Probabilistic Contingency Tables: An Improvement to Verify Probability Forecasts

SARAH GOLD AND EDWARD WHITE
Department of Mathematics and Statistics, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

WILLIAM ROEDER AND MIKE MCALEENAN
45th Weather Squadron, Cape Canaveral Air Force Station, Cape Canaveral, Florida

CHRISTINE SCHUBERT KABBAN
Department of Mathematics and Statistics, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

DARRYL AHNER
Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

(Manuscript received 10 June 2019, in final form 30 December 2019)

ABSTRACT

The 45th Weather Squadron (45 WS) records daily rain and lightning probabilistic forecasts and the associated binary event outcomes. Subsequently, they evaluate forecast performance and determine necessary adjustments with an established verification process. For deterministic outcomes, weather forecast analysis typically utilizes a traditional contingency table (TCT) for verification; however, the 45 WS uses an alternative tool, the probabilistic contingency table (PCT). Using the TCT for verification requires a threshold, typically at 50%, to dichotomize probabilistic forecasts. The PCT maintains the valuable information in probabilities and verifies the true forecasts being reported. Simulated forecasts and outcomes as well as 2015–18 45 WS data are utilized to compare forecast performance metrics produced from the TCT and PCT to determine which verification tool better reflects the quality of forecasts. Comparisons of frequency bias and other statistical metrics computed from both dichotomized and continuous forecasts reveal misrepresentative performance metrics from the TCT as well as a loss of information necessary for verification. PCT bias better reflects forecast verification in contrast to that of TCT bias, which suggests suboptimal forecasts when in fact the forecasts are accurate.

1. Introduction

Effective measures of probability forecasts are required to not only measure performance such that users can assess validity of those forecasts but also to ascertain if improvements have been achieved (Jolliffe and Stephenson 2011). This assessment may also include discussions of whether the effort was cost effective, worth the effort to implement, or of value (Murphy 1993; Brooks and Correia 2018). Many methods already exist to verify probability forecasts, each with strengths and weaknesses, with the ultimate aim of improving forecast methods (Fowler et al. 2012). Some of the more frequently used methods include the reliability diagram and the closely related attributes diagram, the sharpness diagram, and the ratio skill score, which measures the skill of a set of probability forecasts against a baseline method, typically a zero to low skill score. The Brier skill score is a frequently used ratio skill score (Bradley et al. 2008; Wilks 2010) where the baseline is climatology. Other methods of verifying probability forecasts may utilize the receiver operating characteristic (ROC) curve (Mason 1982), the ranked probability score or the linear error in probability space (LEPS) score (Potts et al. 1996) or the root-mean-square error (Stanski et al. 1989). Jolliffe and Stephenson (2011) and Wilks (2011), to include other articles and texts, discuss these and other forecast verification techniques in detail.

The variety of verification techniques speaks to their diverse usage and audience. As noted by Mason (2008),

DOI: 10.1175/WAF-D-19-0116.1

For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
a neglected property of verification scores is their understandability. Consequently, choosing one particular verification method over another might result in a less optimal application, but more intuitively appealing in communicating the quality of forecasts to nonspecialists. As an example, contingency tables provide a common method of verifying probability forecasts. While academia may rarely teach verifying probability forecasts with contingency tables, it is often practiced in operational meteorology (specifically the U.S. military), presumably because of the ease of interpretation of the performance metrics. Air Weather Service (1991) documents the use of contingency tables, and illustrative examples include Gremillion and Orville (1999), Yang and King (2010), Travis (2015), Olsen (2018), and Katuzienski (2019).

Wilks (2011) discusses the conversion of probability forecasts to categorical forecasts. In particular, the customary approach entails converting the probability forecasts into categorical yes/no outcomes, where a probability exceeding a certain threshold is considered a “yes,” the event is forecasted to occur, and less than that threshold is considered a “no,” the event is forecasted not to occur. These categorized forecasts are then entered into the contingency table and the normal suite of forecast performance metrics calculated. Unfortunately, this approach is somewhat flawed. First, most of the information in the probability forecasts is lost in the categorization process (Mason 1979; Zhang and Casey 2000). Second, the designated threshold to which a forecast shall be compared implicitly assumes the forecasts will be unbiased, which is not necessarily true. Last, this cutoff may be arbitrary, for example a 50% threshold or a mean value discriminator (e.g., climatology). However, despite these flaws, this verification method is used because the suite of forecast performance metrics are generally easy to explain. Those easier to understand metrics typically include probability of detection, probability of false alarm or false alarm ratio, and some measure of skill such as true skill statistic, Heidke skill score, critical success index, and others.

