 100-km radius was chosen for this study because it produced the most reliable probabilistic forecasts of warm season 6-h precipitation during 2015–16 with the High Resolution Rapid Refresh Time-Lagged Ensemble (HRRR-TLE; not shown). For the Frac approach, the probability at grid point $i$ was therefore given by

$$P_i = \frac{1}{N_b N_{\text{ens}}} \sum_{m=1}^{N_b} \sum_{k=1}^{N_{\text{ens}}} \text{BP}_{km}. \tag{1}$$

The number of grid points within $r = 100$ km of grid point $i$ was $N_b$, $N_{\text{ens}}$ was the number of ensemble members, and $\text{BP}_{km}$ was the binary probability for point $m$ in $N_b$ for ensemble member $k$. Theis et al. (2005) and Roberts and Lean (2008) utilized a square spatial filter around each grid point, while a circular spatial filter was used in this study, similar to the approach taken by Schwartz et al. (2010).

The third approach, the EAS fractional coverage (EAS) method, is proposed as a possible refinement to the Frac method. The only difference between Frac and its EAS counterpart is that a different radius was applied at each grid point based on the local agreement scale among the member forecasts. The radius of influence for the spatial filter was defined as the smallest scale over which the member forecasts were deemed suitably similar (Dey et al. 2016). The similarity criteria take the following form:

$$D_{ij} = \left\{ \begin{array}{ll} \dfrac{(A-B)^2}{(A^2+B^2)}, & \text{if} \quad A>0 \wedge B>0 \\ 1, & \text{if} \quad A=0 \vee B=0 \end{array} \right\}, \tag{2}$$

$$D_{\text{crit},ij} = \alpha, \quad \text{and} \tag{3}$$

$$D_{ij} \le D_{\text{crit},ij}. \tag{4}$$

The average value of the exceedance grid within the spatial filter for two forecasts was represented by $A$ and $B$ in (2). An exceedance grid comprises ones and zeros for points that did and did not exceed the threshold, respectively. One must calculate $D_{ij}$ for all member–member comparisons to obtain a mean value for $D_{ij}$; there were 28 possible pairs for the eight-member HREFv2 ensemble because ${}_8C_2 = 8!/[2!(8–2)!]$. The similarity criteria parameter $\alpha$ was related to the amount of bias tolerated, where $\alpha = 0$ signified no bias was tolerated at the grid scale and $\alpha = 1$ means any bias was tolerated. Consequently, for a smaller $\alpha$ it was more difficult to satisfy (4) at smaller radii, and vice versa. Dey et al. (2016) chose a value of 0.5 for $\alpha$; here, $\alpha$ was set to 0.1. Dey et al. used a different variation of (3), which yielded different values of $D_{\text{crit}}$ for different radii. Their formulation yielded low radii values because it was relatively easy to achieve (4), which increased the sharpness of the EAS probabilities. Here, we set $D_{\text{crit}}$ to a constant value $\alpha$ independent of radius. Our simpler specification of $D_{\text{crit}}$ made it progressively harder to satisfy (4) as the radius was decreased, and this decreased the sharpness

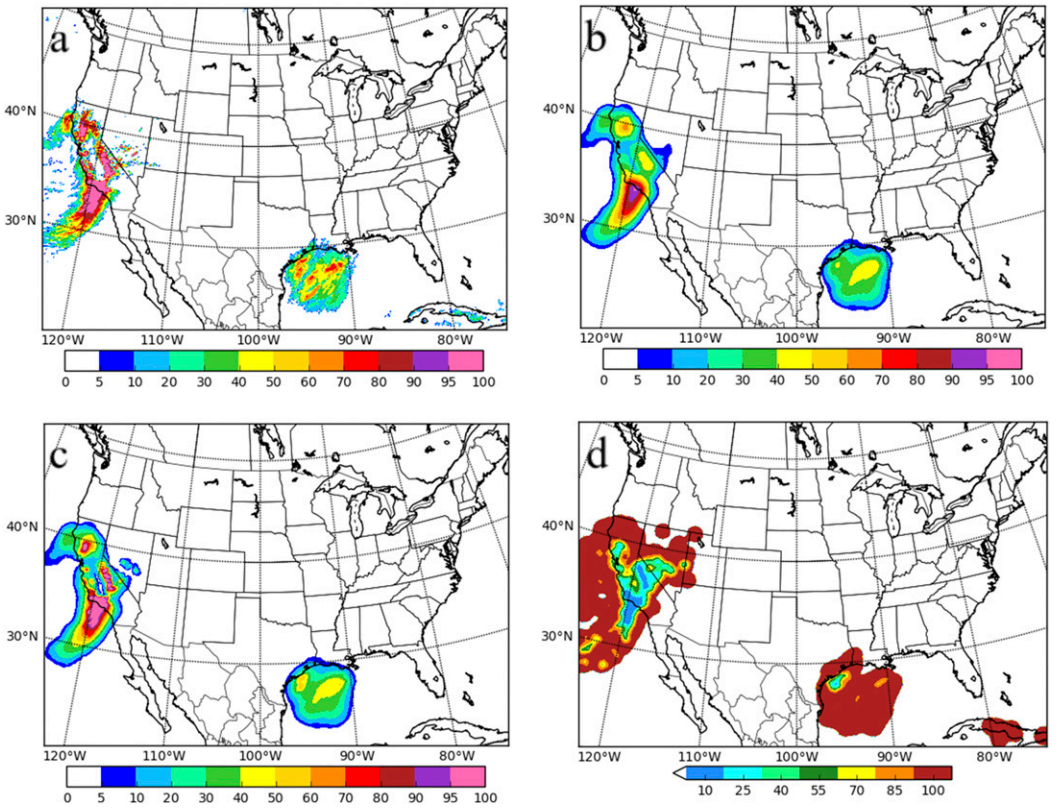


FIG. 1. Plots of probabilities (%) of 0.5 in. of precipitation accumulating over the 6-h period beginning at 1800 UTC 17 Feb 2017 and ending at 0000 UTC 18 Feb 2017 for the (a) Point, (b) Frac, and (c) EAS methods. These probabilities are 24-h forecasts from the 0000 UTC 17 Feb HREFv2 cycle. (d) A plot of the radii values (km) utilized by the EAS method.

of the EAS probabilities. We chose 100 km for the maximum possible radius, which was initially the radius at all grid points. For any grid points where (4) was satisfied, a suitably similar scale had been found, and the previous radii were overwritten. Next, the radius was decreased by $3 \times dx$, and the calculations were repeated for each successive radius. This iteration yielded possible radii of 10, 25, 40, 55, 70, 85, and 100 km. Before the radii were used in (1) they underwent a smoothing procedure to remove any spatial discontinuities. To accomplish the smoothing, a 20-km Gaussian kernel filter was applied to the radii field (Silverman 1986). The final probability field was then obtained via spatial filtering using the variable radii in (1).

An example of what the probability field for each approach looks like is provided in Fig. 1. These probabilities represented the chance of 0.5 in. of precipitation accumulating over the 6-h period beginning at 1800 UTC 17 February 2017 and ending at 0000 UTC 18 February 2017, and they were 24-h forecasts from the 17 February 0000 UTC cycle. The corresponding 6-h Stage IV precipitation is also shown for comparison (Fig. 2). The Point probabilities were quite sharp because they contained more 0% and 100% values than the other methods, especially over the Sierra Nevada and along the Pacific coast. The probabilities generally aligned well with the QPE. The Frac probabilities were much smoother and most of the 95%–100% regions that were in the Point probabilities were not present, except for a solitary maximum right along the California coast. However, most of the 95%–100% regions in the Frac probabilities exceeded 0.5 in. in 6 h. The EAS probabilities preserved many of the 95%–100% values over the Sierra Nevada and along the Pacific coast because the radii of influence were smaller (Fig. 1d), indicating the ensemble members were in good agreement over those locations.

### c. *Verification*

Version 6.0 of the Model Evaluation Toolkit's (MET) Grid Stat tool was used to generate objective verification statistics for the matched forecast and observation grids (Jensen et al. 2017). All of the forecast grid points in the verification region were matched to observation

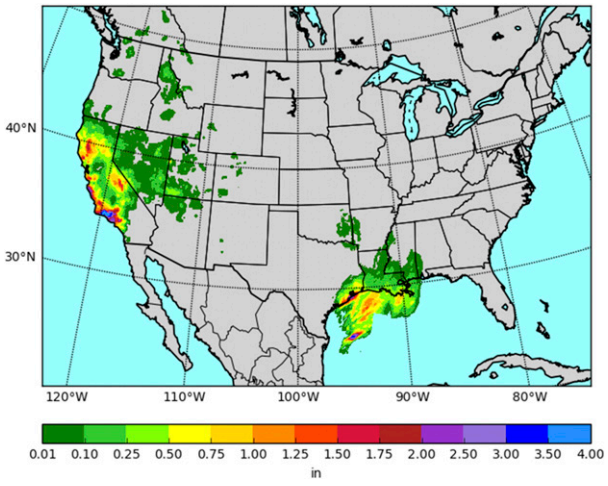


FIG. 2. A plot of the 6-h Stage IV accumulated precipitation (in.) valid from 1800 UTC 17 Feb to 0000 UTC 18 Feb 2017. The accumulated precipitation values were budget interpolated to NCEP grid 227, a 5-km grid covering the CONUS.

