A Comparison of Two Methods for Bias Correcting Precipitation Skill Scores

MATTHEW E. PYLE

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland

KEITH F. BRILL

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

(Manuscript received 28 June 2018, in final form 19 November 2018)

ABSTRACT

A fair comparison of quantitative precipitation forecast (QPF) products from multiple forecast sources using performance metrics based on a 2×2 contingency table with assessment of statistical significance of differences requires accounting for differing frequency biases to which the performance metrics are sensitive. A simple approach to address differing frequency biases modifies the 2×2 contingency table values using a mathematical assumption that determines the change in hit rate when the frequency bias is adjusted to unity. Another approach uses quantile mapping to remove the frequency bias of the QPFs by matching the frequency distribution of each QPF to the frequency distribution of the verifying analysis or points. If these two methods consistently yield the same result for assessing the statistical significance of differences between two QPF forecast sources when accounting for bias differences, then verification software can apply the simpler approach and existing 2×2 contingency tables can be used for statistical significance computations without recovering the original QPF and verifying data required for the bias removal approach. However, this study provides evidence for continued application and wider adoption of the bias removal approach.

1. Introduction

The determination of which quantitative precipitation forecast (QPF) source is best in a statistical sense often is more complicated than it first appears. This is because traditional performance metrics respond both to placement error and to frequency bias. Baldwin and Kain (2006) used a geometrical model to demonstrate some aspects of the interaction of placement error and frequency bias for several performance metrics. Brill (2009) showed that any performance metric computed from the 2×2 contingency table for dichotomous forecasts is sensitive to frequency bias. The Brill (2009) critical performance ratio (CPR) can be computed for any metric and gives the hit rate for added or removed forecasts above or below which the metric shows improvement for an incremental change in frequency bias. Even forecasts having a perfect frequency bias can receive an improved performance score by inflating or deflating the frequency bias. Thus, the comparison of different QPF sources cannot be fairly carried out unless those forecasts have very nearly the same frequency biases at all thresholds considered. The goal of this study is to provide supporting evidence for a standard approach to the problem of comparing QPF sources that do not have similar frequency biases.

The traditional metrics used to evaluate gridded QPFs are based on 2×2 contingency tables of binary outcomes assigned values of 1 (yes) or 0 (no) for precipitation accumulations exceeding or not exceeding given thresholds, respectively, for both forecast and observed precipitation. The binary values are accumulated over space and/or time to populate the cells of the contingency table, an example of which is given by Table 1. The contingency table values may be normalized by dividing each by the total count of pairs of forecast and analyzed values. Performance metrics computed using the 2×2 contingency table values have quantifiable sensitivity to frequency bias (Brill 2009). Frequency bias is the ratio of the frequency of ''yes'' forecasts to the frequency of ''yes'' observations [$(a + b)/(a + c)$, in terms of the contingency table values in Table 1]. In this study, ''frequency bias'' and ''bias'' without the qualifier will have the same meaning.

Corresponding author: Matthew E. Pyle, matthew.pyle@noaa.gov

DOI: 10.1175/WAF-D-18-0109.1

TABLE 1. Example 2 × 2 contingency table of accumulated counts or frequencies (a, b, c, d) for observed and forecast precipitation exceeding or not exceeding some threshold Q.

Outcomes	Observation $\geq Q$ (yes)	Observation $< Q$ (no)
Forecast $\geq Q$ (yes)	a	b
Forecast $< Q$ (no)	c	d

Popular metrics for deterministic QPFs, such as the critical success index (a.k.a. threat score) and Gilbert skill score [GSS; Schaefer (1990), a.k.a. equitable threat score (ETS)], tend to ''reward'' a slightly higher frequency-biased forecast with a higher skill score (Mason 1989; Hamill 1999). But, this reward occurs only if a sufficient fraction of the forecast values enhanced by excessive bias become hits (Brill 2009), where the hit count or fraction is the value of a in Table 1. Thus, forecasting such that the frequency bias exceeds unity may be an advantage since overforecasting that adds enough yes forecasts corresponding to yes observations can increase the value of the metric. This boost from increased bias often is realized for forecasts that place precipitation areas in close proximity to areas of observed precipitation.

Various efforts have been made to provide a level or fair comparison among forecasts with disparate bias characteristics. The contour relabeling approach used by Hamill (1999) has forecast source selection dependency, is not widely used, and is not evaluated here.[1] Mesinger (2008) derived a mathematically based adjustment of the hit count as the forecast count is raised or lowered to achieve unit bias as described in detail below. The quantile mapping approach of (Clark et al. 2009, p. 1134) removes the bias by matching the frequency distribution of each forecast to that of the verifying analysis or point data and is described in detail below. The latter two methods are compared in this study. Eliminating or minimizing the effect of frequency bias is essential to assessing the statistical significance of differences in 2 × 2 contingency table metrics between two forecasts (Hamill 1999). A recent study (Gowan et al. 2018) implicitly removes bias by evaluating forecasts within the context of the model and observation climatologies at a point (e.g., a 95th percentile event), but given the historical utilization of fixed thresholds in operational QPF verification, this implicit approach was not explored in this paper.

The objective of this study is to compare the bias-adjustment (BA) technique of Mesinger (2008) and the bias-removal (BR) technique of Clark et al. (2009, p. 1134). An advantage of the BA technique is that it can be applied to existing contingency tables; the BR technique must be applied when the QPF is verified, as it requires that the hit count after bias removal be saved in addition to the elements shown in Table 1.[2] Of interest to the National Weather Service (NWS) Weather Prediction Center (WPC), the NWS Environmental Modeling Center (EMC), and others is whether the BA and BR techniques yield the same result for the statistical significance of differences in GSS for pairwise comparisons of various QPFs. If the BA method reliably gives the same result as the BR method, then historical contingency table data can be assessed without having to access and modify the original QPFs using verifying analyses, sidestepping the need to apply the BR approach retrospectively. In some cases, perhaps only the contingency table data still exist, and the BR method cannot be applied.

