



**EFFECT OF USING PROBABILISTIC CONTINGENCY TABLES TO MODIFY
FORECAST PREDICTIONS**

THESIS

Sarah A. Gold, Second Lieutenant, USAF

AFIT-ENS-MS-19-M-117

**DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY**

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-MS-19-M-117

EFFECT OF USING PROBABILISTIC CONTINGENCY TABLES TO MODIFY
FORECAST PREDICTIONS

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the
Degree of Master of Science in Operations Research

Sarah A. Gold, BS

Second Lieutenant, USAF

March 2019

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENS-MS-19-M-117

EFFECT OF USING PROBABILISTIC CONTINGENCY TABLES TO MODIFY
FORECAST PREDICTIONS

THESIS

Sarah A. Gold, BS
Second Lieutenant, USAF

Committee Membership:

Dr. Edward D. White
Chair

Dr. Darryl K. Ahner
Member

Dr. Christine M. Schubert-Kabban
Member

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 implemented verification process. For deterministic outcomes, weather forecast analysis typically utilizes a Tradition 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 2015-2018 45 WS data were utilized to compare forecast performance metrics produced from the TCT and PCT to determine which verification tool better supports producing the greatest quality forecasts. Comparisons of frequency bias, reliability, and Brier Score (BS) computed from both dichotomized and continuous forecasts revealed misrepresentative performance metrics from the TCT as well as a loss of information necessary for verification. Exploration of the 45 WS data with a probabilistic verification process revealed a need to verify seasonally as well as slightly unreliable warm season lightning event forecasts.

Acknowledgments

I would like to express my sincerest gratitude to my faculty advisor, Dr. Tony White, for his patience, diligence, and mentorship throughout this thesis process. I would also like to thank my readers, Dr. Darryl Ahner and Dr. Christine Schubert-Kabban, as well as the members of the 45 WS, William Roeder and Mike Mcaleenan, for your many contributions. Lastly I would like to thank my friends and family for their constant support and encouragement.

Sarah A. Gold

Table of Contents

	Page
Abstract	iv
Acknowledgments.....	v
Table of Contents	vi
List of Figures	viii
List of Tables	x
I. Introduction	1
1.1 Background.....	1
1.2 Research Motivation.....	1
1.3 Traditional Contingency Table.....	2
1.4 Probabilistic Contingency Table	4
1.5 Methodology.....	6
1.6 Overview	6
II. Literature Review	8
2.1 Introduction	8
2.2 Evaluating General Classification Performance.....	8
2.3 General Uses of Contingency Tables	14
2.4 Evaluating Forecast Performance.....	15
2.5 Probabilistic Contingency Table	24
III. Methodology	26
3.1 Chapter Overview.....	26
3.2 Simulation Assuming Equal Forecasts.....	26
3.3 Simulation Using 45 WS Empirical Distributions	33
3.4 45 WS Forecasts & Observations.....	39

3.5 Summary.....	40
IV. Analysis and Results.....	41
4.1 Chapter Overview.....	41
4.2 Simulation Results and Trends - Unweighted	41
4.3 Simulation Results and Trends - Weighted	50
4.4 45 WS Forecasts and Observations	55
V. Conclusions and Recommendations	64
5.2 Discussion.....	64
5.3 Summary.....	65
5.4 Future Research Recommendations	66
5.5 Final Remarks.....	67
Appendix A.....	69
Appendix B	71
Appendix C	79
Bibliography	82

List of Figures

	Page
Figure 1. Precision- Recall Curve.....	11
Figure 2. ROC Curve	12
Figure 3. Reliability Diagram Modified from the 45 WS for Lightning Predictions and Observations.....	23
Figure 4. Examples of Varying Forecast Sharpness	24
Figure 5. 45 WS (2015-2018) Rain Forecast Distributions by Month	35
Figure 6. 45 WS (2015-2018) Lightning Forecast Distributions by Month	37
Figure 7. Accuracy Simulation Effect Size.....	43
Figure 8. POD Simulation Effect Size.....	44
Figure 9. POFD Simulation Effect Size.....	45
Figure 10. POFA Simulation Effect Size.....	45
Figure 11. TS Simulation Effect Size	47
Figure 12. KSS Simulation Effect Size.....	47
Figure 13. HSS Simulation Effect Size.....	48
Figure 14. Bias and Absolute Bias Simulation Effect Size	49
Figure 15. Seasonal Rain and Lightning Distributions.....	51
Figure 16. Weighted Cold Season Rain Effect Size	52
Figure 17. Weighted Warm Season Rain Effect Size	53
Figure 18. Weighted Cold Season Lightning Effect Size	54
Figure 19. Weighted Warm Season Lightning Effect Size.....	54
Figure 20. Cold Season Rain 45 WS Metrics	56

Figure 21. Warm Season Rain 45 WS Metrics	57
Figure 22. Cold Season Lightning 45 WS Metrics	58
Figure 23. Warm Season Lightning 45 WS Metrics.....	59
Figure 24. 2015-2018 45 WS Dichotomized Forecast Reliability Diagram.....	62
Figure 25. 2015-2018 45 WS Probabilistic Forecast Reliability Diagram	62

List of Tables

	Page
Table 1. Traditional Contingency Table	3
Table 2. Probabilistic Contingency Table.....	4
Table 3. Example of 7-Day Dichotomous and Probabilistic Forecasts, and Outcomes	5
Table 4. TCT & PCT Example Calculated from Identical Forecasts and Outcomes	5
Table 5. Contingency Table Containing General Classification Terms, and Metrics	9
Table 6. Contingency Table Containing Weather Verification Terms	17
Table 7. Unbiased Thresholds for Likely to Occur Binned Forecast Simulations	27
Table 8. Unbiased Thresholds for Not Likely to Occur Binned Forecast Simulations	27
Table 9. Unbiased Binned Forecast Simulation: Events Observed for Likely to Occur Simulation, Non-events Observed for Not Likely to Occur Simulation, and Associated Dichotomous and Probabilistic Event/Non-event Forecasts	29
Table 10. Unbiased TCT Hits, False Alarms, Misses, and Correct Negatives by Ascending Forecast Bin	29
Table 11. TCTs Produced From Unbiased Binned Forecast Simulations	30
Table 12. Unbiased TCT Hits, False Alarms, Misses, and Correct Negatives by Ascending Forecast Bin	31
Table 13. PCTs Produced From Unbiased Binned Forecast Simulations	31
Table 14. Over-forecast Thresholds for Likely to Occur & Not Likely to Occur Binned Forecast Simulations	32
Table 15. Under-forecast Thresholds for Likely to Occur & Not Likely to Occur Binned Forecast Simulations	33