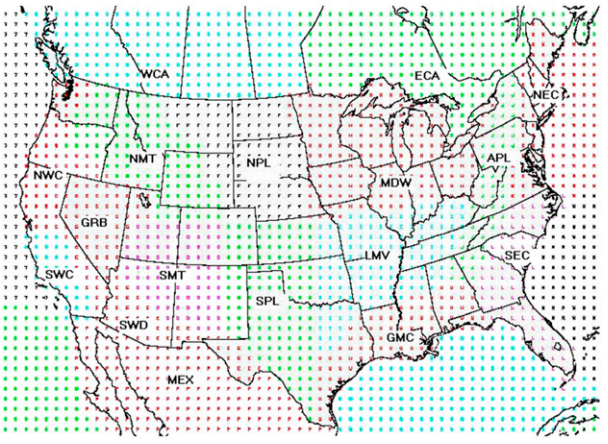


FIG. 3. The current NCEP verification subregions. There are 14 subregions over the CONUS: NEC, APL, SEC, GMC, LMV, MDW, NPL, SPL, NMT, SMT, GRB, NWC, SWC, and SWD. See text for subregion abbreviations.

grid points on the same grid, and all the matched points were then used to compute the statistics. Verification was conducted over three distinct regions, which comprise subsets of the NCEP verification subregions (Fig. 3). The East region included the Appalachian Mountains (APL), Northeast Coast (NEC), Southeast Coast (SEC), Midwest (MDW), Lower Mississippi Valley (LMV), and Gulf of Mexico Coast (GMC). The West included the Northern Mountains (NMT), Southern Mountains (SMT), Great Basin (GRB), Northwest Coast (NWC), Southwest Coast (SWC), and Southwest Desert (SWD). The CONUS region was the union of the East and West regions, and also included Northern Plains (NPL) and Southern Plains (SPL). The verification was confined to land-only points; while the Stage IV grid extended offshore, the predefined NCEP verification regions depicted in Fig. 3 were only over land.

The three primary objective verification techniques used in this study were the fractions Brier score (FBS; Roberts 2005), attributes diagrams (Hsu and Murphy 1986; Wilks 1995; Hamill 1997), and area under the receiver operating characteristic (ROC) curves (AUCs; Mason and Graham 1999; Jolliffe and Stephenson 2003; Hamill and Juras 2006). Each provides different insights into determining the accuracy and reliability of probabilistic forecasts.

The Brier score (Brier 1950; Jolliffe and Stephenson 2003) is often used to compare probabilistic forecasts to a dichotomous observational field. The BS is defined as the mean squared error of a probability forecast, and can be decomposed into reliability, resolution, and uncertainty. It quantifies the accuracy of a probabilistic forecast, which is the degree to which the forecasts and the observations agree. The BS averages the squared differences between pairs of forecast probabilities and the binary observation probabilities, where the probability is 1 if the event occurs and 0 if it does not occur. The BS ranges from 0 to 1, with 0 being a perfect forecast:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2. \tag{5}$$

The total number of grid points is represented by $N$, $i$ is a grid point, $p$ is the forecast probability, and $o$ is the binary observation probability. The BS is sensitive to the base rate or climatological frequency of an event. Forecasts of rare events, such as those exceeding 0.5 or 1.0 in. of accumulated precipitation in 6 h, can have a very small BS without having much actual skill because grid points with zero precipitation in either the observations or model forecast dominate the score. A variation on the BS is the FBS, where the dichotomous observational field is transformed into an analogous field of observation-based fractions. The forecast point probabilities and the observed fractions are then directly compared. Note that the FBS only differs from the traditional BS in that the observation values in (5) are no longer binary and are allowed to vary between 0 and 1 through the application of (1). Like the BS, the FBS is negatively oriented, where a score of 0 indicates perfect performance and a larger FBS indicates poorer correspondence between the model forecasts and the observations.

A reliability diagram is a graphical method for assessing the reliability, resolution, and sharpness of a

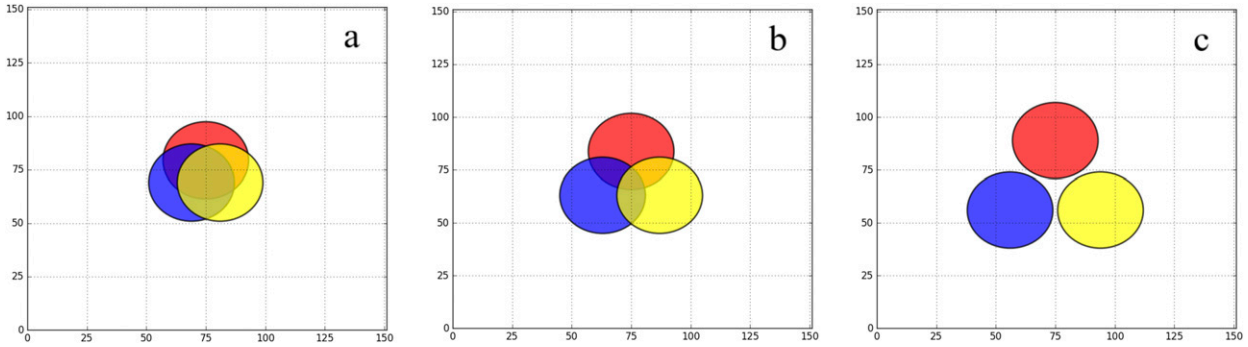


FIG. 4. Plots of three circles representing three different forecast scenarios for three ensemble member forecasts: (a) a large amount of overlap, (b) a small amount of overlap, and (c) no overlap. Each circle is the same size, and the centers of the circles are equidistant from one another.

probability forecast. Reliability is the degree to which the frequency of occurrence of the event agrees with the forecast probability. For instance, a perfect reliability means that a 40% probability forecast would be observed 40% of the time. Resolution is the ability of the forecasts to resolve the set of events into subsets with different relative frequencies of the event, representing the degree to which an event occurs relative to its climatological frequency. Climatology forecasts do not discriminate between events and nonevents and, therefore, have no resolution. Resolution is graphically represented by the distance of a point on the reliability diagram from the climatological frequency or "no resolution" line, and is computed by weighting the distance by the number of forecasts made at that forecast probability. Sharpness refers to the distribution of forecast probabilities over a verification sample. If probabilities of 0% and 100% are often utilized, the forecast is classified as sharp.

To create a reliability diagram, observation relative frequencies are plotted against the forecast probabilities. A perfectly reliable forecast would result in a diagonal line that is oriented from the bottom-left corner to the top-right corner of the plot. When the resultant curve deviates from the perfect reliability line, the forecasts are either underforecasts or overforecasts. Underforecasting is occurring when the curve is above the perfect reliability line and overforecasting is occurring when the curve is below the line. Another version of a reliability diagram includes the no-resolution (climatology) and "no skill" lines, and is referred to as an attributes diagram (Wilks 1995). The no-skill line, located halfway between the perfect reliability and no-resolution lines, depicts where resolution is equal to reliability and is in reference to the Brier score. If the curve drops below the no-skill line, the forecast is said to have no skill.

A ROC curve is utilized for evaluating the discrimination of a forecast. Discrimination is the ability of a forecast system to distinguish between occurrences and nonoccurrences of an event. To create a ROC curve, the probability of detection (POD) is plotted against the probability of false detection (POFD) at each forecast probability threshold. POD, or hit rate, is the fraction of events that were correctly forecast to occur. Conversely, the POFD, also referred to as the false alarm rate, is the proportion of nonevents that were forecast to be events. Here, the AUC is computed via trapezoidal integration (Mason 1982), where an AUC = 1 signifies a perfect forecast and an AUC = 0.5 indicates random forecasts (Marzban 2004). AUC values larger than ~0.7 are generally considered to represent useful probabilistic forecasts that discriminate between events and nonevents (Buizza et al. 1999). Correct forecasts of nonevents are determined over all locations in the domain, such that adding in large areas where little to no precipitation occurred improves the AUC by lowering the POFD. Consequently, in most rare-event forecasting applications, the points on a ROC diagram are located on the far left side (e.g., Schwartz and Sobash 2017); this signifies that the AUC is sensitive to the height of the "top most" point, which is associated with the lowest nonzero probability.