The BA and BR methods cannot be expected to agree precisely on the value of performance metrics. Even disagreements between the two methods as to which of two QPF sources has the better performance can be tolerated so long as the differences in the metric are not statistically significant. Statistical significance is assessed using the resampling method of Hamill (1999) with a test level of 0.05. The proposition is that both the BA and BR methods always agree on which QPF source in a pairwise comparison is better when the difference in the metric is statistically significant for *both* methods. Logically, one counterexample is sufficient to refute the proposition that the BA and BR methods lead to the same result regarding statistically significant differences in metrics. Thus, the purpose of this effort is to seek out situations for which the BA and BR approaches provide contradictory results, both of which are statistically significant.

The WPC has an archive of 2 × 2 contingency table verification data for several sources of QPFs. The hit rate determined by the BR method has been included in this archive for a period of several years. This study makes use of these data to directly compare the BA and BR methods and judge whether or not the two methods can be utilized interchangeably.

[1] The focus of Hamill (1999) is hypothesis testing by resampling, not frequency bias elimination.

[2] After bias removal, the forecast frequency equals the observed frequency; thus, the modified forecast count does not have to be saved unless it is needed as a check on the BR process.

2. Frequency bias treatment methods

The BA and BR methods are very different in how they function to eliminate the effect of bias when comparing performance metrics for differently biased forecasts. To make this point clear, it is necessary to describe each method in some detail.

a. The BA approach

The BA approach of Mesinger (2008) attempts to place the forecasts being compared on an equivalent basis by adjusting all forecasts to unit frequency bias. The Mesinger (2008) version of BA provides a more mathematically sound refinement to the earlier approach of Mesinger and Brill (2004), and is "based on the assumption that the increase in hits per unit increase in false alarms is proportional to the yet unhit area" (Mesinger 2008). This BA approach attempts to impart to scores like the GSS a stronger measurement of "placement accuracy" via the modified hit rate and unit frequency bias.

The quoted assumption in the preceding paragraph is expressed by Mesinger (2008) as a differential equation rendered in terms of forecast, hit (correct forecasts), and observed frequencies, $F = a + b$, $H = a$, and $O = a + c$, respectively, where a, b, and c are frequencies from Table 1. The differential equation is

$$\frac{dH}{dA} = \beta(O - H), \tag{1}$$

where $dA = dF - dH$, and β is a constant determined from the known values of F, H, and O. The solution for the adjusted hit rate [Mesinger (2008); see his Eq. (11)] is

$$H_a = O - \frac{F - H}{\ln\left(\dfrac{O}{O - H}\right)} \text{lambertw}\left[\frac{O}{F - H} \ln\left(\frac{O}{O - H}\right)\right], \tag{2}$$

where H_a is the bias-adjusted hit rate and the "lambertw" function returns the value of w given z in the equation $z = we^w$. In (2), z is the argument of the lambertw function, which can be solved quickly by an iterative method (e.g., file 443 under the Collected Algorithms catalog online at http://www.netlib.org/toms).

In (2) the BA hit rate is a function of F, H, and O and, consequently, the elements a, b, and c of the original 2×2 contingency table. To compute a BA performance metric, H is replaced by H_a and F is replaced by O (giving unit bias), so that the modified contingency table of frequencies has $a' = H_a$, $b' = O - H_a$, $c' = O - H_a$, and $d' = 1 - 2O + H_a$.

To investigate how the BA modifies the GSS as a function of bias, (2) is rewritten in terms of probability of detection (POD) $P = H/O$, and bias $B = F/O$, yielding

$$H_a = O - \frac{O(B - P)}{\ln\left(\dfrac{1}{1 - P}\right)} \text{lambertw}\left[\frac{1}{B - P} \ln\left(\frac{1}{1 - P}\right)\right]. \tag{3}$$

If the event frequency O and POD P are held constant in (3), the effect of changing bias is isolated because H_a becomes a function only of B. But, since O is constant for a specified event frequency, analyzing the bias response with POD constant amounts to only changing F in (2).

Using (3), differences between the BA GSS and GSS can be computed for given event frequencies and probabilities of detection and plotted as a function of bias, as shown in Fig. 1. The requirement that F is greater than or equal to H sets a lower limit on the possible range of biases to evaluate, a lower limit that depends on the P value being considered. The differences between BA GSS and GSS become very large near this lower limit of bias for each P value (Fig. 1). As bias decreases below the value of 1, the differences between the BA GSS and GSS rapidly increase as bias approaches the POD limit. Common events ($O = 0.30$) having a low POD exhibit the least abrupt increase as bias decreases below unity. Above unity, rare events ($O = 0.002$; red curves in Fig. 1) have BA GSS values lower than the GSS values (negative differences), whereas common events have higher BA GSS values with an upward trend as bias increases. If forecast sources tend to be underbiased for rare events and overbiased for common events, Fig. 1 shows that the BA GSS will generally be larger than the GSS value, except in the case of overbiased common events with high POD (solid red curve). It is also possible that events more frequent than 0.002 depicted in Fig. 1 and predicted with greater than unit bias will have lower values for the BA GSS. But the overall impression from the shape of these curves is that a very low bias gets a much stronger inflation of GSS than a high bias gets a reduction of GSS (particularly as common high-biased events also show a modest inflation of GSS).

b. The BR approach

The BR approach of Clark et al. (2009, p. 1134) differs significantly from the BA approach, as it matches the forecast precipitation distribution to the observed distribution using a technique conceptually similar to the computation of a probability-matched mean (Ebert 2001)