The 45th Weather Squadron (45 WS) is one of the few Air Force units that issues probability forecasts. The 45 WS at Cape Canaveral Air Force Station provides weather support prior to and including the day of a space vehicle launch. Their probability forecasts include their daily 24-h and 7-day planning forecasts, which include the probability of precipitation and probability of lightning for various time periods of the day. The 45 WS verifies these probability forecasts to provide forecast performance feedback to their forecasters, the unit leadership, and their customers, and to determine if new methods improve the performance. In this role, the 45 WS has developed an alternative approach to verify probability forecasts, the probabilistic contingency table (PCT). The PCT preserves the simplicity of the traditional contingency table metrics, but preserves more of the probability information in the forecasts, and automatically accounts for biased forecasting. However, the PCT is a nonstandard tool in forecast verification. Therefore, the 45 WS pursued an independent objective verification of this method with the Air Force Institute of Technology. This paper presents the results of that study, which was done as a master’s thesis (Gold 2019).

2. Background

Classification models generally dichotomize continuous responses (forecast probabilities in our case) into a zero or one class output. Comparison of predictions produced from a binary classification model and the observed outcome is usually arranged in a 2 × 2 contingency table and reflect whole number values. These four cells record true positives (hits), true negatives [correct negatives (CN)], false positives [false alarms (FAs)], and false negatives (misses), which consequently generate metrics to evaluate a model’s classification performance. Metrics commonly reported from the table include the following: accuracy, precision, negative predictive value, sensitivity, and specificity. Table 1 illustrates a typical 2 × 2 contingency table along with associated definitions and metrics to include a relational meteorological terminology. Note that every value in the main 2 × 2 table reflects a whole number.

The contingency table metrics used here are broadly described as follows. Precision, also known as positive predictive value, is the proportion of all positive predictions correctly classified. Negative predictive value is the proportion of all negative predictions correctly classified. Sensitivity, also known as recall, is the proportion of all occurring events correctly classified as likely to occur. Specificity is the proportion of all nonoccurring events correctly classified as not likely to occur. Last, accuracy is the ratio of correctly predicted events over total number of events. Many of the same metrics used to evaluate a general classification model are utilized in weather verification, but are often identified by different names.

Translating some of the terms in Table 1 utilizing meteorological terminology results in the following. Accuracy remains the same in both instances. Probability of detection (POD), also known as hit rate or the generalized term of sensitivity, ranges from zero to one; one represents a perfect score. POD is sensitive to hits but ignores false alarms. Probability of false detection (POFD), also known as false alarm rate, is not often reported alone but is used in concurrence with POD. POFD is sensitive to false alarms but ignores misses. POFD is also
the probability complement of specificity from Table 1. Probability of false alarm (POFA), also known as false alarm ratio or the generalized term precision, is sensitive to false alarms but ignores misses. POFA is more informative in conjunction with POD. Frequency bias [see Eq. (1)] measures the ratio of the frequency of forecast events to frequency of observed events (Fowler et al. 2012). The relative frequency does not measure how well the forecasts and observations correspond. The metric does reveal whether the model underforecasts or overforecasts. A perfect score of one indicates no bias, less than one represents under forecasting, and a score over one represents overforecasting (Wilks 2011):

$$\text{bias} = \frac{\text{hit} + \text{FA}}{\text{hit} + \text{miss}}$$  \hspace{1cm} (1)

Additional skill scores, relegated to a contingency table setting and measuring the predictive ability against a reference forecast, may also be used to evaluate performance. As a reference hereinafter, Table 2 lists discussed metrics, ranges, and optimal values. The scores addressed in this paper for comparison purposes include the threat score (TS), the Kuiper skill score (KSS), and the Heidke skill score (HSS). The threat score [see Eq. (2)], also known as critical success index, does not consider the correct negatives and assesses how well the forecasted likely to occur events correspond to the observed events. It ranges from zero to one, with zero being poor, and one being perfect. TS tends to excessively penalize predictions of rare events but is still a more balanced single metric than POD and POFA (Jolliffe and Stephenson 2011). Another way to view this is that POD and POFA by themselves can be extremely misleading. For example, if one were to always forecast the event, the POD would be perfect, even though the forecasts would be of no value to the user (Murphy 1996). Likewise, if one were to never forecast the event, the POFA would be perfect, though again the forecasts would be of no value. Prudent forecasting balances between POD and POFA, which is measured by various skill scores.