In addition to the aforementioned objective verification measures, the three probability methods were subjectively evaluated during the 2017 Flash Flood and Intense Rainfall Experiment (FFaIR; Perfater and Albright 2017). FFaIR is an annual experiment hosted by the Hydrometeorological Testbed (HMT) at the Weather Prediction Center (WPC). The experiment was conducted over the course of four weeks, beginning 19 June 2017 and ending 21 July 2017; the experiment was not held the week of 4 July. Each morning, participants were shown the three types of

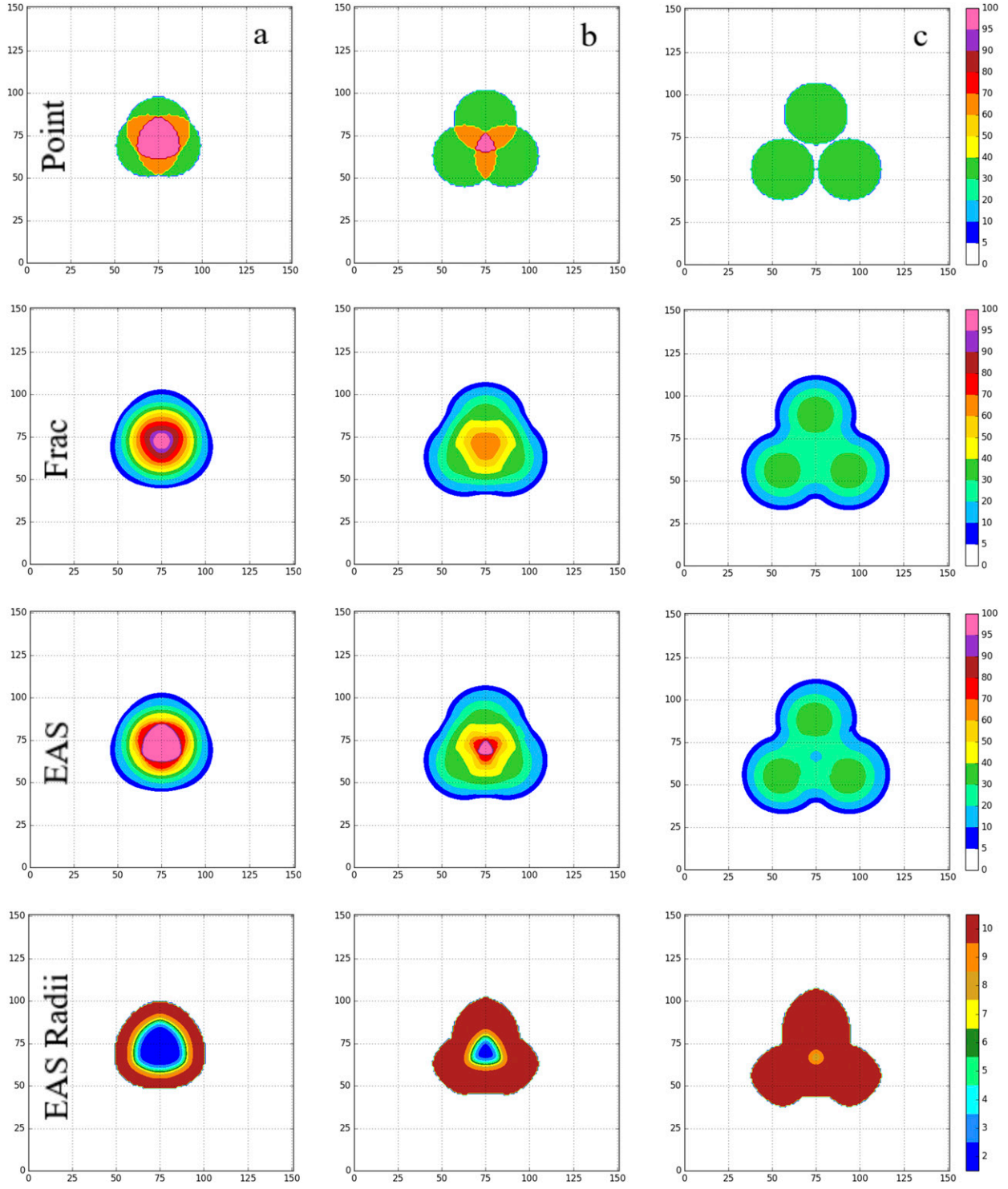


FIG. 5. Plots of the Point, Frac, and EAS probabilities (%) for each of the three scenarios depicted in Figs. 4a–c. We considered the arbitrary threshold to be exceeded for any grid points that fell inside the circle. Plots of the radii values utilized by the EAS method are also displayed.

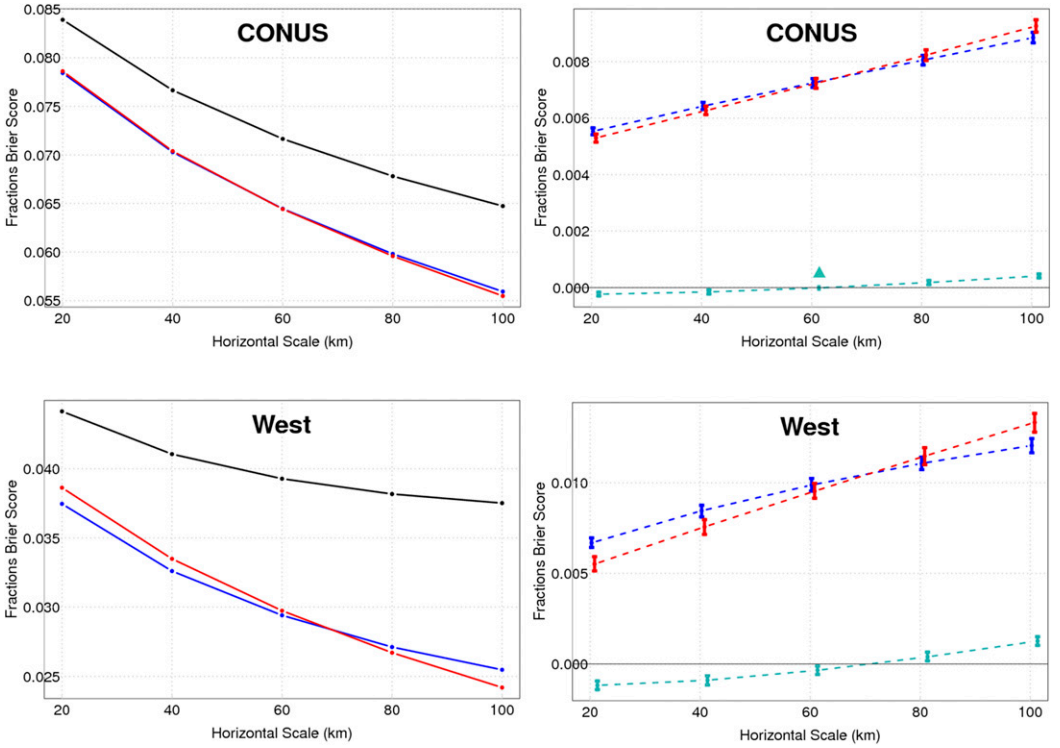


FIG. 6. Graphs of FBS as a function of spatial scale (km) for the Point (black line), Frac (red line), and EAS (blue line) approaches at the 0.5-in. threshold over the (top) CONUS and (bottom) West verification regions for 3 Feb–30 Sep 2017. Pairwise difference curves for Point − EAS (dashed blue line), Point − Frac (dashed red line), and EAS − Frac (dashed teal line) are also displayed. The 95% bootstrap confidence intervals for the difference curves were obtained using 1000 bootstrapping replications. If the differences are statistically significant, the confidence intervals are depicted in boldface. Triangles are associated with confidence intervals where the differences are not statistically significant.

experimental HREFv2 probabilities for 0.5 in. of QPF valid from 1800 to 0000 UTC on day 1 over a limited domain chosen by the testbed participants for the 6-h probabilistic flash flood (PFF) forecast. A 6-h Multi-Radar/Multi-Sensor (MRMS) system QPE was also displayed for verification. Participants were asked to comment on the utility of each approach and to subjectively rank the performance of each method on a scale from 1 to 10 based on how well the probabilistic values represented what occurred. A score of 1 indicated a very poor forecast, and a score of 10 represented a great forecast. Participants also utilized the methods in the daily experimental 6-h PFF forecast process.

## 3. Results

### *a. Idealized experiment*

An idealized experiment was conducted in order to better understand the strengths and weaknesses of each approach. Three circles were plotted for three different degrees of overlap: a large amount of overlap, a small amount of overlap, and no overlap (Fig. 4). Each circle was the same size, and the centers of the circles were equidistant from one another in each scenario. The different degrees of overlap were designed to represent three different types of forecast scenarios for an ensemble of three member forecasts. Large overlap between the circles represented a forecast where all three members were in good agreement with each other. On the other hand, less overlap between the circles signified a forecast where the members were not in good agreement. Here, we considered the arbitrary threshold to be exceeded for any grid points that fell inside the circle. Each probability method was applied to each scenario.