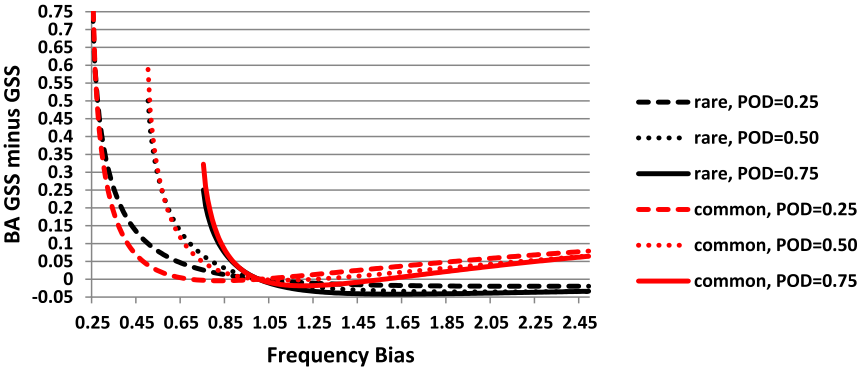

FIG. 1. BA GSS minus GSS difference as a function of frequency bias for two event frequencies O [$(a + c)/(a + b + c + d)$ in Table 1]—common ($O = 0.30$; red curves) and rare ($O = 0.002$; black curves)—and three different probabilities of detection [$(a/(a + c)$ in Table 1]—0.25 (long dash), 0.50 (short dash), and 0.75 (solid).

and identical to the method of transformation of frequency distributions described by Panofsky and Brier (1968, 40–42), achieving unit bias without an assumed mathematical response among the contingency table (Table 1) counts. In contrast to the BA method, the BR approach retains the actual QPF placement information rather than inferring it from contingency table values. The BR approach replaces forecast values with matching elements from the distribution of analysis values; due to this manipulation of the original gridded forecast by the verifying analysis, it must be performed while this gridded information remains available. As discussed in section 1, this requirement limits the applicability for historical verification datasets for which only contingency counts have been retained.

To demonstrate how the BR method works, synthetic (fictitious) forecast and observed data are created as shown in Table 2 for a spatial domain represented by 10 points. The quantile mapping proceeds in left-to-right order of the columns in Table 2. First, the forecast and observed values (first and second columns) are sorted in ascending order, to form the forecast and observed order statistics (third and fourth columns). Then, the quantile mapping is done as follows: for each forecast value, find its position (quantile) in the forecast order statistics, then locate the observed value that has the same position in the observed order statistics and replace the forecast value with that observed value. For example, consider the first forecast value in Table 2: 0.25. The value 0.25 is the 40th percentile (0.4 quantile) in the sorted forecast order statistics. The 40th percentile in the sorted observed order statistics is 0.43. Thus, the first forecast value is replaced by 0.43, while the observed value at the spatially located point is 0.48. This replacement operation continues through all of the forecast values, ending with the final and largest forecast

TABLE 2. Synthetic quantile mapping demonstration data. The first and second columns give the forecast and observed data at their original spatial locations. The third column contains the forecast data sorted from low to high, and the fourth column contains similarly sorted observed data. The fifth column shows quantile mapped forecast data back in proper spatial position (see section 2 for details on this process). Finally, the last column repeats the second column to show the modified forecast data in proper association with the observed data.

Forecast	Observed	Sorted forecast	Sorted observed	Quantile mapped forecast	Observed
0.25	0.48	0.09	0.09	0.43	0.48
0.11	0.09	0.11	0.17	0.17	0.09
1.02	1.85	0.15	0.22	1.41	1.85
0.09	0.22	0.25	0.43	0.09	0.22
0.77	0.62	0.33	0.48	0.84	0.62
0.95	1.12	0.62	0.62	1.12	1.12
0.33	0.43	0.77	0.84	0.48	0.43
0.15	0.17	0.95	1.12	0.22	0.17
0.62	0.84	1.02	1.41	0.62	0.84
1.32	1.41	1.32	1.85	1.85	1.41