Table 16. Number of Cold Season Rain Forecasts for 2015-2018	36
Table 17. Number of Warm Season Rain Forecasts for 2015-2018	36
Table 18. Tukey’s Comparisons of Lightning and Rain Forecasts by Season	36
Table 19. Number of Cold Season Lightning Forecasts	38
Table 20. Number of Warm Season Lightning Forecasts	38
Table 21. Metric Ranges and Associated Perfect Scores	42
Table 22. Cold Season Rain Unbiased Weighted Simulation Results	56
Table 23. Percent Difference in Unbiased Weighted Simulation and 45 WS Data (Cold Season Rain)	56
Table 24. Warm Season Rain Unbiased Weighted Simulation Results	57
Table 25. Percent Difference in Unbiased Weighted Simulation and 45 WS Data (Cold Season Rain)	57
Table 26. Cold Season Lightning Unbiased Weighted Simulation Results	58
Table 27. Percent Difference in Unbiased Weighted Simulation and 45 WS Data (Cold Season Lightning)	58
Table 28. Warm Season Lightning Unbiased & 5% Under-Forecasted Weighted Simulation Results	60
Table 29. Percent Difference in Unbiased and 5% Under-Forecasted Weighted Simulations and 45 WS Data (Warm Season Lightning)	60
Table 30. Brier Scores for Dichotomized and Probabilistic 45 WS Forecasts	61

EFFECT OF USING PROBABILISTIC CONTINGENCY TABLES TO MODIFY FORECAST PREDICTIONS

I. Introduction

1.1 Background

Proper weather forecasts are integral to maintaining the 45th Weather Squadron's (45 WS) mission of safe access to air and space. The space program is very dependent on strict weather requirements for rocket launch preparation and execution (Harms et al., 1999). According to Roeder et al. (2014), the leading cause of launch delays and cancellations is due to weather not meeting specified criteria. Additionally, poor weather impacts ground processing in preparation for the launches, influencing billions of dollars in equipment. Forecast performance directly impacts the safety of launches and influences resource spending. Weather in east central Florida is very difficult to forecast due to subtle influences, requiring constant model verification and forecast improvement. Weather verification monitors prediction performance, identifies and corrects flaws, and results in improved decision making (Fowler et al., 2012).

1.2 Research Motivation

Probabilistic forecasts are reported daily at the 45 WS for rain and lightning events. Outcomes are later recorded based on specific criteria discussed in Chapter 2. Constant verification then follows to ensure maintenance of consistent and appropriate

forecasts. Common weather verification practices involves changing probabilistic predictions to binary predictions based on a fixed threshold cutoff. If truncating the probabilistic forecasts results in improper verification, the 45 WS could produce unrepresentative forecasts, endangering both safety and cost.

Altman and Royston (2006) fault dichotomizing continuous results with reducing statistical power in detecting an outcome. This conversion to dichotomous predictions is common due to an aversion to probabilities because of the difficulty in interpretation and communication of the value's meaning (Altman & Royston, 2006). The conversion increases the number of false positive results, underestimates variation within a group, and ignores linear relationships between variables and the outcome (American Meteorological Society [AMS], 2008). However, assuming all forecasts are deterministic yields simple interpretation but ignores important components of a forecast (Abramson & Clemen, 1995). Additionally, the dichotomous transformation does not account for the degree to which the forecasts differ from the threshold (Zhang & Casey, 2000). Maintaining the probabilistic forecast would provide quantitative information regarding the uncertainty.

1.3 Traditional Contingency Table

Weather forecast predictions provide probabilistic values, which are converted to binary classifications for verification purposes. The models are verified using a Traditional Contingency Table (TCT). Table 1 displays the customary form of a contingency table having two rows (yes/no) as the forecast and two columns (yes/no) as the observed outcome (Agresti, 2002). The four cells are counts of events forecasted to

occur actually having occurred (a), number of events forecasted to occur that actually did not occur (b), number of events forecasted not to occur actually having occurred (c), and the number of events forecasted not to occur that actually did not occur (d).

The performance metrics such as Probability of Detection (POD), Probability of False Alarm (POFA), and frequency bias calculated from the TCT may not capture the weather model's true forecasts and therefore may not be a proper validation tool. POD addresses what fraction of occurring events were correctly forecast. POFA answers what fraction of forecast likely to occur events did not occur. Frequency bias compares the frequency of likely to occur forecasts and the frequency of events which occurred (Fowler et al., 2012). Model corrections based on such metrics may not represent the accurate relationship between the model's probabilistic forecasts and real-world outcomes. Another approach to the TCT may be considered desirable to overcome such weaknesses.

Table 1. Traditional Contingency Table

		<u>Outcome</u>		
		Yes	No	
<u>Forecast</u>	Yes	a	b	a+b
	No	c	d	c+d
		a+c	b+d	a+b+c+d

1.4 Probabilistic Contingency Table

The contingency table's cells do not represent the true probabilistic forecasts; so to possibly improve the verification, this research compares an alternative tool, a Probabilistic Contingency Table (PCT) displayed in Table 2. The row vector $\mathbf{p}$ represents probabilities of an event occurring ($p_1, p_2 \dots p_n$) during sample size $i=1$ to n total days. Row vector $\mathbf{k}$ is the corresponding binary outcomes ($k_1, k_2 \dots k_n$); one representing an event occurred and zero representing an event did not occur. Using the PCT, probabilistic forecasts are maintained and compared to event outcomes.

Table 2. Probabilistic Contingency Table

	<u>Outcome</u>	
	Yes	No
<u>Forecast</u>		
Yes	$E_{11} = \mathbf{p}'\mathbf{k}$	$E_{12} = \mathbf{p}'(\mathbf{1}-\mathbf{k})$
No	$E_{21} = (\mathbf{1}-\mathbf{p})'\mathbf{k}$	$E_{22} = (\mathbf{1}-\mathbf{p})'(\mathbf{1}-\mathbf{k})$
	$E_{11} + E_{21}$	$E_{12} + E_{22}$
	$E_{11} + E_{12} + E_{21} + E_{22}$	

To depict the differences between the TCT and PCT, consider the following example: During a seven-day forecast, an event of rain has probability p_i and outcome k_i displayed in Table 3, where one represents an observed event and zero represents an observed non-event.

Table 3. Example of 7-Day Dichotomous and Probabilistic Forecasts, and Outcomes

Day	PCT Prediction (p_i)	TCT Prediction	Outcome (k_i)	$1-p_i$	$1-k_i$
1	.60	1	1	.40	0
2	.55	1	0	.45	1
3	.70	1	1	.30	0
4	.10	0	0	.90	1
5	.20	0	1	.80	0
6	.30	0	0	.70	1
7	.20	0	1	.80	0

Table 4 displays the TCT and PCT from the seven-day rain forecast. The observed event and non-event totals for both are equal, while the forecast event and non-event total differ from using the 50% cutoff based predictions and the true probabilistic predictions. Further comparison of the tables will be addressed in Chapter 3 after evaluating performance metrics such as POD, POFA, and frequency bias.