For the case where there was the most overlap, the three methods were all effective at identifying the region where the forecasts were in good agreement (Fig. 5a). The region of 95%–100% probability was smaller for Frac than for the Point and EAS methods. Note that the Point probabilities inherently have

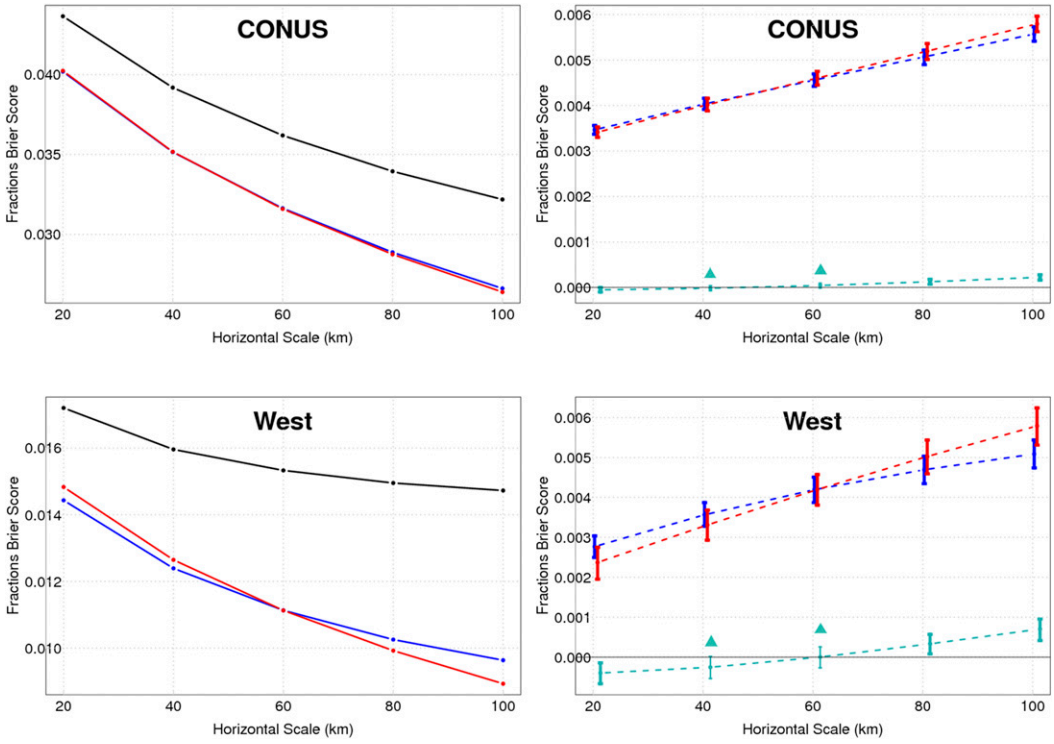


FIG. 7. As in Fig. 6, but for the 1.0-in. threshold.

sharp transitions between 0%, 33.3%, 66.7%, and 100%, while the Frac and EAS probabilities appeared smoother owing to their computations over a spatial filter via (1). Other groups have dealt with these sharp transitions by applying a Gaussian filter to smooth the probabilities (e.g., Harless et al. 2010; Hitchens et al. 2013). When there was a small degree of overlap (Fig. 5b), the Point method produced a region of 100% confidence where all three circles overlap. The values sharply dropped off to 66.7% and 33.3% once outside the region of interest. The Frac method depicted maximum probabilities of 60%–70% where all three circles overlap, and placed 30%–60% probabilities over the regions where two of the three circles overlap. EAS highlighted a small region of 95%–100% confidence where all three circles overlap, and depicted 30%–80% probabilities over the regions where two of the three circles overlap. For the scenario where there was no overlap between the circles (Fig. 5c), the Point approach had an area of 0% probability in the center where none of the circles were located. The Frac and EAS methods produced low probabilities of 10%–30% in the center because they took neighboring grid points into account, and they were nearly indistinguishable, except for a local minimum of 10%–20% in the EAS field corresponding to the lower radii values.

A forecaster looking at three forecasts that do not overlap but are relatively close to one another would likely place an emphasis on the center of those three forecasts or along the edges where they are closest together. The idealized circles highlight a major flaw with the traditional point probabilities, which depicted a 0% probability over the region where none of the circles overlap. The Frac method produced a much smoother probability field, resulting in nonzero probabilities over the region where none of the circles overlap, but it was difficult to achieve high probabilities with this approach. This was evident in the case of small overlap; Frac depicted 60%–70% probabilities where all three circles overlap, and the Point and EAS methods depicted 95%–100% probabilities. This scenario could be analogous to an orographic precipitation event, where points within 100 km are not located in the mountains and therefore should not be included in the fractional coverage calculation. The EAS approach was a compromise between Point and Frac, producing a smoother field than the Point method but a sharper field than the Frac method.

### b. *CPE forecasts*

To assess the mean squared error of the HREFv2 probability forecasts, plots of FBS as a function of

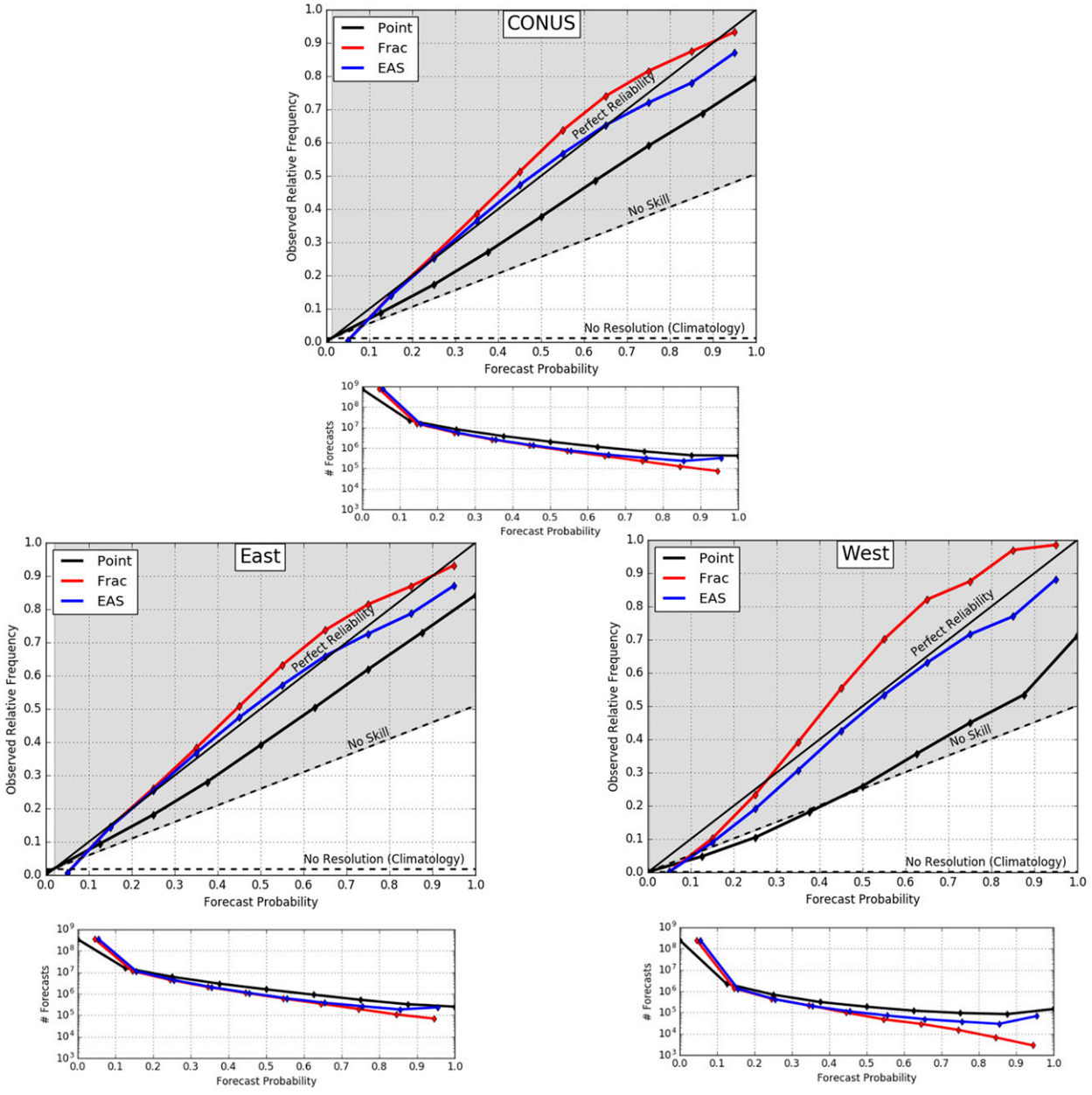


FIG. 8. Attributes diagrams for the Point (black line), Frac (red line), and EAS (blue line) approaches at the 0.5-in. threshold over the CONUS, East, and West verification regions for 3 Feb–30 Sep 2017. The perfect reliability line (solid black), the no-skill line (dashed black), and the no-resolution line (dashed black) are depicted. The total number of forecasts at each discrete probability value or within each probability bin are also plotted beneath each attributes diagram.