The Kuiper skill score [see Eq. (3)] compares forecast skill to random chance, with a score of zero representing random chance. The formulation is equivalent to POD minus POFD, and for rare events POFD is very small, resulting in KSS converging to POD (World Meteorological Organization 2014). Last, the Heidke skill score [see Eq. (4)] compares the prediction performance to a reference accuracy measure. The reference measure in the HSS is the proportion correct that would be achieved with random forecasts independent of observations. The score ranges from zero to one. Forecasts equivalent to reference forecasts produce a HSS of zero and perfect forecasts receive a HSS of one (Wilks 2011). It is unlikely that a single metric provides enough information to make optimal forecasting adjustments; therefore, all of the discussed weather metrics should be addressed collectively:

$$\text{TS} = \frac{\text{hit}}{\text{hit} + \text{miss} + \text{FA}}$$  \hspace{1cm} (2)

$$\text{KSS} = \frac{(\text{hit} \times \text{CN}) - (\text{miss} \times \text{FA})}{(\text{hit} + \text{miss})(\text{FA} + \text{CN})}$$, and  \hspace{1cm} (3)

$$\text{HSS} = \frac{2 \times (\text{hit} \times \text{CN}) - (\text{miss} \times \text{FA})}{(\text{hit} + \text{miss})(\text{hit} + \text{CN})(\text{hit} + \text{FA})(\text{FA} + \text{CN})}$$  \hspace{1cm} (4)

Weather forecast predictions may provide probabilistic values, which can be converted to binary classifications.
for verification purposes. Forecast accuracies can then be verified using a traditional contingency table (TCT); however, all four main elements of the $2 \times 2$ table in Table 1, hits, misses, CNs, and FAs, entail whole numbers. That is, all the probabilities have been converted to a 0 or 1. Consequently, performance metrics such as POD, POFA, and frequency bias calculated from the TCT may not capture the weather model’s true forecasts and therefore may not be an appropriate validation tool.

**Probabilistic contingency table**

Given the constraint of maintaining the user friendliness offered by the presentation of the TCT format (as required by the customers of the 45 WS), we present an alternative tool, a probabilistic contingency table (PCT) displayed in Table 3 that incorporates probabilistic forecasts in order to possibly improve upon model verification. The row vector $\mathbf{p}$ represents probabilities of an event occurring ($p_1, p_2, \ldots, p_n$) during sample size $i = 1$ to $n$ total days. Row vector $\mathbf{k}$ is the corresponding binary outcomes ($k_1, k_2, \ldots, k_n$), one representing an event occurred and zero representing an event did not occur. Therefore, through vector products between $\mathbf{p}$ and $\mathbf{k}$, a PCT can be created where the resulting probability values are maintained and compared to event outcomes in lieu of whole numbers related to hits, misses, FAs, and CNs. For example, the vector dot product of $\mathbf{p}$ and $\mathbf{k}$ represents hits. We next document the procedures to compare the TCT to the proposed PCT approach using both simulated and empirical data.

### Table 3. Probabilistic contingency table cell values. Vector $\mathbf{p}$ represents probabilities of an event occurring, while $\mathbf{k}$ is the corresponding binary vector of outcomes. All expected cell counts involve vector products, as noted by the $\cdot$ operator.

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed outcome</th>
<th>Yes</th>
<th>No</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$E_{11} = p_k$</td>
<td>$E_{12} = p(1-k)$</td>
<td>$E_{11} + E_{12}$</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>$E_{21} = (1-p)k$</td>
<td>$E_{22} = (1-p)(1-k)$</td>
<td>$E_{21} + E_{22}$</td>
<td></td>
</tr>
<tr>
<td>Column total</td>
<td>$E_{11} + E_{21}$</td>
<td>$E_{12} + E_{22}$</td>
<td>$E_{11} + E_{12} + E_{21} + E_{22}$</td>
<td></td>
</tr>
</tbody>
</table>

3. **Methodology**

This section describes the three different ways in which we compare TCT and PCT metrics using simulated as well as real-world forecasts and observations from the 45 WS. The first comparison only entails simulating forecasts and accompanying observations with varying bias. This process assumes that each forecast occurs an equal number of times. Although completely unrealistic in practice, such uniformity establishes a controlled baseline. The second comparison addresses and incorporates seasonal trends in the data as provided by the 45 WS. Using these empirical distributions, the number of simulated observations for each forecast varied according to the appropriate weight, which provides a more realistic comparison in practice. The last comparison of TCT and PCT metrics utilizes only empirical (real-world) data from forecasts and observations for 4 years of rain and lightning events as reported by the 45 WS. No simulated data are involved in this final comparison.