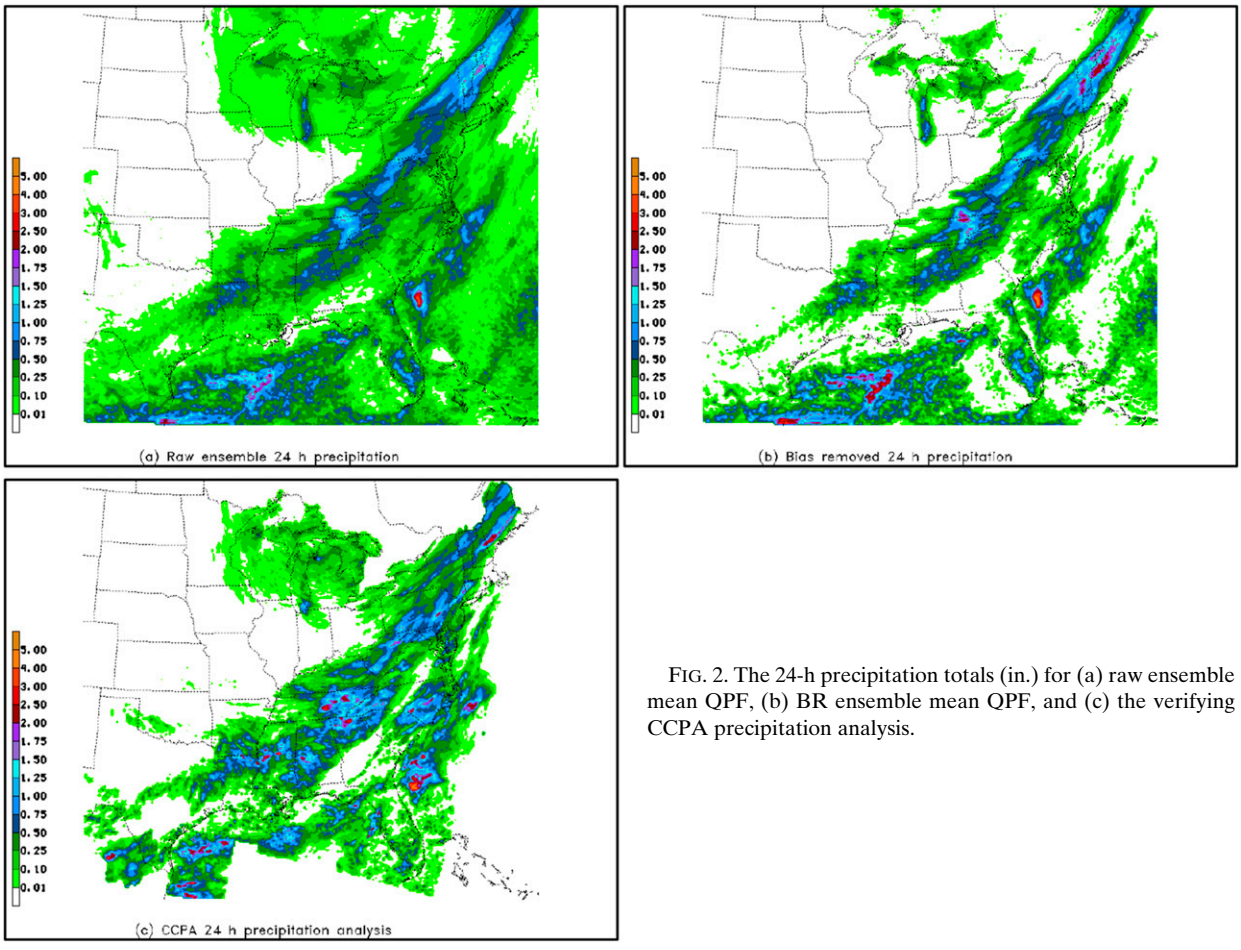

FIG. 2. The 24-h precipitation totals (in.) for (a) raw ensemble mean QPF, (b) BR ensemble mean QPF, and (c) the verifying CCPA precipitation analysis.

value, 1.32, being replaced by the largest observed value, 1.85. In the synthetic data in Table 2, the forecasts have a low bias, and thus the forecast values tend to be replaced by larger values from the observed order statistics. The BR method operates independently of the thresholds ultimately used to construct 2×2 contingency tables. The BR method does not alter the placement of heavy or light forecast precipitation; it only modifies the amounts.

Given the substitution nature of the BR approach, it cannot be analyzed using an equation like the BA approach, but a graphical demonstration of how the BR approach operates on a real gridded data case is illustrative and is provided in Fig. 2. The demonstration case here uses the mean QPF output from a high-resolution ensemble system as the forecast on which to apply bias removal. The smoothing that results from averaging members together to produce an ensemble mean QPF typically leads to a (spatial) overprediction of light precipitation, and an underprediction of heavier amounts.

Comparing the raw 24-h ensemble mean (Fig. 2a) with the BR 24-h mean (Fig. 2b) shows a notable reduction in coverage of measurable (greater than or equal to 0.01 in.) rainfall. The increased coverage of heavier amounts in the BR QPF is evident along the Kentucky–Tennessee border, over northern New England, and over the Gulf of Mexico, and better reflects the high-end intensity of the precipitation analysis being matched (Fig. 2c). Even though the BR approach replaces the true QPF values with corresponding values from the analysis, the orientations and spatial distribution (placement) of the raw model precipitation forecast features are unchanged by the BR process. How the BR approach modifies the raw precipitation forecast depends on the bias properties of the forecast, but the intensity frequency distribution of the modified QPF matches that of the observations, while placement errors remain intact. This neutral matching of forecast intensities to verifying analyses without any additional assumptions is a desirable property of the BR approach.

3. Description of data

For the purposes of this study, almost any archive of QPFs from several different sources along with the verifying analyses could be used. This is not a study designed to argue that one particular forecast is better than another. The focus here is on evaluating two frequency bias treatment methods in making fair assessments of statistically significant differences. Since this work was a collaboration of two national centers, WPC and EMC, it was convenient to use the WPC database in which the BR method has been utilized for several years.

The verifying precipitation analyses are from the climatology-calibrated precipitation analysis (CCPA; Hou et al. 2014). The 4-km resolution analyses are remapped onto the 20-km-resolution WPC QPF grid domain covering the contiguous United States (CONUS). The remapping method preserves area averages. The verification is done at 20-km resolution.

The 6-h QPF verification database utilized in this study is maintained by WPC and contains summary contingency table statistics for QPFs from a variety of operational modeling systems, a WPC-produced multimodel mean, and the WPC operational human 6-h QPFs [described by Novak et al. (2014)]. This database contains the additional contingency table value (the hit count after the BR technique has been applied) from January 2014 onward, enabling comparisons between BR and BA over this relatively recent period.

The European Centre for Medium-Range Weather Forecasts (ECMWF) verification data utilized for this study were generated from a $1^\circ \times 1^\circ$ output grid. This spatially degraded ECMWF data have a lower precipitation bias for heavier precipitation than exists for a higher-resolution ECMWF product also available from the WPC verification database. However, for the purposes of this study, the very low bias of the lower-resolution ECMWF product demonstrates some of the BA behavior explored in section 2 and thus is utilized here. Verification data for a multimodel weighted ensemble mean (WEM) QPF were also used. The WEM process combines deterministic model QPFs with the QPFs from an ensemble prediction system. Where forecast uncertainty is large, the WEM QPF is weighted more toward the mean of all the QPFs. Where uncertainty is small, the WEM QPF is weighted more toward the mean of the deterministic model QPFs. The uncertainty is measured in terms of the coefficient of variation over all the QPFs at each grid point. Further details on this product may be found in Novak et al. (2014, appendix A). The WEM QPF is generated within WPC to provide guidance for WPC QPF forecast products.