spatial scale were created for the 0.5-in. (Fig. 6) and 1.0-in. (Fig. 7) thresholds over the CONUS and West verification regions. Plots for the East region (not shown) were very similar to the CONUS plots. The observation probabilities were generated on each spatial scale, while the forecast probabilities were not modified; in other words, the EAS probabilities were calculated using different radii at each grid point, while the Frac probabilities were obtained via 100-km radii at all grid points. Pairwise difference curves for Point − EAS, Point − Frac, and EAS − Frac are also displayed. Bootstrap confidence intervals at the 95% level for the difference curves were obtained using 1000 replications in order to assess statistical significance. If the confidence interval of the difference curve did not encompass 0, then the differences were considered to be

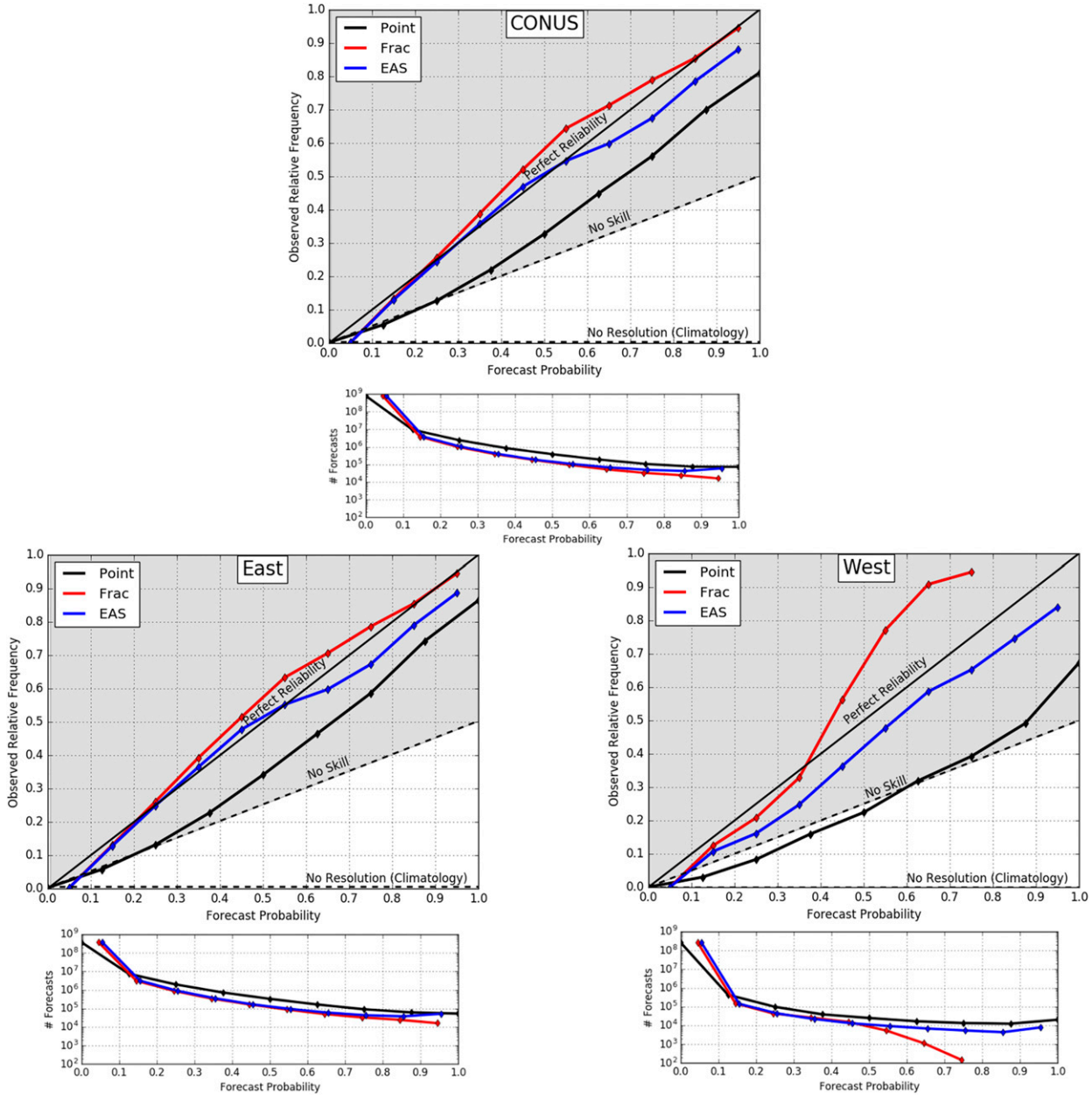


FIG. 9. As in Fig. 8, but for the 1.0-in. threshold.

statistically significant (Wilks 1995); in such instances, the confidence intervals were depicted in boldface. A triangle was placed over the confidence intervals that encompassed 0, or those where the differences were not statistically significant.

The Point probabilities had the highest FBS in all regions for both thresholds, indicating they were consistently associated with the largest forecast errors. The confidence intervals indicate that the differences between the Point approach and both fractional coverage methods were statistically significant in every region for all spatial scales on the observed grid. Generally, Frac and EAS had a similar FBS over the CONUS. Over the West, EAS had the lowest FBS at the 20-, 40-, and 60-km spatial scales for the 0.5-in. threshold, while Frac had the lowest FBS at the 80- and 100-km scales (Fig. 6)—scales which were closest to that used to compute the Frac probabilities (100 km). The differences between Frac and EAS were statistically significant at all spatial scales. The behavior at the 1.0-in.

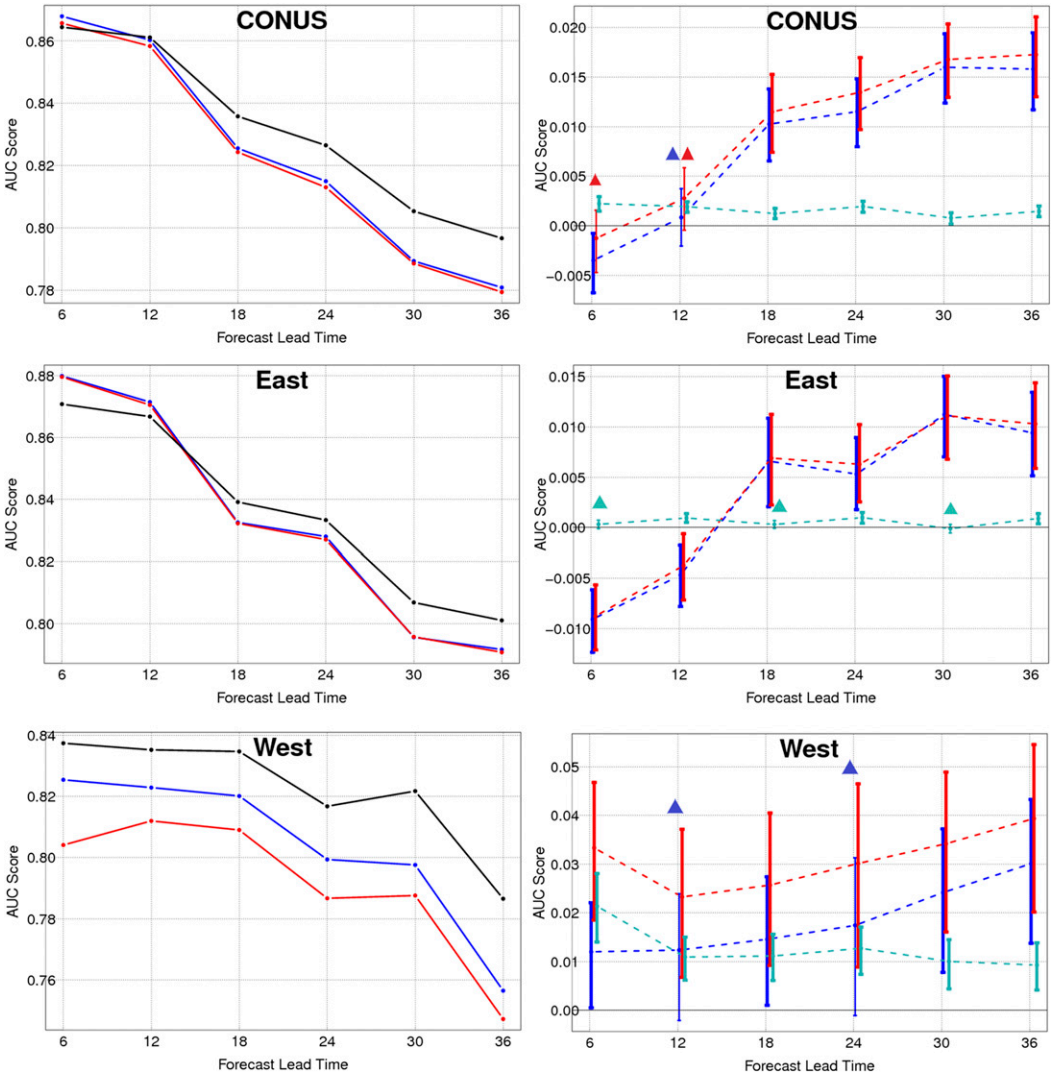


FIG. 10. Graphs of AUC scores as a function of forecast lead time for the Point (black line), Frac (red line), and EAS (blue line) approaches at the 0.5-in. threshold over the (top) CONUS, (middle) East, and (bottom) West verification regions for 3 Feb–30 Sep 2017. Pairwise difference curves for Point − EAS (dashed blue line), Point − Frac (dashed red line), and EAS − Frac (dashed teal line) are also displayed. The 95% bootstrap confidence intervals for the difference curves were obtained using 1000 bootstrapping replications. If the differences are statistically significant, the confidence intervals are depicted in boldface. Triangles are associated with confidence intervals where the differences are not statistically significant.

threshold was similar (Fig. 7); the differences between Frac and EAS were statistically significant at 20, 80, and 100 km, but they were not significant at 40 and 60 km.