To establish fidelity in the comparison of the TCT and PCT with real data, simulated weather observations were first used to establish a controlled evaluation. Using R programming language and the RStudio integrated development software environment, observations were generated with respect to forecasts ranging from 0% to 100%, incremented by 10%. Reliable (i.e., unbiased forecasts) were first considered followed by overforecasting and underforecasting by 5% increments varying from 5% to 20% (i.e., the bias). A standard uniform distribution, $U(0, 1)$, was adopted to generate all unbiased observations. A hit occurs if the randomly generated number was less than or equal to the forecast percent. A correct negative occurs if the randomly generated number is less than or equal to 100% minus the forecast percent. For example, if the forecast percent for an event to occur equaled 20%, a hit would only occur if the randomly generated uniform distribution value for that simulation run was equal to 0.20 or less. Conversely, if the forecast percent of an event to not occur was 80% (the complement of a 20% forecasted event), a correct negative would occur only if the randomly generated uniform distribution for that simulation run was equal to or less than 0.80. In total, 30 replications of the 10,000 observations were generated separately for each forecast level (0%, 10%, $\ldots$, 100%) and its complement (100%, 90%, $\ldots$, 0%).

To calculate the predicted values for each TCT, we adopted the 50% threshold for dichotomizing percents into whole numbers. That is, any forecast less than 50% was predicted as not likely to occur and coded as a “0”. Therefore, hits were recorded as either zero for all forecasts under 50% or as the number of observations for forecasts greater than or equal to 50%. False alarms...
were also zero for all forecasts under 50% and 10 000 minus the number of hits for forecasts greater than or equal to 50%. The number of correct negatives simulated for each TCT generated from a less than 50% forecast was equal to the number of simulations where the randomly generated value was less than or equal to the forecasted percent of an event not likely to occur. The number of correct negatives for the TCTs from greater than or equal to 50% forecasts was zero. Misses were recorded as 10 000 minus the number of correct negatives for forecasts less than 50% and zero for all forecasts greater than or equal to 50%. From each replication for the unbiased forecasts, 11 TCTs and PCTs were produced.

The eight metrics—accuracy, POD, POFD, POFA, TS, KSS, HSS, and frequency bias—were calculated for each contingency table. Following this computation, the metrics were averaged for the thirty replications of each forecast level. Last, the averaged TCT metrics for each forecast were subtracted from the corresponding PCT metrics to calculate the effect sizes of the differing tables. The effect size was then graphed against the forecasts to compare the performance of TCT and PCT for each forecast.

The biased forecast simulations follow the same process as the unbiased cases, but with varying to occur/not to occur forecast thresholds. At each level of bias, 30 replications were again simulated for each forecast level. Not all forecasts are possible with the different levels of bias. For example, it is not possible to predict a forecast of 10%, over biased by 20% as you cannot predict under 0%. Contingency tables were produced for each applicable forecast in the same manner as the unbiased simulations. The eight metrics were calculated and used to determine effect sizes to graph the trends in the disparity between TCT and PCT for increasing biases of both over- and underforecasting.

For the second method of comparing metrics from the TCT and PCT, we incorporated both simulated as well as real-world forecasts and observations since it is not realistic to assume only one level of a forecast will be predicted repeatedly. Additionally, it is also not reasonable to assume all forecast levels are predicted the same number of times during the year. To relax these assumptions, four calendar years of data (2015–18) from the 45 WS involving daily 24-h forecasts of rain and 24-h forecasts of lightning were used to fit representative empirical distributions in order to incorporate seasonal trends. Statistical testing outside the scope of this paper indicates a warm season (June–September) and a cold season (October–May) for the 45 WS forecasting areas with respect to rain and lightning.