This analysis of QPF performance assessment is based on pairwise comparisons of two QPF sources utilizing biased (uncorrected) GSS and the two bias-corrected (BA and BR) GSS metrics, focusing on 6-month periods (an April–September warm season and an October–March cool season) over western and eastern CONUS verification regions[3] (Fig. 3). The focus on different geographical areas and different seasons affords a greater possibility for the emergence of contradictory indications by the BA versus the BR method for statistically significant differences. Hypothesis testing of the statistical significance of the difference in skill scores between pairs of forecasts uses the random resampling technique described by Hamill (1999), with 2000 samples testing the hypothesis that the two skill scores are effectively the same at the 95% confidence level (0.05 test level). Frequency bias also was computed for each system in these comparisons, as it provides a framework for understanding the bias-corrected GSS behavior described in the next section.

4. Analysis results

Comparing the WPC human-generated forecast product (guided in large part from blending solutions from multiple modeling systems) against a well-respected deterministic modeling system, such as the ECMWF, or against a carefully constructed blend of multiple sources of QPFs, such as in the WEM product, addresses an increasingly common question: does the human forecaster add value over and above model guidance (e.g., Novak et al. 2014)? But as these forecasts typically have different bias characteristics, this question is not straightforward to answer without applying a robust method of removing the influence of the frequency bias.

For the six-month 2015/16 cool season and limiting the verification to the western United States, a comparison between WPC human-generated forecasts and ECMWF (truncated to ECM in the plot) deterministic forecasts (Fig. 4) shows that identifying the better of the two forecasts varies with the intensity threshold and depends on which one of biased GSS, BA GSS, or BR GSS is considered. At the 0.01-in. threshold, both the WPC and ECMWF forecasts have a high frequency bias. In both the ''biased'' (uncorrected) GSS and BA GSSs, neither system is significantly better than the other, as their

[3] Note that the BR technique is applied separately over each subregion to ensure that the BR bias is unity when multiple subregions are combined, such as is done here with the aggregation into western and eastern CONUS regions.

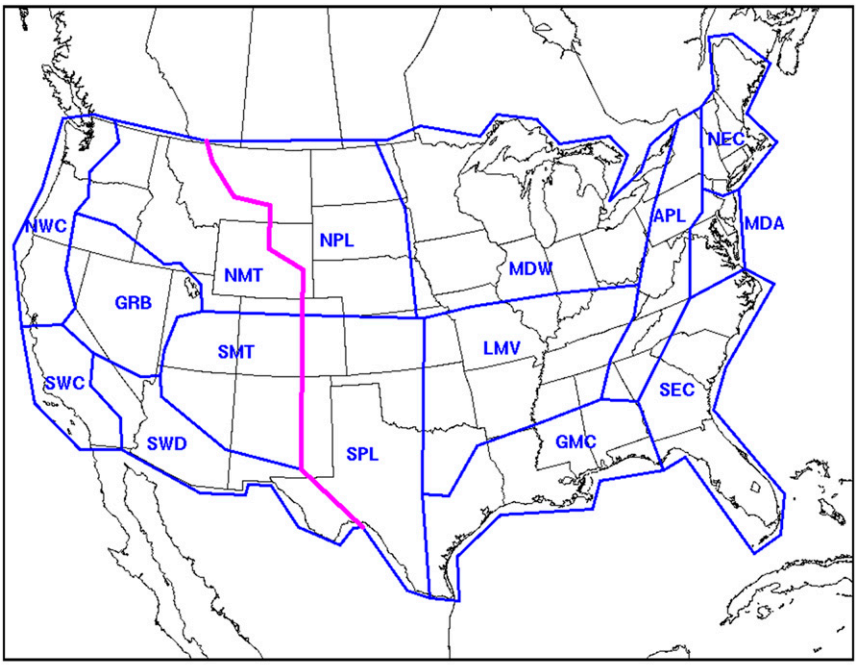

FIG. 3. Map showing the subregions utilized in the WPC QPF database, with the heavy magenta line separating the subregions that are combined to form the western CONUS and eastern CONUS verification regions.

performance is comparable by these two metrics. From the BR GSS perspective the ECMWF forecast is significantly better, indicating superior placement of precipitation by the ECMWF, but with a higher bias. This result evokes an important point: BR favors an overbiased forecast if the forecasts removed are not hits, in other words, if placement is good. Therefore, it remains imperative to always show the frequency bias when reporting performance metrics computed from the 2×2 contingency table, even if a bias-removal treatment has been applied. At the 0.1-in. threshold, the biased GSS and BA GSSs provide a similar sense of relative skill (ECMWF significantly better than WPC), while the BR GSSs show very little difference in skill between the two (equivalent placement accuracy). The frequency biases of the two forecasts are fairly similar at the 0.1-in. threshold. The 0.25–1-in. thresholds show comparable patterns of behavior: the biased and BR GSSs both indicate that the WPC forecast is significantly better (not quite significant for biased GSS at 0.25 in.), while the BA GSS increases the skill score of the low-biased ECMWF forecast by a larger amount than does the BR GSS, leading to the ECMWF forecast being significantly better from the BA GSS perspective. The large increase in the BA GSS at the 1-in. threshold relative to the biased GSS appears to be a real-world example of the strong inflation of GSS that the BA approach can provide for rare events at sufficiently low frequency bias (Fig. 1). The sharp disagreement between the BA and BR approaches at this threshold provides clear evidence that the BA approach cannot be relied upon to replicate the BR result.