Attributes diagrams were constructed for all three regions at the 0.5-in. (Fig. 8) and 1.0-in. (Fig. 9) thresholds. Note that the curves for the Point method have values at each discrete point probability (e.g., 0, 0.125, and 0.25), while the curves for the two postprocessed techniques have values corresponding to each probability bin (e.g., 0–0.1, 0.1–0.2, and 0.2–0.3). For 0.5 in., the Point probability curves were below the perfect reliability line for all regions, which was indicative of overforecasting. The overforecasting bias was most prevalent over the West, where the curve fell below the no-skill line for the 0.125, 0.25, and 0.375 probabilities. The Frac probabilities reduced the overforecasting bias of the Point probabilities. However, the corresponding probability curves were generally above the perfect reliability line, which was indicative of underforecasting. In addition, the number of forecasts plots illustrates that the Frac probabilities had a very low event frequency at higher probability values. Consequently, while Frac often appeared to be the closest to the perfect reliability line for the 0.8–0.9

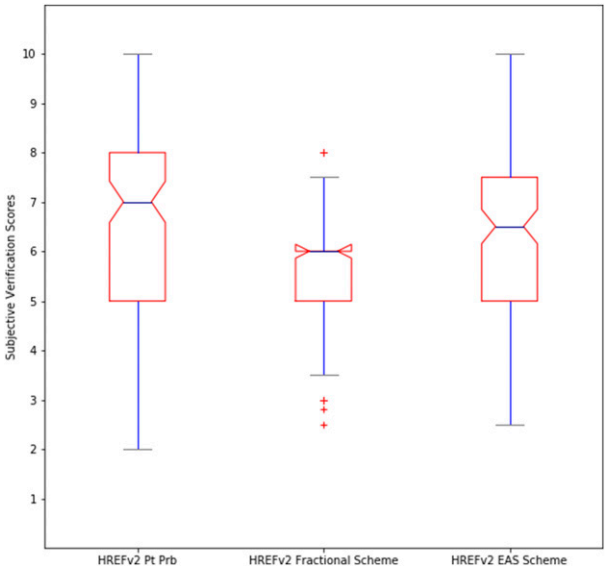


FIG. 11. Box plot of all the subjective verification scores from the 2017 FFaIR experiment for each of the three probability methods for 19 Jun–21 Jul 2017.

and 0.9–1.0 probability bins, the sample size was nearly an order of magnitude smaller than that of the other methods. For instance, there were 146 708, 2878, and 68 913 total forecasts over the West for a 100% Point probability and within the 0.9–1.0 probability bin for Frac and EAS, respectively. The EAS probability curves lied relatively close to the perfect reliability line. Over the CONUS and the East, EAS achieved nearly perfect reliability for all but the highest probabilities. EAS had a slight overforecasting bias, most noticeably at the higher probability values, and at times a slight underforecasting bias, particularly for the lower probability values. Nevertheless, EAS yielded a significant improvement compared to the overforecasting bias of Point, and it reduced the underfore-casting bias of Frac. The superior performance of the EAS method over the West illustrates that it was the most reliable over complex terrain.

For the higher impact precipitation events at the 1.0-in. threshold, the trends were generally the same but magnified (Fig. 9). The overforecasting of Point is clear; in the CONUS and the East, the curve was located below the no-skill line for the 0.125 probability, and in the West, the curve was located below the no-skill line for the 0.125, 0.25, 0.375, and 0.5 probabilities. The underforecasting bias of Frac had also increased; for instance, there were no forecasts over the West that fell into the 0.8–0.9 or 0.9–1.0 probability bins. The EAS method was overall closest to the perfect reliability line. The exception is for the higher probabilities over the CONUS and the East, where Frac was the most reliable at the 0.7–0.8, 0.8–0.9, and 0.9–1.0 probability bins.

Analyses of the AUC score as a function of forecast lead time were constructed for the three regions in order to evaluate the discriminatory ability of each method. Recall that an AUC score greater than 0.7 is indicative of a useful probabilistic forecast, and that for rare-event forecasting applications AUC is sensitive to the height of the top-most point. As with the FBS plots, the pairwise difference curves and the corresponding 95% bootstrap confidence intervals are included. For 0.5 in., Point generally had the highest AUC, except at earlier lead times over the CONUS and the East (Fig. 10). This result is consistent with the Point method preserving the highest amount of detail, making it easier to achieve higher probability values; thus, it had more sharpness and a higher discriminatory ability than the fractional coverage techniques. The Frac and EAS approaches had similar AUC scores, except over the West where EAS consistently outperformed Frac. The confidence intervals illustrate that the differences between Frac and EAS were statistically significant at all lead times over the CONUS and the West. The AUC scores for the 1.0-in. threshold (not shown) were worse than those for 0.5 in., and at longer lead times the scores for Frac and EAS fell below 0.7.

### c. *The 2017 FFaIR experiment*

To complement the objective verification scores presented in the previous section, subjective verification statistics from the 2017 FFaIR experiment were compiled. Figure 11 is a box plot of all the subjective scores from the 2017 FFaIR experiment for each of the three probability methods (Fig. 26 in Perfater and Albright 2017). The standard deviation of the FFaIR scores represents a measure of the variance or spread from the mean score. The Point method had the highest mean score of 6.44 out of 10 with a standard deviation of 1.83. The EAS method had a comparable mean score of 6.31 with a standard deviation of 1.69. The Frac method had the lowest mean score, a 5.42, with a standard deviation of 1.31. The notches on the box plot represent the 95% confidence intervals around the medians, which were 7, 6, and 6.5 for Point, Frac, and EAS, respectively. The confidence interval for Frac did not overlap with the confidence intervals for Point or EAS, indicating the median of Frac was significantly different from the medians of Point and EAS.

The participants also provided valuable feedback on the different probability schemes. Since the probabilities were dependent on the underlying HREFv2 forecast, the comments focused more on how the three different methods visually represented the probabilistic