Table 4. TCT & PCT Example Calculated from Identical Forecasts and Outcomes

<u>Outcome</u>				<u>Outcome</u>					
		Yes	No	Total			Yes	No	Total
<u>TCT</u>					<u>PCT</u>				
<u>Forecast</u>					<u>Forecast</u>				
Yes	a=	b=	3.0	Yes	E ₁₁ =	E ₁₂ =	2.65		
	2.0	1.0			1.7	0.95			
No	c=	d=	4.0	No	E ₂₁ =	E ₂₂ =	4.35		
	2.0	2.0			2.3	2.05			
Total	4.0	3.0	7.0	Total	4.0	3.0	7.0		

1.5 Methodology

Using uniformly simulated forecasts and observations, contingency tables were first used to compare performance metrics of the same simulated data, utilizing the traditional approach and proposed alternative. The comparison of metrics was first conducted for each individual forecast at varying levels of bias. The forecasts were biased by intentionally simulating 5% to 20% (incremented by 5%) greater or fewer observed events than forecasted likely to occur events. The bias levels were set according to the empirical forecast data. An unbiased forecast occurs with the same number of observed events as forecasted as likely to occur. An over biased forecast occurs when an event occurs less than is predicted by the forecast. An under biased forecast occurs when an event occurs more than is predicted by the forecast. The next step in the simulation involved fitting the forecasts to empirical distributions from data provided by the 45 WS. Similar comparisons were examined in the disparity of metric values calculated from the TCT and PCT using the weighted tables. Following the preliminary simulation results, data obtained from the 45 WS was evaluated by comparing the different verification approaches for rain and lightning events.

1.6 Overview

This research seeks to investigate the usefulness of probabilistic forecasts to produce informative performance metrics and representative forecast verification. The research addresses how the performance metrics calculated using the TCT and proposed alternative compare for varying levels of bias. Chapter 2 explores metrics used to evaluate general classification models, specifically emphasizing weather performance

metrics. The chapter also briefly introduces the proposed alternative method of the PCT used to avoid information loss from a probabilistic threshold cutoff. Chapter 3 explains the process of simulating weather events to calculate performance metrics for the two types of contingency tables. Chapter 3 then addresses the comparison of the TCT and proposed alternative with respect to the 45 WS data. Chapter 4 discusses the results and analysis of the simulation and real-world data, exploring the differences and trends in metric effect sizes by contingency table type. Chapter 5 presents, implications of the results, proposes future research, and final conclusions.

II. Literature Review

2.1 Introduction

Monitoring and maintaining adequate forecast performance requires an accurate and sound verification process. Given this thesis' exploration of dichotomous classification models and methods of performance evaluation, this chapter focuses on characterizing categorical responses. In addition, this chapter describes the tool of contingency tables for weather verification purposes. Based on the possible problems associated with verification of classification models using the TCT, the benefits of the alternative PCT method are discussed.

2.2 Evaluating General Classification Performance

Classification models generally dichotomize continuous responses into a zero or one class output. With the absence of a continuous response, model evaluation often differs from classical regression metrics such as R^2 (coefficient of determination). Instead the comparison of predictions produced from a binary classification model and the observed outcome is usually arranged in a 2 x 2 contingency table, later used to calculate model performance metrics. Terms associated with classification and the table include true positives, true negatives, false positives, and false negatives; Table 5 displays these. True positives occur when the prediction of an event matches the actual occurrence of the event. Conversely, true negatives occur when the predicted lack of an event matches the event not occurring. With classification errors, false positives happen when a predicted

event does not transpire, while false negatives occur when an event takes place when it was not predicted to transpire (Tamilvanan & Bhaskaran, 2017).

Table 5. Contingency Table Containing General Classification Terms, and Metrics

Forecast	<u>Outcome</u>		<u>Metrics</u>
	Yes	No	
Yes	True Positives (TP)	False Positives (FP)	Precision $\frac{TP}{TP + FP}$
No	False Negatives (FN)	True Negatives (TN)	Negative Predictive Value $\frac{TN}{FN + TN}$
Metrics	Sensitivity $\frac{TP}{TP + FN}$	Specificity $\frac{TN}{FP + TN}$	Accuracy $\frac{TP + TN}{TP + FP + FN + TN}$

The four cells of the table are utilized to evaluate the model's classification performance. Metrics commonly reported from the table include: accuracy, precision, negative predictive value, sensitivity, and specificity. Accuracy is the proportion of all predictions correctly classified by the model. Precision, also known as positive predictive value, is the proportion of all positive predictions correctly classified by the model. Precision provides valuable information when the cost of a false positive is high. It addresses the question of all events classified as likely to occur, how many actually occurred? Negative predictive value is the proportion of all negative predictions correctly classified by the model. It addresses the question of all events classified as not likely to

occur, how many actually did not transpire? Sensitivity, also known as recall is the proportion of all occurring events correctly classified as likely to occur by the model. It addresses the question of all occurring events, how many were correctly predicted by the model? Specificity is the proportion of all non-occurring events correctly classified as not likely to occur by the model. It addresses the question of all non-occurring events, how many were correctly predicted by the model? Equations one through five display the calculations for the commonly associated values of general classification models.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Negative Predictive Value = \frac{TN}{FN + TN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{FP + TN} \quad (5)$$

Additional performance evaluation measures using these metrics help balance the weaknesses of the metrics alone. Such measures include the F_1 score, area under the Receiver Operating Characteristic (ROC) curve, area under the Precision-Recall curve (AUPRC), and Youden's J statistic.

The F_1 score (see Equation 6) is a metric useful for imbalanced classes, meaning the event or non-event occurs far more than the other. It is calculated by determining the harmonic mean of precision and sensitivity.

$$F_1 = 2 * \frac{\text{precision} * \text{sensitivity}}{\text{precision} + \text{sensitivity}} \quad (6)$$

The score ranges between zero and one, zero being the worst score and one the best score. The F_1 score punishes extreme values so it tends to zero with either small values of precision or sensitivity. A disadvantage of the F_1 score is the tendency to favor models with similar values for precision and sensitivity. Depending on classification objectives, higher precision or sensitivity may be more important, but the tradeoff between the two metrics results in a score close to zero.

The Precision-Recall curve plots the tradeoff of recall also known as sensitivity on the x-axis against precision on the y-axis, displayed in Figure 1. The figure shows a descending line, losing precision and gaining recall as it decreases. A graph of perfect precision and recall would display two straight lines going across the top and right side.

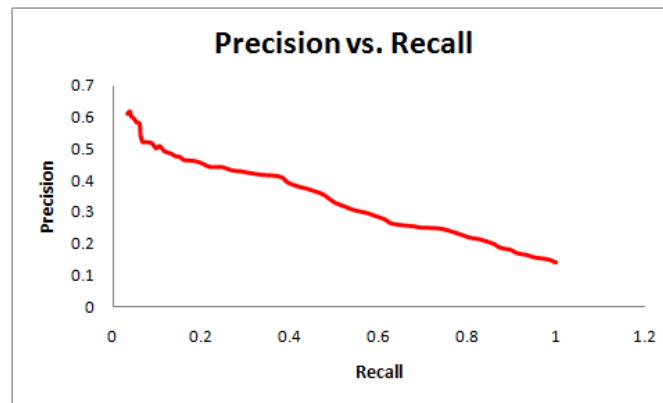


Figure 1. Precision- Recall Curve

A visualization of classification performance depicting the tradeoff of sensitivity and specificity at varying probability classification thresholds can be displayed by the ROC curve. The curve plots the false positive rate on the x-axis, which is also equivalent to 100-specificity against the true positive rate (sensitivity) on the y-axis displayed in Figure 2. The area under the curve (AUC) is one summary measure that quantifies a model's classification performance using the ROC. The AUC ranges from zero to one, with lower numbers indicating poor classification. A value of 0.50 represents random classification and depending on use of the classification model, an AUC of 0.90 indicates the model is excellent at classifying (Ozenne et al., 2015).