After determining seasons (warm/cold) according to rain and lightning forecasts, the data were divided by the respective seasons. The true observed 45 WS counts of each forecast in the given season, along with its distribution, are displayed in Table 4 and used to determine the empirical proportion of times a rain (or lightning) forecast was predicted 24 h prior to the event. To calculate the proportions for unbiased forecasts, the count was simply divided by the total number of forecasts in the seasonal data. For biased forecasts, the count of forecasts marked as N/A were removed from the total number of forecasts to determine the proportional values.

With the seasonal distributions, four different simulations were generated to compare the TCT and PCT. The likely to occur and not likely to occur matrices were calculated by simulating the weighted forecasts. Rather than simulating 10 000 runs for each forecast, the proportional number of runs were simulated. As an example, the 10 000 unbiased likely to occur forecasts for cold season lightning events were multiplied by the respective count and divided by the total number of cold season lightning events.

The number of hits, false alarms, misses, and correct negatives were calculated as before for each individual
forecast. The respective cells of the contingency tables were then summed across all forecasts and averaged for the thirty replications resulting in one TCT and one PCT for the unbiased simulation. This was done for all bias levels of overforecast and underforecast, adjusting the number of runs simulated for each level of bias. The effect sizes from the PCT minus the TCT metric values were then graphed against the bias level for each seasonal event.

The varied simulations allow for comparing TCT and PCT in a controlled environment, but the last comparison strictly adopts only observed empirical data. For this, only data provided by the 45 WS were used to compare the TCT and PCT for both rain and lightning events in the cold and warm seasons. The data divided by seasons was used to determine hits, false alarms, misses, and correct negatives of the TCT and PCT. Afterward, we calculated the same eight metrics as in the simulations: accuracy, POD, POFD, POFA, TS, KSS, HSS, and frequency bias. Similarly, the effect sizes of the metrics calculated from the PCTs minus the TCTs were graphed separately for cold season rain forecasts, warm season rain forecasts, cold season lightning forecasts, and warm season lightning forecasts. In the next section, we present the results of comparing TCT and PCT metrics under the three aforementioned scenarios.

4. Results

We first address the overall trends for the unweighted simulation for the effect sizes for POD, POFD, TS, accuracy, and absolute bias (described shortly) for both over and underforecasts; POFA, KSS, and HSS had negligible differences of less than $10^{-2}$ effect sizes. The effect size for each particular metric was calculated by subtracting the value produced using the TCT from the value produced using the PCT. The effect size of the
frequency bias metric is not directly informative. Since a perfect score for frequency bias is one, a modified comparison of biases, denoted as absolute bias, was calculated as the effect size of the differences from one. A positive value for this metric indicates the TCT has values closer to the perfect score.

Overall, the metrics calculated after dichotomizing the forecasts for the TCT did not converge to the probabilistic results, which is not unexpected. As anticipated, the bias of over- or underforecasts produced a shifted trend as displayed by the unbiased case. The greatest discrepancy occurs when comparing TCT to PCT around the threshold of 50%, which is also to be expected since TCT rounds or truncates at this arbitrary boundary. Furthermore, the TCT frequency bias differed from the anticipated perfect value of one even for the unbiased result. Although unexpected, this finding suggests that frequency bias might be a key metric to investigate further. Figure 1 highlights the finding with respect to bias.

Next, we examine the metric effect sizes according to the four empirical distributions: warm and cold seasons.
rain and lightning events; we separate these results by over- and underforecasting. Figures 2 and 3 illustrate the PCT and TCT metric comparisons when incorporating the empirical seasonal 45 WS forecasts trends. Overall, the TCT metrics display values closer to the ideal scores in regard to accuracy, POD, POFD, POFA, TS, KSS, and HSS for all varying biased forecasts for both seasons. The real differences lie with respect to the absolute frequency bias. PCT frequency bias appears closer to one for all unbiased cases regardless of season or event. For the cold season, PCT frequency bias appears closer to one for all underforecasts, while TCT appeared closer for all biased overforecasts. For the warm season, this pattern inverts. That is, PCT frequency bias appears closer to one for all overforecasts, while TCT appears closer for all biased underforecasts.

As noted in the first simulation results, there appears to be a consist pattern about how TCT and PCT compare with respect to frequency bias. PCT appears to track closer to the simulated truth than TCT does for this particular metric with respect to unbiased forecasts.
Table 5 displays the unbiased weighted simulation results for both seasons (cold/warm) for both weather events (rain/lightning). As evident by the numbers, TCT metrics are closer to the optimal values in comparison to the PCT metrics except for frequency bias (highlighted in bold). There, PCT bias often equals the optimal value of 1.