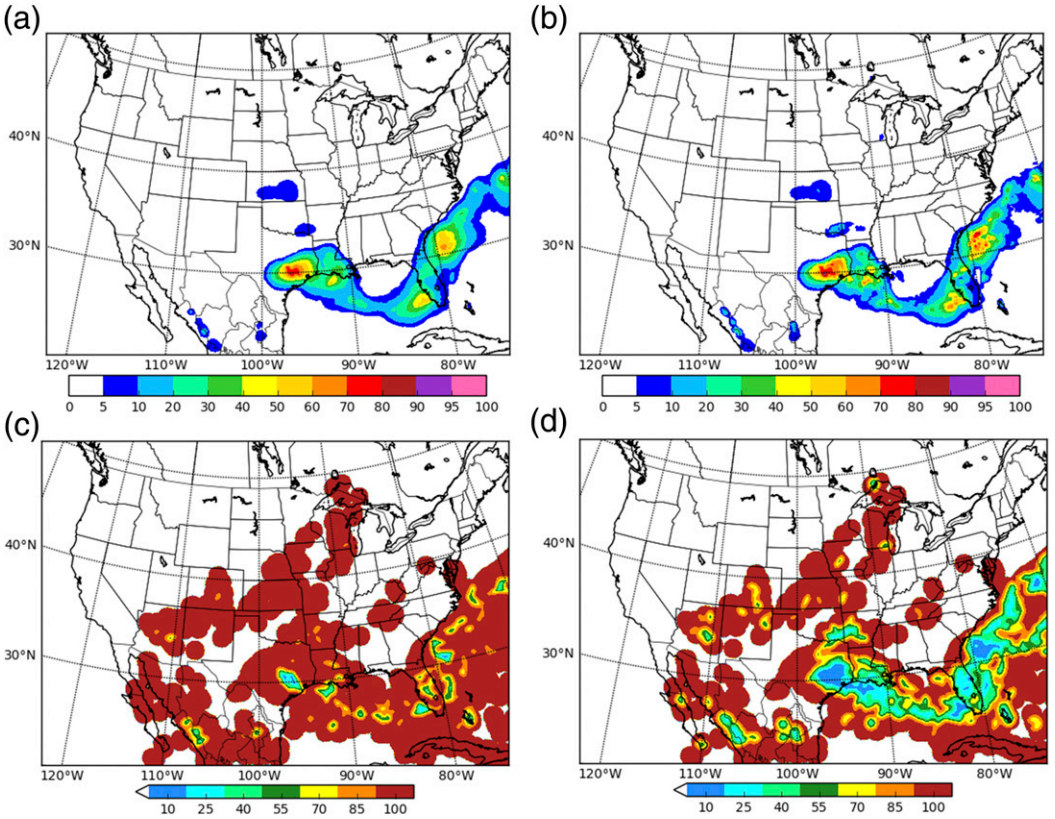


FIG. 12. Plots of probabilities (%) of 0.5 in. of precipitation accumulating over the 6-h period beginning at 1800 UTC 27 Aug 2017 and ending at 0000 UTC 28 Aug 2017 for the (a) EAS and (b) EAS 0.5 methods. These probabilities are 36-h forecasts from the 1200 UTC 26 Aug HREFv2 cycle. (c),(d) Corresponding plots of the radii values (km) utilized by the EAS methods are also displayed.

field and the utility they would each provide to a forecaster. The traditional point probabilities were generally favored due to the large amount of detail they contained and were considered particularly useful for smaller-scale features, such as diurnally driven or sea-breeze-induced convection that often occurred in the Southeast. In these cases, the Point probabilities typically gave some indication that a precipitation event was possible whereas the other two schemes showed little to no indication. One forecaster stated that they "would rather have a certain amount of overforecasting and more detail even if the verification wasn't as great."

The Frac approach was generally the least favored because participants felt that it smoothed out the probability field too much, producing very low probabilities. A related key finding from the 2016 HMT–WPC Winter Weather Experiment was that narrow lake-effect snowbands were often not associated with high probabilities when the Frac method was employed (Perfater and Albright 2016). Conversely, the EAS approach garnered positive feedback from participants as the variable radii allowed for more detail in certain situations that was often appreciated by the forecasters. Both Frac and EAS were preferred by forecasters as a tool to use when drawing deterministic QPF contours because Point was described as too noisy. One suggestion for modifying the EAS method was to reduce the lower bound of the radii below 10 km in order to produce a greater compromise between the extreme detail of Point and the smoother field of Frac. This suggestion was attempted by setting the lower bound to 0 km, which yields a traditional point probability, but the resultant probability field looked too sharp to be reliable with abrupt transitions from low to high values (not shown).

As an alternative solution, the EAS method was run using $\alpha = 0.5$ for (3), hereafter referred to as EAS 0.5. This modification made it easier to achieve the similarity criteria, lowering the radii values and increasing the sharpness of the probability field (Fig. 12). Attributes diagrams from 1 August to 30 September 2017 for the 0.5-in. threshold (Fig. 13) reveal that EAS 0.5 was comparably reliable to EAS. In general, the underdispersion associated with EAS 0.5 was larger than EAS.

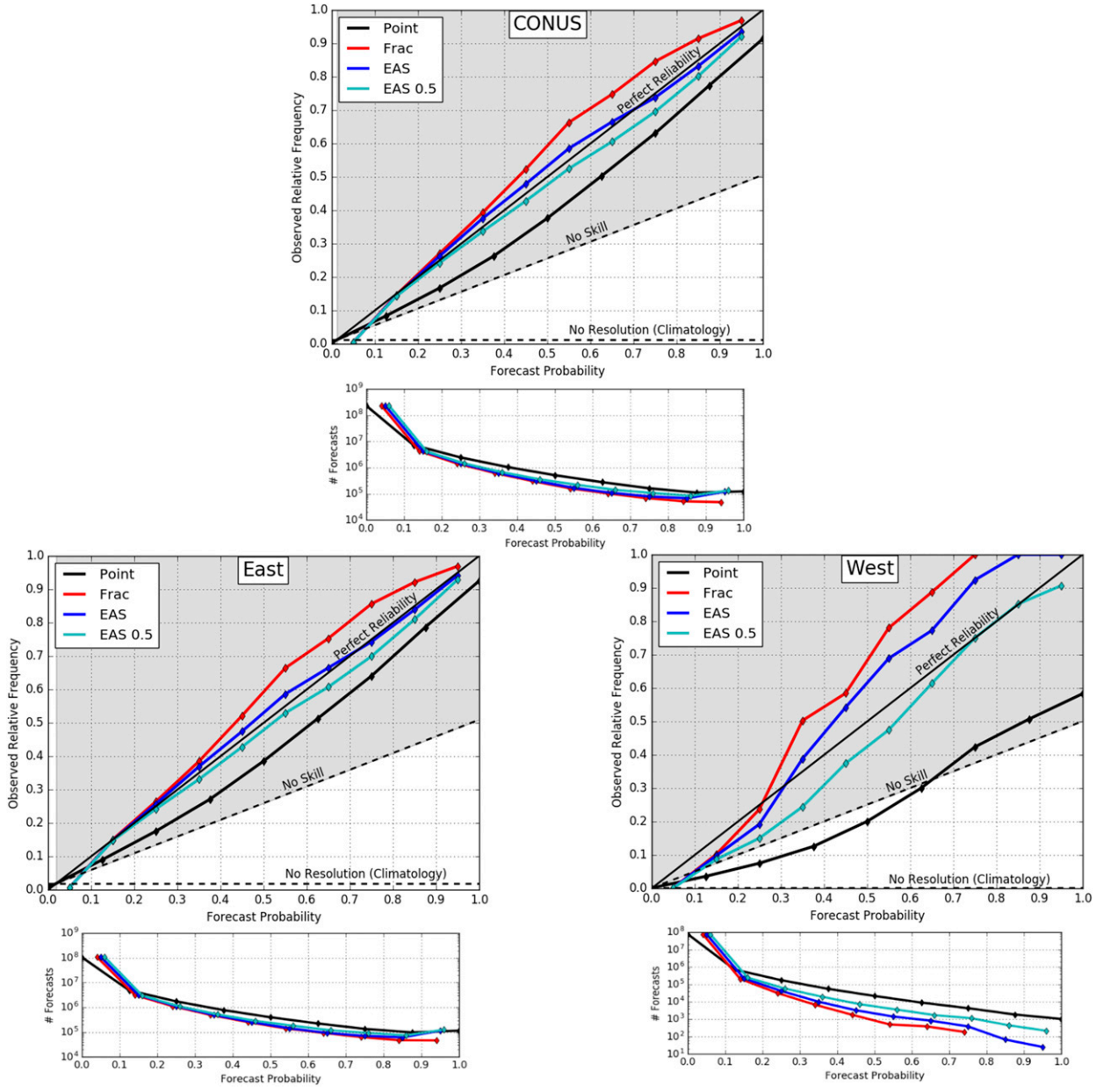


FIG. 13. Attributes diagrams for the Point (black line), Frac (red line), EAS (blue line), and EAS 0.5 (teal line) approaches at the 0.5-in. threshold over the CONUS, East, and West verification regions for 1 Aug–30 Sep 2017. The perfect reliability line (solid black), the no-skill line (dashed black), and the no-resolution line (dashed black) are depicted. The total number of forecasts at each discrete probability value or within each probability bin are also plotted beneath each attributes diagram.

However, EAS 0.5 was arguably the most reliable over the West, the region most dominated by complex terrain. The behavior at the 1.0-in. threshold was similar (not shown). Applying the EAS technique with $\alpha = 0.5$ yielded a substantial improvement over the Point method, which is encouraging since the forecasters had a strong preference for sharp forecasts with a high amount of detail. Based on the subjective verification presented herein, the HMT–WPC staff recommended the EAS probability technique be transitioned to WPC operations (Perfater and Albright 2017).

## 4. Discussion and conclusions

Traditional ensemble point probabilities are very sharp because CPEs tend to be underdispersive and are

of limited size. Member forecasts of small-scale events often do not overlap. For instance, individual model forecasts that do not overlap but are relatively close to one another will produce a 0% point probability over the region of interest (Fig. 5). Fractional coverage methods attempt to address these issues by taking the spatial uncertainty of a forecast into account. The Frac method produces a much smoother probability field, making it more difficult to achieve high probabilities as observed in the reliability diagrams (Figs. 8, 9, and 13). While an inability to predict high probabilities is not always detrimental, probabilistic forecasts using the Frac method are often not sharp enough to make deterministic decisions with high confidence. The underlying assumption with Frac is that the event could occur at any point within a 100-km radius around the point of interest. With locally forced events with little uncertainty, such as orographic precipitation or lake-effect snowbands, this assumption is often physically impossible. A 100-km radius would include points that are not in the mountains or downstream of a lake.