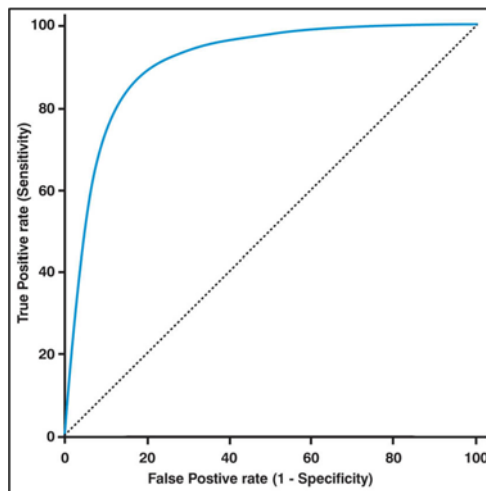


Figure 2. ROC Curve

The ROC curve, however, is insensitive to imbalanced classes and in such cases the Precision-Recall curve better represents performance of a classification model (Saito & Rehmsmeier, 2017). Also in the instance of rarely occurring events, the AUC reports

an overly optimistic classification performance, while the AUPRC reports more representative results (Lobo et al., 2008).

Youden declared the need for a statistic to reduce the four cells displayed in a contingency table to one value to adequately characterize a classification model (Youden, 1950). He did so with the calculation in Equation 7.

$$J = \frac{TP}{TP + FP} + \frac{TN}{TN + FP} - 1 = \textit{sensitivity} + \textit{specificity} - 1 \quad (7)$$

The statistic ranges from zero to one with zero representing a model predicting an event occurrence for the same proportion of non-occurrences and occurrences. Youden describes such a model as obviously worthless, garnering a value of zero. A value of one represents a model producing no false positives or false negatives. The statistic is independent of the absolute sizes of the two classes, useful with imbalanced data classes.

The logarithmic loss function is the last evaluation tool discussed in this section and is utilized to improve probabilistic classification models. The metric is similar to accuracy yet maintains the uncertainty of the probabilities. Equation 8 displays the calculation

$$\textit{logloss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(p_i)) + (1 - y_i) \log(1 - p_i)) \quad (8)$$

where y is the outcome of zero or one, p_i is the predicted probability produced by the model for the i th sample, and N is the total number of samples. To have the greatest classification performance, the logloss function must be minimized. If the outcome is observed ($y=1$), high probabilities minimize the function, and if the outcome is not observed ($y=0$), low probabilities minimize the function. If the logloss value is very high,

the model being used to produce the probabilities should be improved to better represent the real-world outcomes.

2.3 General Uses of Contingency Tables

As described, contingency tables are used for evaluating classification models which serve the purpose of descriptive modeling and predictive modeling (Tan et al., 2018). Descriptive models are a useful explanatory tool for distinguishing objects into classes and predictive models can be useful in predicting class discrimination beyond the given data. Contingency tables describe the interrelationship between any two categorical variables and can be useful in many fields. A variety of current uses include survey research, business intelligence, medical diagnosis, and meteorology.

Survey research compares the expected count gained from the surveys to the actual count. An example of contingency use for survey data is evident in a 2000 American National Election Study in which Census data was used to predict voting on defense spending by region (Burns et al., 2016). The contingency table in this case provides a good tool to evaluate future campaigning needs and trends in region population opinions.

An example of business intelligence use of contingency tables involves a binary classification. An insurance company classifies potential new customers to accept or decline, and the associated observation is whether or not that person was truly low or high risk (Anderson et al., 2010). The bank evaluates whether they made a mistake issuing insurance to a high risk person or declining insurance to a low risk person. A more serious error is issuing insurance to the high risk person and based on the number of

errors represented in the contingency table, banks improve classification models to have the lowest error.

To evaluate medical diagnostic tests, contingency tables are employed to improve models classifying diseases. The two errors made in diagnostic tests include informing a person they have the disease when they do not and informing a person they do not have the disease when they are diseased. The cost of not informing the diseased is much greater as they will not seek treatment, which may be fatal. Based on the findings from the table, diagnostic models aim to improve detection results

The focus of this research is the use of contingency tables in the field of Meteorology, focusing on the forecasts and observations of precipitation and lightning strikes. Depending on the user, the cost of an incorrect classification varies (Jolliffe & Stephenson, 2003). Cancelling a non-refundable event believing the forecast of rain results in a monetary loss for such a user. Another user, believing a no rain forecast, may, for instance, ride a motorcycle that day and get into an accident from the observed rain, injuring the user. Pertaining to the 45 WS, this organization constantly works to improve misclassifications to prevent unnecessary launch cancellations and avoid any unsafe launches. To achieve such a goal, they prioritize iterative weather verification, evaluating performance metrics calculated from contingency tables.

2.4 Evaluating Forecast Performance

Many of the same metrics used to evaluate a general classification model are utilized in weather verification, but are often identified by different names. This research only explores applicable metrics used for weather verification. Commonly calculated

weather metrics include accuracy, POD, probability of false detection (POFD), POFA, and frequency bias. In addition, many measures of skill are used for weather verification, but for the purpose of this research only Threat Score (TS), Kuiper Skill Score (KSS), and Heidke Skill Score (HSS) are discussed.

Deterministic Evaluation.

To calculate any of the metrics, however, first the observation must be defined to represent the forecast event (Fowler et al., 2012). In the cases of rain and lightning forecasts, the 45 WS has existing criteria to identify an event. A lightning event is recorded when a Phase-2 warning is issued for any area of Kennedy Space Center (KSC) or Cape Canaveral Air Force Station (CCAFS). Phase-2 warnings indicate lightning is imminent or occurring within the specified area (Roeder, 2017). A precipitation event is recorded when any one of the gauges reads more than .03 inches or at least three adjacent gauges reports greater than .01 inches (McAleenan, personal communication).

To evaluate the forecast to the defined observation, the cells of Table 6 are utilized for interpretation. The contingency table contains the associated weather verification terminology.

Accuracy is calculated just as it is from the general contingency table (see Equation 9). Once again, accuracy is simple and easy to interpret but does not represent the model's ability to predict rare weather events. The metric inflates models' true performances by only disproportionately capturing the ability to predict common events.

$$Accuracy = \frac{Hit + CN}{Hit + CN + Miss + FA} \quad (9)$$

Table 6. Contingency Table Containing Weather Verification Terms

	Phenomenon Observed	Phenomenon Not Observed	Total
Forecast ≥50%	Hit	False Alarm (FA)	Forecast Yes
Forecast <50%	Miss	Correct Negative (CN)	Forecast No
Total	Observed Yes	Observed No	Total

Probability of Detection (see Equation 10), also known as hit rate or the generalized term of sensitivity, ranges from zero to one, one being a perfect score. POD is sensitive to hits but ignores false alarms. Rare events are captured with POD and with the simultaneous use of POFD, the ROC curve provides a good evaluation tool (Jolliffe & Stephenson, 2003).

$$POD = \frac{Hit}{Hit + Miss} \quad (10)$$

Probability of False Detection (see Equation 11), also known as false alarm rate is not often reported alone but is used in concurrence with POD to produce the ROC curve. POFD is sensitive to false alarms but ignores misses.

$$POFD = \frac{FA}{CN + FA} \quad (11)$$

Probability of False Alarm (see Equation 12), 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. A high POFA is inevitable in forecasting rare events (Olson, 1965).

$$POFA = \frac{FA}{FA + Hit} \quad (12)$$

Frequency Bias (see Equation 13) measures the ratio of the frequency of forecast events to frequency of observed events. The relative frequency does not measure how well the forecasts and observations correspond. The metric does reveal whether the model under forecasts or over forecasts. A perfect score of one indicates no bias, less than one represents under forecasting, and a score over one represents over forecasting (Schwartz, 2016).

$$Bias = \frac{Hit + FA}{Hit + Miss} \quad (13)$$

The ROC curve is conditioned on the observations, while reliability diagrams are conditioned on forecasts, making the pair of visuals very informative. A reliability diagram is also needed to complement the ROC curve as the curve is not sensitive to bias (Brown et al., 2015). False alarm rate is graphed against hit rate for the ROC curve.