Comparing WPC forecasts against the WEM blended forecast product for the 2015/16 cool season over the eastern United States (Fig. 5) shows sample responses for forecasts that have somewhat different frequency bias characteristics than seen in Fig. 4 and provides multiple examples of high-biased forecasts. At the 0.01- and 0.1-in. thresholds, WEM is significantly better than WPC by all three metrics shown. The BA and BR GSSs move the scores in opposite directions relative to the biased GSS, though. For high-biased light precipitation, the BA tends to reduce the score by eliminating hits, while the BR often boosts the score despite the potential for the removal of hits. The BR and BA responses to bias reduction are examined in terms of the Brill (2009) CPR for the GSS in the following discussion of the 0.25-in. threshold in Fig. 5.

At the 0.25-in. threshold, the biased GSS and BR GSSs both show the WEM being significantly better than the WPC forecast, whereas the BA reduces the

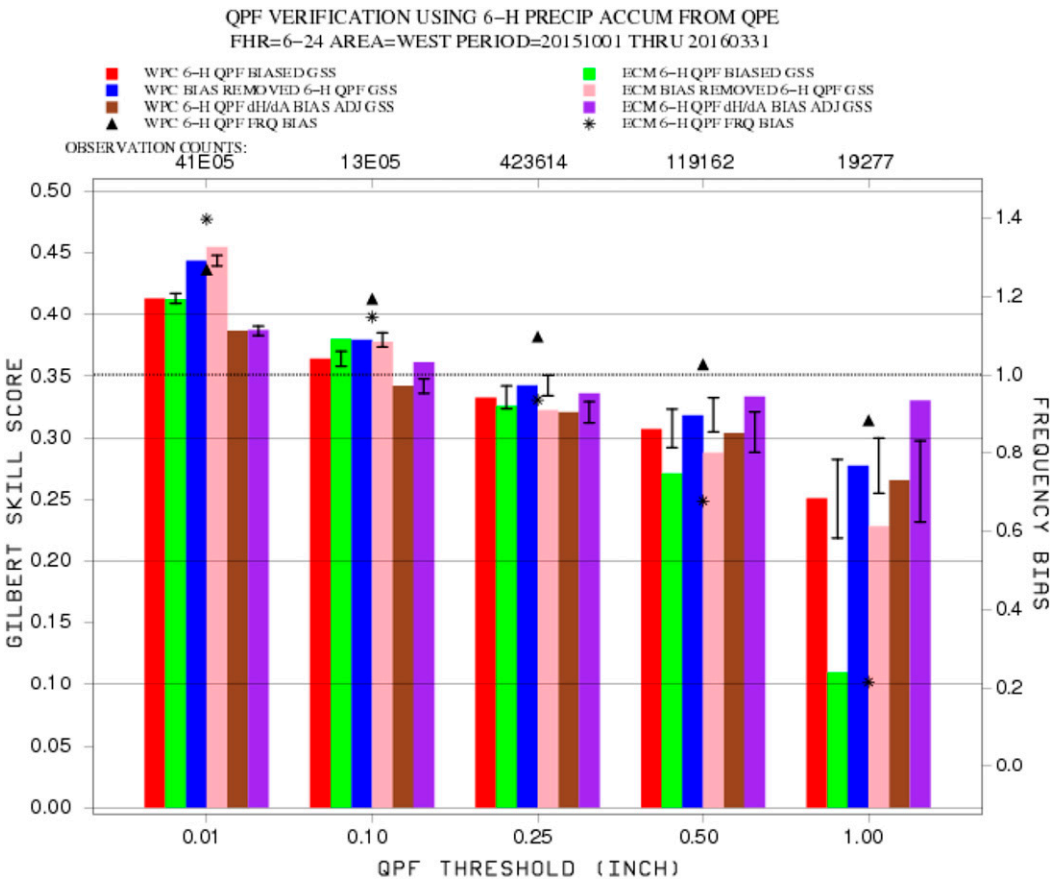

FIG. 4. Pairwise comparisons of WPC and ECMWF (ECM in figure) 6-h QPF forecasts for the 6–24-h forecast range over the 2015/16 cool season and verified over the western CONUS. For each QPF threshold along the abscissa, there are three pairs of comparisons of WPC vs ECM GSS skill shown as color bars plotted against the left-hand axis: biased, BR, and BA. In the right-hand member of each pair, the 95% confidence interval (CI) for the pair is shown as a vertical barred line segment. If the GSS color bar of the right-hand member completely contains the CI segment, it is significantly better than the left-hand member; if the GSS color bar of the right-hand member is completely below the CI segment, it is significantly worse. The frequency biases for the two forecasts are plotted with symbols centered on each QPF threshold value, and are plotted along the right-hand axis. Though CIs are not shown for bias, the frequency biases of the two forecasts differ significantly for all thresholds.

GSS for the more highly biased WEM forecast, making it significantly worse than the WPC forecast. This is a case of opposite results for statistically significant differences and merits closer inspection. A useful tool for this inspection is the Brill (2009) CPR. For reduction in bias to unity, the GSS CPR is an upper bound on the fraction of removed forecasts that are hits for the GSS to show improvement. If the hit fraction for the removed forecasts exceeds the CPR, the score will be degraded. The CPR for the GSS is given on the "ETS" row of Table 2 in Brill (2009). The following equation derived from Brill (2009) expresses the GSS CPR in terms of the fractions F, H, and O:

$$\text{GSS CPR} = \frac{H + O^2 - 2OH}{F + O - 2OF}. \tag{4}$$

Table 3 shows the complete CPR analysis for the 0.25-in. threshold for both the WPC and WEM forecasts compared in Fig. 5. For the WPC forecasts, the CPR value indicates that if more than about 31% of the removed forecasts are hits, then the GSS will decrease. In applying the BR method, only a little more than 3.5% ($\Delta H/\Delta F = 0.03567$) of the removed forecasts are hits; so, the GSS shows an increase from 0.3871 to the BR GSS value of 0.4219. For the BA method, more than 44% of the removed forecasts are hits, exceeding the 31% limit and causing the GSS to decrease from 0.3871 to the BA GSS value of 0.3708. For the WEM forecasts, the CPR value is only slightly higher compared to the WPC, indicating that if more than about 31.5% of the removed forecasts are hits, then the GSS will decrease. But only about