The EAS approach attempts to retain the value from both the Point and Frac approaches in an adaptive way by increasing spread, via inflating the radius for probability calculation, while preserving the relative magnitude of the spread–skill relationship. As demonstrated for a high-impact precipitation event (Fig. 1) and in an idealized sense (Fig. 5), the EAS technique produces a smoother field than the Point method, but a sharper field than the Frac method. Both fractional coverage methods have a lower FBS than the traditional point probabilities over all verification regions, indicating the fractional coverage techniques are associated with smaller forecast errors, and EAS has a lower FBS than Frac over the West at smaller spatial scales (Figs. 6 and 7). Figures 8, 9, and 13 depict how EAS is more reliable than Point or Frac, and especially so over the West, a region dominated by complex terrain. Furthermore, while Point has the highest overall AUC scores, EAS consistently has a higher AUC than Frac over the West (Fig. 10).

The 2017 HMT–WPC FFaIR experiment brought together individuals from a wide variety of backgrounds and institutions across the meteorological field, including operational forecasters. Having a variety of perspectives in the same room fostered great discussions, which directly led to product enhancements to the EAS technique. Figure 11 illustrates that participants scored the Point and EAS methods higher than the Frac method. The traditional point probabilities were often favored because of the large amount of detail they contained, but they were described as too noisy for drawing deterministic QPF contours; the participants therefore suggested increasing the amount of detail present in the EAS probabilities. To accomplish this, the similarity criteria parameter was modified to allow for more finer-scale details in the probability field, making it easier to achieve (4). It remains unclear whether the primary goal of these postprocessing techniques is to make forecasts look good in order to better please the users, or to verify well. As a consequence, future work on probabilistic forecast calibration needs to be cognizant of the fact that some user groups actually prefer forecasts that have less-than-ideal verification statistics.

Based on the eight months of objective verification and the subjective feedback and scores obtained through the 2017 FFaIR experiment, the authors recommend the EAS postprocessing technique for transition to National Weather Service (NWS) operations. The technique is ensemble agnostic, meaning it is applicable to any CPE and not just HREFv2. However, the authors note that different ensembles might work best with different settings whether they are more or less dispersive. While the EAS technique herein has only been applied to 6-h QPF, it can be applied to other variables, including but not limited to accumulated snowfall, precipitation type, composite reflectivity, and updraft helicity. Efforts to accommodate these variables are ongoing.

*Acknowledgments.* This work is part of the Automated High-Resolution Ensemble-Based Hazard Detection Guidance Tool project, which is funded by the U.S. Weather Research Program (USWRP). The project is a collaborative effort between ESRL/GSD, EMC, WPC, and NCAR/DTC. Specific individuals we thank include Tressa Fowler (NCAR/DTC), Tatiana Burek (NCAR/DTC), John Halley Gotway (NCAR/DTC), Tara Jensen (NCAR/DTC), Curtis Alexander (ESRL/GSD), Adam Clark (NSSL), Josh Kastman (WPC/CIRES), Binbin Zhou (IMSG/EMC), Ying Lin (EMC), Logan Dawson (IMSG/EMC), Jun Du (EMC), and three anonymous reviewers.

## REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Alcott, T., I. Jankov, C. Alexander, S. Weygandt, S. Benjamin, J. R. Carley, and B. T. Blake, 2017: Calibrated, probabilistic hazard forecasts from a time-lagged ensemble. *33rd Conf. on Environmental Information Processing Technologies*, Seattle, WA, Amer. Meteor. Soc., 7B.2, https://ams.confex.com/ams/97Annual/webprogram/Paper311242.html.

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.*, **139**, 3887–3905, https://doi.org/10.1175/MWR-D-10-05013.1.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168–189, https://doi.org/10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Clark, A. J., W. A. Gallus Jr., and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

Dey, S. R. A., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

——, N. M. Roberts, R. S. Plant, and S. Migliorini, 2016: A new method for the characterization and verification of local spatial predictability for convective-scale ensembles. *Quart. J. Roy. Meteor. Soc.*, **142**, 1982–1996, https://doi.org/10.1002/qj.2792.

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

Gallus, W. A., Jr., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296–1302, https://doi.org/10.1175/1520-0434(2002)017<1296:IOVGBS>2.0.CO;2.

Gebhardt, C., S. E. Theis, M. Paulat, and Z. B. Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, https://doi.org/10.1016/j.atmosres.2010.12.008.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741, https://doi.org/10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2.

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Harless, A. R., S. J. Weiss, R. S. Schneider, M. Xue, and F. Kong, 2010: A report and feature-based verification study of the CAPS 2008 storm-scale ensemble forecasts for severe convective weather. *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., 13.2, https://ams.confex.com/ams/pdfpapers/175883.pdf.

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.

Hohenegger, C., and C. Schär, 2007: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467–4478, https://doi.org/10.1175/2007JAS2143.1.

——, A. Walser, W. Langhans, and C. Schär, 2008: Cloud-resolving ensemble simulations of the August 2005 Alpine flood. *Quart. J. Roy. Meteor. Soc.*, **134**, 889–904, https://doi.org/10.1002/qj.252.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Jensen, T. L., J. H. Gotway, R. Bullock, T. L. Fowler, B. Brown, B. Strong, L. Nance, and Y. H. Kuo, 2017: Recent advancements in verification within the Developmental Testbed Center. *28th Conf. on Weather Analysis and Forecasting/24th Conf. on Numerical Wea. Prediction*, Seattle, WA, Amer. Meteor. Soc., 586, https://ams.confex.com/ams/97Annual/webprogram/Paper312489.html.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., P9.137, https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html.

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

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

Lean, H. W., A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136**, 3408–3424, https://doi.org/10.1175/2008MWR2332.1.

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, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, https://doi.org/10.1175/825.1.

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, https://doi.org/10.1175/JAS-D-11-0315.1.

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

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084, https://doi.org/10.1175/2008WAF2222142.1.

Peralta, C., Z. B. Bouallègue, S. E. Theis, C. Gebhardt, and M. Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *J. Geophys. Res.*, **117**, D07108, https://doi.org/10.1029/2011JD016581.

Perfater, S., and B. Albright, 2016: The 2016 HMT-WPC Winter Weather Experiment. Weather Prediction Center Final Rep., 39 pp., http://www.wpc.ncep.noaa.gov/hmt/WWE_2016_final_report.pdf.

——, and ——, 2017: 2017 Flash Flood and Intense Rainfall Experiment. Weather Prediction Center Rep., 94 pp., http://www.wpc.ncep.noaa.gov/hmt/2017_FFaIR_final_report.pdf.

Radhakrishna, B., I. Zawadzki, and F. Fabry, 2012: Predictability of precipitation from continental radar images. Part V: Growth and decay. *J. Atmos. Sci.*, **69**, 3336–3349, https://doi.org/10.1175/JAS-D-12-029.1.

Roberts, N. M., 2005: An investigation of the ability of a storm scale configuration of the MET Office NWP model to predict flood-producing rainfall. Met Office Tech. Rep. 455, 80 pp.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale Modeling System: Recent changes and future plans. *23rd Conf. on Weathrt Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 2A.4, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154114.htm.

——, and Coauthors, 2017: Mesoscale modeling development at the National Centers for Environmental Prediction: Version 4 of the NAM forecast system and scenarios for the evolution to a high-resolution ensemble forecast system. *28th Conf. on Weather Analysis and Forecasting/24th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 3B.4, https://ams.confex.com/ams/97Annual/webprogram/Paper311212.html.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Saito, K., and Coauthors, 2006: The operational JMA nonhydrostatic mesoscale model. *Mon. Wea. Rev.*, **134**, 1266–1298, https://doi.org/10.1175/MWR3120.1.

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

Schwartz, C. S., and R. A. Sobash, 2017: Applications of neighborhood verification approaches to convection-allowing ensembles: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, https://doi.org/10.1175/2009MWR2924.1.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Mon. Wea. Rev.*, **139**, 976–991, https://doi.org/10.1175/2010MWR3425.1.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 175 pp.

Smith, T. L., S. G. Benjamin, J. M. Brown, S. Weygandt, T. Smirnova, and B. Schwartz, 2008: Convection forecasts from the hourly updated, 3-km High Resolution Rapid Refresh (HRRR) model. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 11.1, https://ams.confex.com/ams/24SLS/techprogram/paper_142055.htm.

Tang, Y., H. Lean, and J. Bornemann, 2013: The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteor. Appl.*, **20**, 417–426, https://doi.org/10.1002/met.1300.

Tennant, W., 2015: Improving initial condition perturbations for MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **141**, 2324–2336, https://doi.org/10.1002/qj.2524.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, https://doi.org/10.1017/S1350482705001763.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423, https://doi.org/10.1175/2010MWR3487.1.

Weisman, M. L., C. A. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, https://doi.org/10.1175/WAF-D-13-00135.1.