This last section explores the comparison of the TCT and PCT metrics using the 45 WS forecasts and outcomes for calendar years 2015–18. Table 6 displays this empirical data, which involves no simulations. When comparing these empirical data to the weighted simulation results in Table 5, the numbers track reasonably close to each other lending support to our methodology and consistency. The noted exception is the bias for the warm season lightning, which is highlighted in bold in Table 6. This inconsistency with simulated results reveals an inversion in that the TCT bias reflects a closer value to 1 in lieu of the PCT bias.

Consequently, the TCT frequency bias metric suggests future forecasting should not be altered. However, our simulations suggest otherwise given the PCT frequency bias should be higher. When comparing the bias from 5% underforecast simulation (see Table 7), these results are consistent with the real-world data and reveals a forecast unreliability, which is what the bias metric of PCT metric is highly suggesting. Investigating further, Fig. 4 shows what has been occurring. Lightning has been occurring more often than predicted at the 45 WS, starting around a forecasted prediction of 40% to approximately 75%. So overall, the 45 WS empirical metrics align better with the 5% underforecasting than the unbiased simulation. The PCT frequency bias metric revealed this inconsistency. We recognize there is not one universal best meteorological metric to focus on. However, when comparing TCT to PCT, at least in our reported data, the PCT verification methodology did detect an underforecast bias when predicting lightning during the warm season that the TCT overlooked or did not detect.

Last, we now compare the PCT results to the TCT approach again; however, this time the dichotomizing percentage is not 50%. Instead, we compare how the PCT results compare to using climatology (Climo) as the basis of determining a “0” or “1” (weather event will not occur/weather event will occur). That is, if probabilistic forecast equals or exceeds the climatological event frequency, then it is coded a 1 else it is coded a 0. In addition, we present how TCT, PCT, and Climo compare with respect to the Brier score. Table 8 presents these consolidated findings. Overall, we see the
same performance of PCT with respect to frequency bias even when compared to dichotomizing via climatology. Also, the PCT approach presented the closest optimal value of the Brier score, 0, in all incidents.

5. Conclusions

In this paper, we compared a traditional $2 \times 2$ contingency table used for verifying binary weather events to a proposed alternative tool termed a probabilistic contingency table. This comparison entailed three stages. First, we simulated uniform weather events and computed hits, false alarms, misses, and correct negatives. This simulation revealed the greatest discrepancy around the threshold of 50% when comparing TCT to PCT. It also first revealed the TCT frequency bias differed from the expected perfect value of one. Varying levels of bias were tested to determine the trends and differences in metric values derived from TCT and PCT approaches assuming the forecasts were not perfectly reliable. Bias levels did not affect POD, POFD, POFA, KSS, or HSS for the individual forecast simulations.

Next, we incorporated simulating seasonal patterns based on the number of rain or lightning events forecasted by the 45 WS during calendar years 2015–18. The averaged metrics revealed values closer to the perfect scores utilizing the TCT for all except again frequency bias. Different levels of bias were again added to the simulation, resulting in increasing or decreasing effect size trends for each over- and underforecasting simulation by season and event type.

Following the simulations, we used just the empirical data that produced the simulation weights and investigated the same metrics for comparison purposes. The metric values for three of the four seasonal events were consistent with the unbiased weighted simulations. For the warm season lightning events, empirical results did not completely agree with the associated unbiased simulations. The disconnect lay with the frequency bias. Based on a reliability diagram, the 45 WS warm season lightning events appear to be slightly underforecasted; however, the TCT produced a frequency bias closer to one whereas the PCT indicated a possible prediction forecast issue to investigate further.

Overall, comparing various meteorological metrics for the TCT and PCT reveal an apparent difference between the two tools in analyzing the 45 WS data. The effect sizes do not reveal which tool is “better,” but rather emphasize that the TCT and PCT metrics calculated from the same data are different. The metrics