Additional skill scores, measuring the predictive ability against a reference forecast, are necessary to evaluate model performance. The scores addressed in this research include: TS, KSS, and HSS.

The Threat Score (see Equation 14), also known as Critical Success Index, does not consider the correct negatives and answers how well the forecasted likely to occur events correspond to the observed events. It ranges from zero to one, 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 & Stephenson, 2003).

$$TS = \frac{Hit}{Hit + Miss + FA} \quad (14)$$

Kuiper Skill Score (see Equation 15) 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 [WMO], 2014).

$$KSS = \frac{(Hit * CN) - (Miss * FA)}{(Hit + Miss)(FA + CN)} \quad (15)$$

Heidke Skill Score (see Equation 16) 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). No single metric provides enough information to make proper adjustments; therefore all of the discussed weather metrics should be addressed collectively for forecast adjustments.

$$HSS = \frac{2 * (Hit * CN) - (Miss * FA)}{(Hit + Miss)(Miss + CN) + (Hit + FA)(FA + CN)} \quad (16)$$

Probabilistic Evaluation.

Statistical forecasting can be classified as objective or subjective forecasting.

Objective forecasts are produced by automatic means while subjective forecasts integrate and interpret information from the objective forecast. Subjective forecasts depend on human judgement of the objective data such as deterministic forecast information, dynamic integrations, and guidance from Model Output Statistics (MOS). Forecasters also use atmospheric observations such as surface maps and radar images, in addition to prior information on persistence, climatology, and previous experiences. A good subjective forecast conveys the measure of a forecaster's uncertainty which can be most accurately captured with in probability terms. (Wilks, 2011).

There are two possible ways to interpret a probabilistic forecast; using a Frequentist approach or a Bayesian approach. The Frequentist approach considers the long-run relative frequency. A 30% forecast means the event should occur 30% of the time for that given forecast bin. The Bayesian approach uses a subjective interpretation to convey the probability as a degree of belief regarding the uncertain event. Such beliefs are founded on previous knowledge and experiences. Bayesian interpretation is heavily utilized for the subjective forecasts.

The Brier Score (BS) (See Equation 18) is a commonly used metric for probabilistic forecasts of dichotomous events. It measures probabilistic prediction accuracy with a mean squared probability error. The score ranges from zero to one, zero being perfect and one being poor. BS can be decomposed into reliability (REL), resolution (RES), and uncertainty. Perfect reliability (see Equation 19) exists when predicted forecast probability p_i is equivalent to the observed frequency of the event, $\bar{o}_i$.

Resolution (see Equation 20) is the ability of the forecast to distinguish different situations. The measure determines the distance between observed relative frequency and sample climatological base rate (see Equation 17).

$$\text{Sample Climatology} = \frac{\text{Hit} + \text{Miss}}{\text{Hit} + \text{FA} + \text{Miss} + \text{CN}} \quad (17)$$

Uncertainty (see Equation 21) measures the variability in observations, indicating the difficulty in which situations can be climatologically predicted. Uncertainty cannot be influenced by anything the forecaster can do and is much higher when the sample climatological probability is close to 0.5 (Wilks, 2011). The BS score decomposes as follows:

$$BS = \frac{1}{N} \sum_{k=1}^N (p_k - o_k)^2 = REL - RES + \text{Uncertainty} \quad (18)$$

$$REL = \frac{1}{N} \sum_{i=1}^I n_i (p_i - \bar{o}_i)^2 \quad (19)$$

$$RES = \frac{1}{N} \sum_{i=1}^I n_i (\bar{o}_i - \bar{o})^2 \quad (20)$$

$$\text{Uncertainty} = \bar{o}(1 - \bar{o}) \quad (21)$$

with N equal to the total number of forecasts, k indexing the forecast-event pairing, n_i the number of forecasts with the same probability category, I the number of unique

forecasts, $\bar{o}_i$ the observed frequency given the forecast probability p_i , and $\bar{o}$, the climatological base rate (Wilks, 2011).

To provide a more informative representation of forecast performance, reliability diagrams display a greater diagnosis of strengths and weaknesses of the forecasts. Reliability measures the agreement between the probability forecasts and the average observed frequency of events, employing the Frequentist approach. In Figure 3 perfect reliability of lightning forecasts is plotted with the diagonal line. The deviation from the diagonal line indicates the conditional bias. Curves below the line represent over forecasting, and above represents under forecasting. A flatter curve means less resolution. The climatology lines are horizontal because climatology forecasts have no resolution as they do not discriminate between events and non-events. The points between the no skill line and diagonal line contribute positively to the BS (Ferro & Fricker, 2012). Reliability plays an important role in subjective forecast verification as it can indicate bias from particular forecasters. No single verification metric dictates forecast adjustments, but reliability diagrams do provide obvious visuals of weaknesses in the probabilistic forecasts. Chapter 4 discusses the importance of maintaining the probabilistic forecasts in relation to reliability and bias.