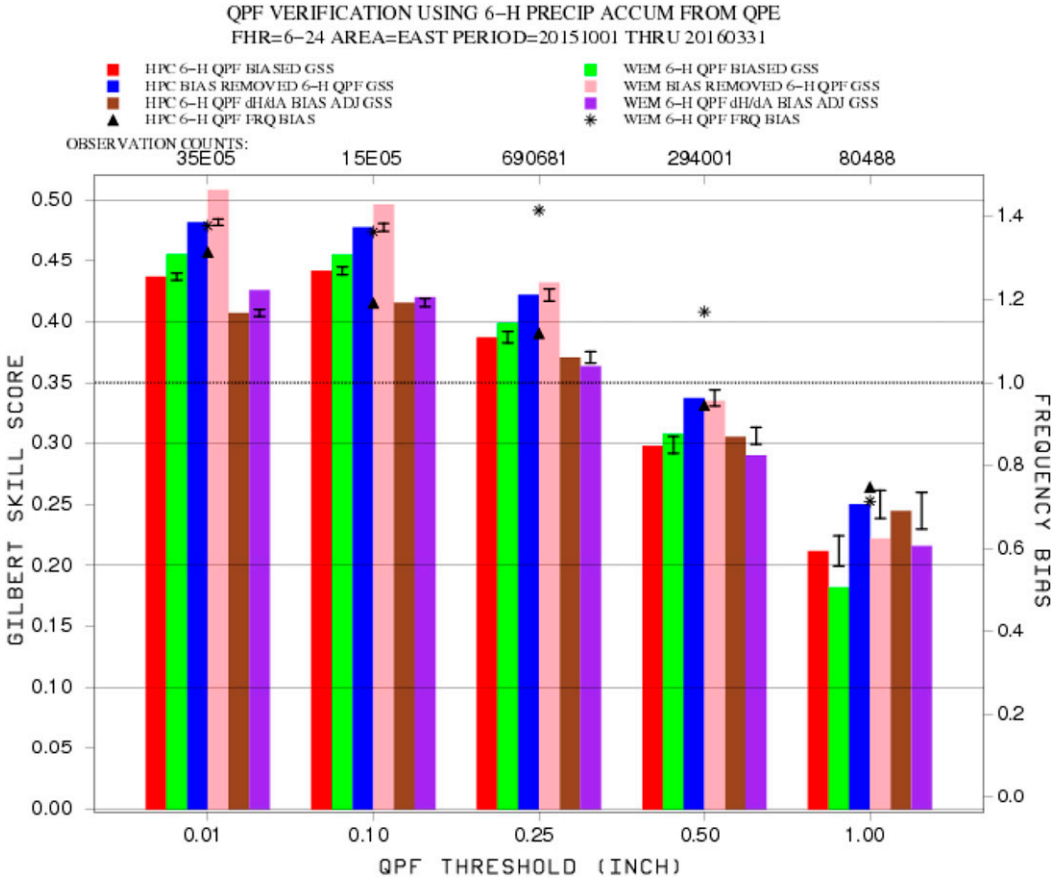

FIG. 5. As in Fig. 4, but for WPC and WEM for the 2015/16 cool season verified over the eastern CONUS. Though CIs are not shown for bias, the frequency bias of the two forecasts differs significantly for all thresholds less than 1 in.

24% of the forecasts removed by the BR method are hits; so, the BR GSS shows an increase to 0.4321 from the biased value of 0.3989 because 24% is well below the 31.5% limit indicated by the CPR. On the other hand, more than 39% of the forecasts removed by the BA method are hits, and the BA GSS value of 0.3634 is consequently less than the original biased value of 0.3989.

Had the CPR analysis been done for a case of bias increase to unity, the CPR value would serve as the lower bound for the fraction of added forecasts that must be hits for the score to improve. For the BR

TABLE 3. CPR analysis for threshold = 0.25 in. in Fig. 5. The first column gives the source of QPF, the columns labeled *H, F,* and *O* give the hit, forecast, and observed fractions, respectively. The frequency bias and GSS are shown in columns labeled Bias and GSS, respectively. The BR column pertains to the bias removal method, with H_r giving the hit rate after bias removal. The BA column pertains to the bias adjustment method, with H_a giving the bias adjusted hit rate from applying (2) or (3). The hit fraction for the removed forecasts is $\Delta H/\Delta F = (H_x - H)/O - F$, where H_x is either H_r or H_a. The $\Delta H/\Delta F$ fraction is to be compared to the value in the CPR column as discussed in the text.

QPF	*H*	*F*	*O*	Bias	GSS	CPR	BR	BA
WPC	0.044 02	0.078 69	0.070 28	1.120	0.3871	0.3101	$H_r = 0.043\,72$	$H_a = 0.040\,29$
							$\frac{\Delta H}{\Delta F} = 0.035\,67$	$\frac{\Delta H}{\Delta F} = 0.4435$
							GSS = 0.4219	GSS = 0.3708
WEM	0.051 41	0.099 48	0.070 28	1.415	0.3989	0.3153	$H_r = 0.044\,38$	$H_a = 0.039\,78$
							$\frac{\Delta H}{\Delta F} = 0.2408$	$\frac{\Delta H}{\Delta F} = 0.3983$
							GSS = 0.4321	GSS = 0.3634

method, the fraction of hits for forecasts removed or added directly depends on the geometric placement of the forecast precipitation relative to the analyzed placement. This is not true of the BA method, for which the hit fraction of added or removed forecasts depends on the contingency table values that are only indirectly related to the original relative placement of the forecast and analyzed precipitation. In the preceding analysis, the WEM, although overbiased, has better placement of its QPF than does the WPC forecast for the 0.25-in. threshold. In any case, the CPR formulation, which depends on the performance metric, quantifies the condition for improvement or degradation of that metric when forecasts are added or removed to bring the frequency bias to unity regardless of the method used.