| TABLE 6. Empirically generated metrics for both seasons (cold and warm) and weather events (rain and lightning) for calendar years 2015–18. For warm season lightning, the bold bias values indicate a discrepancy in comparisons between the TCT and PCT. |
|---|---|---|---|---|
| Cold season | Warm season | Rain | Lightning | Rain | Lightning |
| TCT | PCT | TCT | PCT | TCT | PCT | TCT | PCT |
| Accuracy | 0.836 | 0.886 | 0.759 | 0.881 | 0.915 | 0.775 | 0.881 | 0.799 |
| POD | 0.683 | 0.759 | 0.544 | 0.647 | 0.710 | 0.550 | 0.667 | 0.710 |
| POFA | 0.210 | 0.348 | 0.316 | 0.470 | 0.412 | 0.388 | 0.461 | 0.390 |
| POFD | 0.089 | 0.450 | 0.078 | 0.390 | 0.077 | 0.316 | 0.089 | 0.390 |
| TS | 0.578 | 0.478 | 0.450 | 0.367 | 0.573 | 0.478 | 0.450 | 0.367 |
| KSS | 0.590 | 0.472 | 0.531 | 0.524 | 0.573 | 0.472 | 0.531 | 0.524 |
| Bias | 0.865 | 0.982 | 0.831 | 1.026 | 1.109 | 0.986 | 0.930 | 1.026 |
| HSS | 0.615 | 0.475 | 0.573 | 0.475 | 0.573 | 0.475 | 0.573 | 0.475 |
calculated after dichotomizing the forecasts did not converge to the probabilistic results. The TCT is artificially favored for the metrics examined with the exception of frequency bias. The unbiased simulations conducted should have produced a frequency bias of one, but the dichotomization of the probabilities resulted in metrics indicating biased forecasts. The frequency bias from the PCT for the same simulations, however, did show a bias of one, revealing the PCT as more representative of the true forecasts and outcomes.

Forecast adjustments made as a result of utilizing the TCT for verification would incorrectly result in bias driven modifications. For example, the warm season lightning frequency bias from the TCT indicates the 45 WS forecasters should not alter the frequency of predicting lightning events when the PCT correctly indicated an underforecast of approximately 5% and that the prediction for lightning should be adjusted upward. Given that lightning plays a key role in space launch delays at Cape Canaveral, forecast verification and adjustment of model prediction of lightning is imperative.

Dichotomizing can often result due to an aversion to probabilities (Altman and Royston 2006), however, it diminishes the available information (Abramson and Clemen 1995; Zhang and Casey 2000). PCTs evaluate forecast performance based on the actual forecasts that are reported rather than the arbitrarily dichotomized forecasts. The resulting metrics from the PCT reveal a representative evaluation of the forecasts. To promote better forecasts, the verification process should evaluate the true probabilistic prediction determined by the forecaster through utilization of the PCT and seasonal reliability diagrams.

Verification techniques should not overestimate or underestimate forecast performance. The metrics addressed in this research all attempt to measure the strength in forecast predictions. Since each metric addresses different considerations of performance, the importance

| Unbiased | TCT | Accuracy | 0.757 |
|          |     | POD      | 0.802 |
|          |     | POFA     | 0.243 |
|          |     | POFD     | 0.294 |
|          |     | TS       | 0.638 |
|          |     | KSS      | 0.509 |
| Bias     | 1.060 |
| HSS      | 0.511 |

| 5% underforecast | TCT | Accuracy | 0.729 |
|                 |     | POD      | 0.728 |
|                 |     | POFA     | 0.252 |
|                 |     | POFD     | 0.270 |
|                 |     | TS       | 0.585 |
|                 |     | KSS      | 0.457 |
| Bias            | 0.973 |
| HSS             | 0.457 |

**TABLE 7.** Comparison of warm season lightning unbiased and 5% underforecasted weighted simulation results. Frequency bias values are bolded.

![Fig. 4. Warm season reliability curve for the 45 WS for calendar years 2015–18. Forecast probabilities are on the x axis, while the y axis represents the occurrence percentage at the corresponding forecast probabilities.](image-url)
of each metric differs depending on the user’s priorities. An overestimation would provide a false sense of adequacy and dismiss areas needing to be improved. An underestimation may result in unnecessary changes and/or adjustments and may potentially reduce forecast efficacy and performance. Therefore, optimal techniques must be implemented to ensure appropriate and productive verification to make certain event outcomes appropriately match predicted forecasts.

Acknowledgments. Many thanks to friends and colleagues who assisted in this research and in particular the reviewers for their constructive comments and suggestions. This work was partially supported by the 45th Weather Squadron.

REFERENCES


Downloaded from http://journals.ametsoc.org/waf/article-pdf/35/2/609/4924514/wafd190116.pdf by guest on 17 September 2020