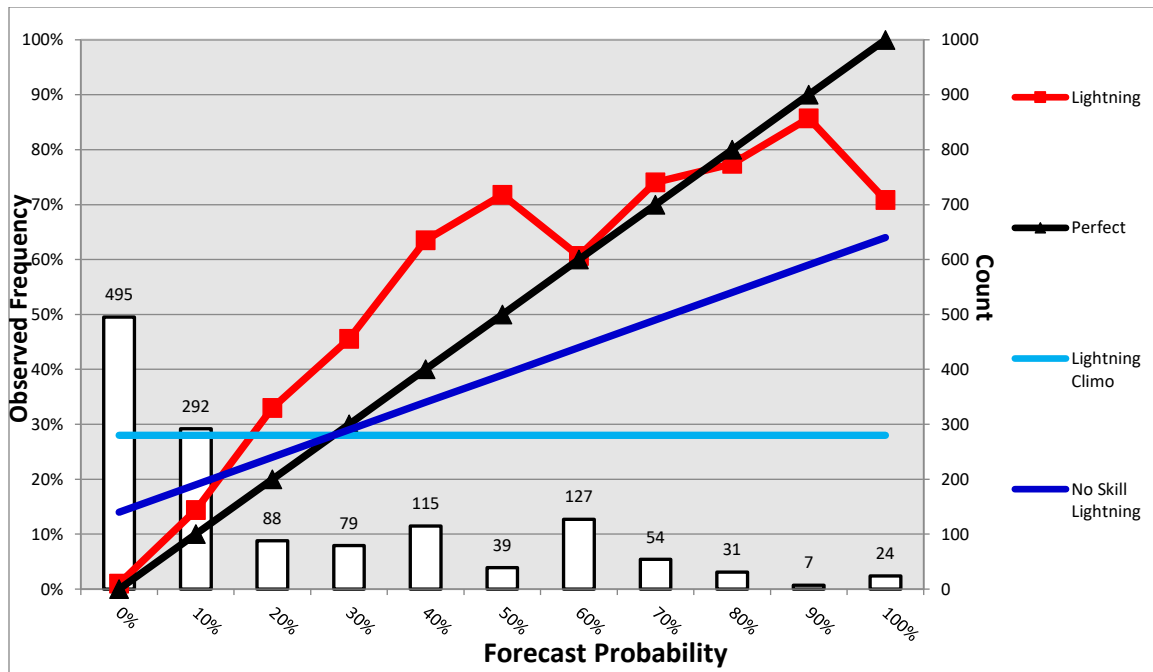


Figure 3. Reliability Diagram Modified from the 45 WS for Lightning Predictions and Observations

Another aspect of forecasts evaluated is the sharpness, an attribute of only the forecasts and not the observations. Sharp forecasts favor the extremes, meaning only forecasts of 0% and 100% would have perfect sharpness. Forecasters must balance all aspects of the forecasts, as perfectly sharp predictions could yield inaccurate and unreliable forecasts. Figure 4 references varying degrees of forecast sharpness. Dichotomizing probabilistic forecasts would result in perfect sharpness, but not represent the true uncertainty in the forecasts. This research expands on the differences and weaknesses of verifying forecasts such that all forecasts were made with perfect sharpness.

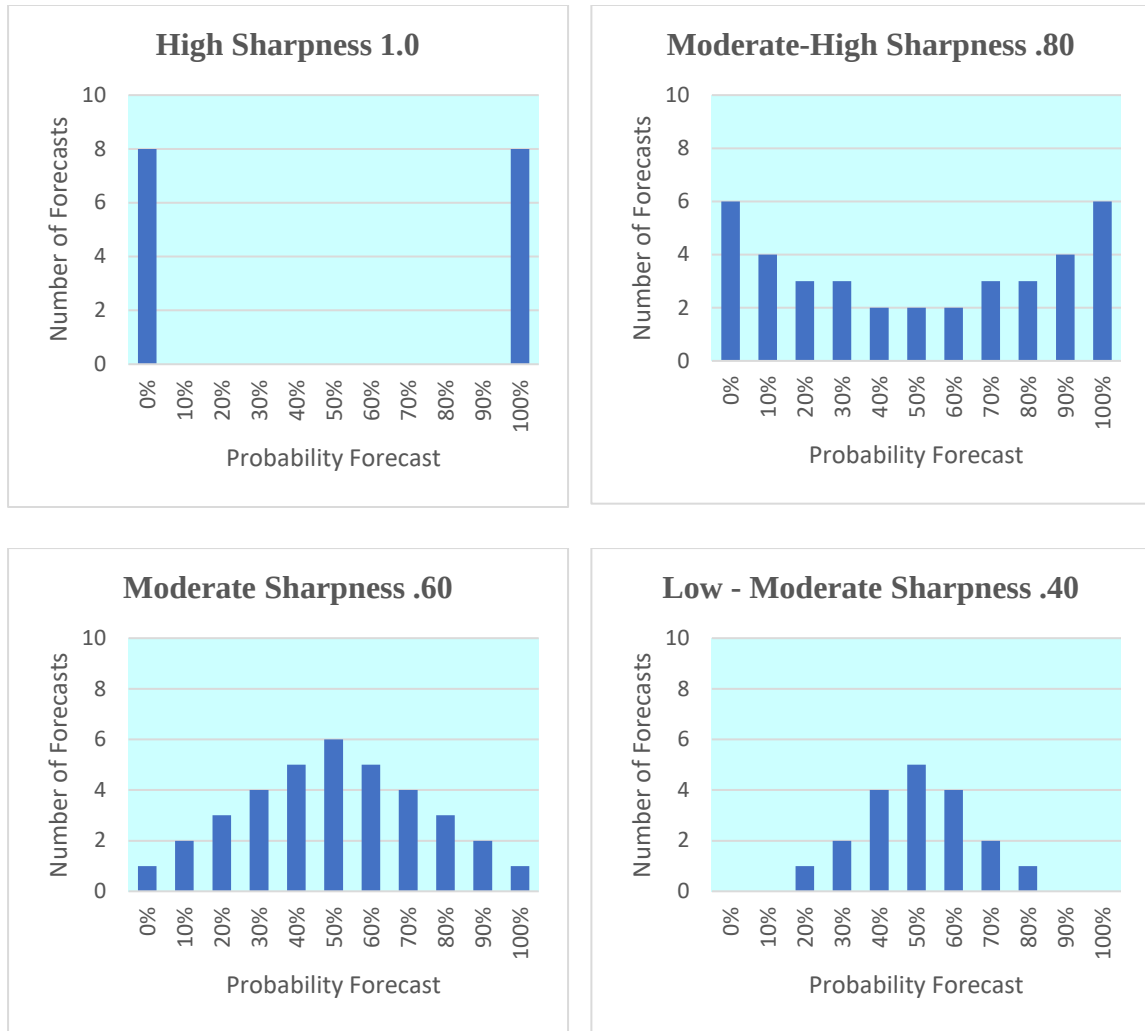


Figure 4. Examples of Varying Forecast Sharpness

2.5 Probabilistic Contingency Table

While the contingency table is a relatively straight-forward tool for evaluating dichotomous model performance, the current practices of dichotomizing the probabilistic forecasts for verification causes a great loss of information. Forecasters are unable or unwilling to use the probabilistic forecasts and default to the yes or no predictions, losing the quantifiable uncertainty associated with probabilities (Mason, 1979; Zhang & Casey,

2000). The conversion to deterministic forecasts is considered unfortunate information degradation and to the detriment of the forecast users (Wilks, 2011). The proper probability threshold is dependent on the users, but is decided by forecasters often arbitrarily. Extensive literature exists encouraging against dichotomizing, so much that the editors of *Medical Decision Making* have enforced a policy limiting the practice with submissions to the journal (Dawson & Weiss, 2012).

All other fields suffer the same consequences of ignoring the information provided in a probabilistic forecast, but still tend to the binary yes or no forecasts. Binary classification models are simple and intuitive to evaluate using a contingency table, explaining the tendency to dichotomize, but an alternative tool of the PCT maintains the simplicity and probabilistic information. One example of the PCT in circulation is evident in (Vêncio & Shmulevich, 2017). The next chapter lays the foundation of exploring if maintaining the probabilistic forecast with a PCT produces different results than the TCT by evaluating the performance metrics associated with the tables.