At the 1-in. threshold in Fig. 5, the biases of the WEM and WPC forecasts are very similar, and it is reassuring that the biased, BA, and BR GSSs all are consistent in saying the WPC forecast is significantly better and that the magnitude of the correction relative to the biased GSS is comparable for both the BA and BR approaches. While the focus of this section is on scenarios where the BA and BR techniques lead to significantly different results, a majority of verification comparisons are more like this one, where the underlying differences in skill are large enough that different bias correction approaches give the same qualitative answer as to which forecast is better. However, comparisons of two systems with comparable skill and differing bias characteristics are of greatest interest, and are the comparisons for which the two bias-correcting approaches are most likely in disagreement as to which QPF is ''better.''

5. Summary and conclusions

Given the frequency bias sensitivity of 2×2 contingency table–based metrics such as the GSS, there long has been an interest in ways to reduce or eliminate any advantage (or disadvantage) that bias may provide to a system being evaluated. This study examined a pair of techniques (BA and BR) used to provide bias-corrected QPF skill scores, motivated by an interest in seeing how consistent they are in determining the more skillful of a pair of forecasts from a bias-corrected perspective with the assessment of statistical significance of differences. Each technique has some inherently desirable traits, along with some inherent problems or limitations.

The BA is readily applicable to any set of 2×2 contingency table data, and how it changes a skill score such as GSS can be understood unambiguously from the mathematical formulation. However, it is disconnected from the spatial distribution of the original forecast and is subject to the assumptions made in formulating the technique. The BR adjusts the QPF to exactly match the frequency distribution of the verifying analysis, and by doing so in the spatial domain the actual bias-corrected forecast can be viewed, providing graphical clarity on what is being changed by the removal of bias. However, it lacks easy applicability to historical contingency table statistics and cannot be mathematically analyzed in the clear way that the BA approach can be. Also, since the BR technique replaces forecast values with analysis values, it can in theory produce a perfect skill score (e.g., GSS = 1.0) at some threshold without any actual forecast hits, making it potentially troublesome in a conceptual sense (F. Mesinger 2017, personal communication). However, the conceptually troublesome possibility requires perfect relative placement of forecast and observed precipitation. This theoretical possibility and other scenarios discussed in this analysis highlight the importance of always showing frequency bias along with performance metrics, regardless of whether or not a correction is made for bias.

In terms of response, differences between the two approaches seemed to fall into two primary realms for the examples examined here. For high-biased light precipitation, such as the 0.01-in. threshold in Fig. 4, the BA approach lowered the GSS due to removing hits in reducing the bias to one, while the BR approach of trimming light precipitation in a statistically consistent and placement preserving way (Fig. 2) generated an increase in GSS. Countering the inherent advantage of high-biased forecasts has been a motivating factor for seeking bias-corrected approaches, and the BR approach often rewards slightly high-biased forecasts, but only if the relative (spatial) placement is good. In the case of correcting low-biased QPFs, the BA tends to provide a larger increase in GSS than does the BR approach, and this scenario is often where conclusions about which forecast is better are reversed due to this discrepancy between the BA and BR bias corrections. As the low-biased QPFs often are produced for higher-intensity thresholds from blended or ensemble mean products, these contradictory results based on choice of BA or BR bias correction often apply to verifications of the most impactful or significant weather events.

The different responses of the two bias-correction approaches, even if limited to specific scenarios, create limitations for historical cases for which the BR approach is not possible. The results here show that the BA approach on historical verification 2×2 contingency table data will not provide a bias-corrected result that would agree with the preferred BR technique for all bias scenarios. The authors recommend that the BR approach become the standard technique for bias-correcting QPF for verification. This recommendation

may imply the need to modify verification systems, such as the Model Evaluation Tools (MET) package (Fowler et al. 2018), which is becoming a standard verification tool across the breadth of the United States weather enterprise.

Acknowledgments. The authors wish to thank the three anonymous reviewers whose comments and suggestions motivated a more quantitative approach and greatly helped improve the clarity in the final paper. The first author began this work as a NOAA Rotational Assignment Program project in the Weather Prediction Center, which also funded the publication of this work. Any opinions expressed herein are those of the authors only, and not necessarily of the funding agency.

REFERENCES

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.

Brill, K. F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting*, **24**, 307–318, https://doi.org/10.1175/2008WAF2222144.1.

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

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

Fowler, T., J. Halley Gotway, K. Newman, T. Jensen, B. Brown, and R. Bullock, 2018: The Model Evaluation Tools v7.0 (METv7.0) user's guide. Developmental Testbed Center, 407 pp., http://www.dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v7.0.pdf.

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

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

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

Mesinger, F., 2008: Bias adjusted precipitation threat scores. *Adv. Geosci.*, **16**, 137–142, https://doi.org/10.5194/adgeo-16-137-2008.

——, and K. Brill, 2004: Bias normalized precipitation scores. *17th Conf. on Probability and Statistics/20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., J12.6, https://ams.confex.com/ams/84Annual/techprogram/paper_69561.htm.

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.

Panofsky, H. W., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology*. Pennsylvania State University Press, 224 pp.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.