III. Methodology

3.1 Chapter Overview

This chapter provides an understanding of the techniques used to compare the resulting metrics from the TCT and PCT using simulated as well as real-world forecasts and observations. The comparison of the contingency tables was done in three main steps. The first section explains simulating the forecasts and observations with varying bias. The individual forecasts were used to calculate and evaluate the differences in the TCT and PCT. This process assumed that each forecast occurred an equal number of times. The second section addresses seasonal trends in the data provided by the 45 WS and the empirical distributions of the forecasts during the seasons. Using the distributions, the number of simulated observations for each forecast varied according to the appropriate weight. The final section details the method comparing TCT and PCT using forecasts and observations for four years of rain and lightning events reported by the 45 WS.

3.2 Simulation Assuming Equal Forecasts

To establish fidelity in the comparison of the TCT and PCT with respect to real data, simulated weather observations were first used to establish a controlled evaluation. Using R programming language and the RStudio integrated development environment, observations were generated from forecasts ranging from 0% to 100%, incremented by 10%. Reliable, (i.e., unbiased forecasts) were first considered followed by over biasing

and under biasing. The biases varied by 5% increments from 5% to 20% to reflect realistic representation from 45 WS.

To produce the observations for the likely to occur forecasts we used a standard uniform distribution, $U(0, 1)$, to generate unbiased observations based on the thresholds displayed in Table 7. To produce the observations for the not likely to occur forecasts, we used the thresholds in Table 8. Thirty replications of the 10,000 observations were generated for each forecast level (0%, 10%...100%) at each forecast prediction (likely to occur, not likely to occur).

Table 7. Unbiased Thresholds for Likely to Occur Binned Forecast Simulations

Forecast	<i>Event Occurs if $U \leq$</i>
0%	0
10%	0.1
20%	0.2
30%	0.3
40%	0.4
50%	0.5
60%	0.6
70%	0.7
80%	0.8
90%	0.9
100%	1.0

Table 8. Unbiased Thresholds for Not Likely to Occur Binned Forecast Simulations

Forecast	<i>Event Occurs if $U \leq$</i>
0%	1.0
10%	0.9
20%	0.8
30%	0.7
40%	0.6
50%	0.5
60%	0.4
70%	0.3
80%	0.2
90%	0.1
100%	0

To calculate the values in each TCT, we maintained the standard that any forecast less than 50% was predicted as not likely to occur. Contingency tables were created for each forecast level and replication. For the TCT, hits were recorded as either zero for all forecasts under 50% or as the number of observations from the likely to occur for forecasts greater than or equal to 50%. False alarms were also zero for all forecast 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 all observations produced from 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%. Table 9 depicts an example of the simulated values and associated probabilities generated from unbiased forecasts used to produce forecast specific TCTs and PCTs.

Table 9 contains the forecast in the first column, results from the likely to occur simulations in the second column, results from the not likely to occur simulations in the third column, the associated binary scalars for each likely to occur forecast in column four, the associated binary scalars for each not likely to occur forecast in column five, the probabilities associated with the likely to occur forecast in column six, and the probabilities associated with the not likely to occur forecast in column seven.

Table 10 displays the TCT column vectors calculated for one replication of the simulation representing the number of hits, false alarms, misses, and correct negatives. Hits were calculated with the Hadamard multiplication of the likely to occur vector and the accompanying binary vector $\mathbf{V}_{yes}$. False alarms were determined by subtracting the vector of hits from the sample size 10,000 and calculating the Hadamard product of the resulting vector and the accompanying binary vector $\mathbf{V}_{yes}$. Correct negatives were calculated with the Hadamard multiplication of the not likely to occur vector and the accompanying binary vector $\mathbf{V}_{no}$. Misses were determined by subtracting the vector of correct negatives from 10,000 and calculating the Hadamard product of the resulting vector and the accompanying binary vector $\mathbf{V}_{no}$.

Table 9. Unbiased Binned Forecast Simulation: Events Observed for Likely to Occur Simulation, Non-events Observed for Not Likely to Occur Simulation, and Associated Dichotomous and Probabilistic Event/Non-event Forecasts

Forecast	Likely to Occur	Not Likely to Occur	V _{yes}	V _{no}	p	1-p
0%	0	10000	0	1.0	0	1.0
10%	1049	8941	0	1.0	0.1	0.9
20%	2020	7970	0	1.0	0.2	0.8
30%	2971	7019	0	1.0	0.3	0.7
40%	4042	6025	0	1.0	0.4	0.6
50%	4958	4922	1.0	0	0.5	0.5
60%	5992	4018	1.0	0	0.6	0.4
70%	7042	2932	1.0	0	0.7	0.3
80%	7978	2082	1.0	0	0.8	0.2
90%	8966	989	1.0	0	0.9	0.1
100%	10000	0	1.0	0	1.0	0

Table 10. Unbiased TCT Hits, False Alarms, Misses, and Correct Negatives by Ascending Forecast Bin

<i>Hits= (Likely to Occur)°(V_{yes})</i>	<i>False Alarms= (10,000-Hits)°(V_{yes})</i>	<i>Misses=(Not Likely to Occur)°(V_{no})</i>	<i>Correct Negatives= (10,000-Misses)°(V_{no})</i>
0	0	0	10000
0	0	1059	8941
0	0	2030	7970
0	0	2981	7019
0	0	3975	6025
4958	5042	0	0
5992	4008	0	0
7042	2958	0	0
7978	2022	0	0
8966	1034	0	0
10000	0	0	0

From each replication for the unbiased forecasts, eleven TCTs and PCTs were produced. Table 11 and Table 13 display them respectively. This process was repeated

thirty times to capture the variability in the simulations. The example of each TCT was derived from the equivalent rows of each vector in Table 10.

Table 11. TCTs Produced From Unbiased Binned Forecast Simulations

0%	10%	20%	30%																
<table><tr><td>0</td><td>0</td></tr><tr><td>0</td><td>10000</td></tr></table>	0	0	0	10000	<table><tr><td>0</td><td>0</td></tr><tr><td>1059</td><td>8941</td></tr></table>	0	0	1059	8941	<table><tr><td>0</td><td>0</td></tr><tr><td>2030</td><td>7970</td></tr></table>	0	0	2030	7970	<table><tr><td>0</td><td>0</td></tr><tr><td>2891</td><td>7019</td></tr></table>	0	0	2891	7019
0	0																		
0	10000																		
0	0																		
1059	8941																		
0	0																		
2030	7970																		
0	0																		
2891	7019																		
40%	50%	60%	70%																
<table><tr><td>0</td><td>0</td></tr><tr><td>3975</td><td>6025</td></tr></table>	0	0	3975	6025	<table><tr><td>4958</td><td>5042</td></tr><tr><td>0</td><td>0</td></tr></table>	4958	5042	0	0	<table><tr><td>5992</td><td>4008</td></tr><tr><td>0</td><td>0</td></tr></table>	5992	4008	0	0	<table><tr><td>7042</td><td>2958</td></tr><tr><td>0</td><td>0</td></tr></table>	7042	2958	0	0
0	0																		
3975	6025																		
4958	5042																		
0	0																		
5992	4008																		
0	0																		
7042	2958																		
0	0																		
80%	90%	100%																	
<table><tr><td>7978</td><td>2022</td></tr><tr><td>0</td><td>0</td></tr></table>	7978	2022	0	0	<table><tr><td>8966</td><td>1034</td></tr><tr><td>0</td><td>0</td></tr></table>	8966	1034	0	0	<table><tr><td>10000</td><td>0</td></tr><tr><td>0</td><td>0</td></tr></table>	10000	0	0	0					
7978	2022																		
0	0																		
8966	1034																		
0	0																		
10000	0																		
0	0																		

Table 12 displays the PCT column vectors calculated for one replication of the simulation representing the number of hits, false alarms, misses, and correct negatives. Hits were calculated with the Hadamard multiplication of the likely to occur vector and the accompanying probability vector, $\mathbf{p}$. False alarms were determined by subtracting the vector of hits from 10,000 and calculating the Hadamard product of the resulting vector and the accompanying probability vector, $\mathbf{p}$. Correct negatives were calculated with the Hadamard multiplication of the not likely to occur vector and the accompanying probability vector, $\mathbf{1-p}$. Misses were determined by subtracting the vector of correct negatives from 10,000 and calculating the Hadamard product of the resulting vector and the accompanying probability vector, $\mathbf{1-p}$.

Table 12. Unbiased TCT Hits, False Alarms, Misses, and Correct Negatives by Ascending Forecast Bin

Hits= (Likely to Occur)°(p)	False Alarms= (10,000-Likely to Occur)°(p)	Misses= (Not Likely to Occur)°(1-p)	Correct Negatives= (10,000-Not Likely to Occur)°(1-p)
0	0	0	10000
97.7	902.3	945	8055
402.2	1597.8	1623.2	6376.8
879.9	2120.1	2083.9	4916.1
1592.4	2407.6	2428.2	3571.8
2498	2502	2491.5	2508.5
3583.8	2416.2	2408	1592
4932.2	2067.8	2101.5	898.5
6411.2	1588.8	1600.2	399.8
8073	927	901.2	98.8
10000	0	0	0

The four cells of the PCT sum to 10,000, equivalent to the TCT sum, but hits and false alarms are not restricted to zero for forecasts less than 50%, and misses and correct negatives are not restricted to zero for forecasts of 50% or greater. The example of each PCT in Table 13 was derived from the equivalent rows of each vector in Table 12.

Table 13. PCTs Produced From Unbiased Binned Forecast Simulations

0%	10%	20%	30%
0	97.7	402.2	879.9
0	945	1623.2	2083.9
0	8055	6376.8	4916.1
40%	50%	60%	70%
1592.4	2498	3583.8	4932.2
2428.2	2491.5	2408	2101.5
3571.8	2508.5	1592	898.5
80%	90%	100%	
6411.2	8073	10000	
1600.2	901.2	0	
399.8	98.8	0	

The seven 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. Lastly, 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 forecasts follow the same process of simulations, but the thresholds vary for the likely to occur forecasts and not likely to occur forecasts. At each level of bias, thirty replications were 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%. Table 14 displays the thresholds for likely to occur forecasts in the left of the column and not likely to occur forecasts on the right of the column for each over-forecast.

Table 14. Over-forecast Thresholds for Likely to Occur & Not Likely to Occur Binned Forecast Simulations

Bias	5%		10%		15%		20%	
Forecast	<i>Event Occurs if $U \leq$</i>		<i>Event Occurs if $U \leq$</i>		<i>Event Occurs if $U \leq$</i>		<i>Event Occurs if $U \leq$</i>	
0%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
10%	0.05	0.95	0	1.0	N/A	N/A	N/A	N/A
20%	0.15	0.85	0.1	0.9	0.05	0.95	0	1.0
30%	0.25	0.75	0.2	0.8	0.15	0.85	0.1	0.9
40%	0.35	0.65	0.3	0.7	0.25	0.75	0.2	0.8
50%	0.45	0.55	0.4	0.6	0.35	0.65	0.3	0.7
60%	0.55	0.45	0.5	0.5	0.45	0.55	0.4	0.6
70%	0.65	0.35	0.6	0.4	0.55	0.45	0.5	0.5
80%	0.75	0.25	0.7	0.3	0.65	0.35	0.6	0.4
90%	0.85	0.15	0.8	0.2	0.75	0.25	0.7	0.3
100%	0.95	0.05	0.9	0.1	0.85	0.15	0.8	0.2

Table 15 displays the thresholds for likely to occur forecasts in the left of the column and not likely to occur forecasts on the right of the column for each under-forecast. The cells labelled not applicable (N/A) are restrained by probability ranging from 0% to 100%.

Table 15. Under-forecast Thresholds for Likely to Occur & Not Likely to Occur Binned Forecast Simulations

Bias	5%		10%		15%		20%	
Forecast	<i>Event Occurs if $U \leq$</i>		<i>Event Occurs if $U \leq$</i>		<i>Event Occurs if $U \leq$</i>		<i>Event Occurs if $U \leq$</i>	
0%	0.05	0.95	0.1	0.9	0.15	0.85	0.2	0.8
10%	0.15	0.85	0.2	0.8	0.25	0.75	0.3	0.7
20%	0.25	0.75	0.3	0.7	0.35	0.65	0.4	0.6
30%	0.35	0.65	0.4	0.6	0.45	0.55	0.5	0.5
40%	0.45	0.55	0.5	0.5	0.55	0.45	0.6	0.4
50%	0.55	0.45	0.6	0.4	0.65	0.35	0.7	0.3
60%	0.65	0.35	0.7	0.3	0.75	0.25	0.8	0.2
70%	0.75	0.25	0.8	0.2	0.85	0.15	0.9	0.1
80%	0.85	0.15	0.9	0.1	0.95	0.05	1.0	0
90%	0.95	0.05	1.0	0	N/A	N/A	N/A	N/A
100%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Contingency tables were produced for each applicable forecast in the same manner as the unbiased simulations. The seven 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 under forecasting. The comparison of the TCT and PCT simulations for each forecast are discussed in the next chapter.

3.3 Simulation Using 45 WS Empirical Distributions

Many trends were discovered from the first simulations assuming equal number of forecasts, however, it was not realistic to assume only one level of a forecast will be

predicted repeatedly. For example, a 10% forecast of an event will not be made every day of the year. It is also not reasonable to assume all forecast levels are predicted the same number of times during the year. To overcome these assumptions, data from the 45 WS was used to fit representative empirical distributions.

Four years of data from 2015-2018 were provided from the 45 WS in three different Excel workbooks containing many formulations not utilized in this research. The information that was used included the daily date, 24-hour forecast of rain, and 24-hour forecast of lightning. The observation of “yes” or “no” for both the rain and lightning events are addressed in the next section. Looking first at rain forecasts, each forecast was binned by the month an observation occurred. Figure 5 displays the rain forecast distributions by month, suggesting two seasons. Months June through September were defined as the warmer season and months October through May were defined as the colder season. To confirm the seasonal divide a Tukey’s comparison test was conducted to test the equivalency of the mean rain forecasts among the two seasons. Table 18 confirms the mean rain forecasts have a statistically significant difference. Addressing the multiple comparisons of each month (output in Appendix C), at an alpha of 0.05, May and October are not statistically different from either season. This will be addressed in Chapter 5 for future recommendations.

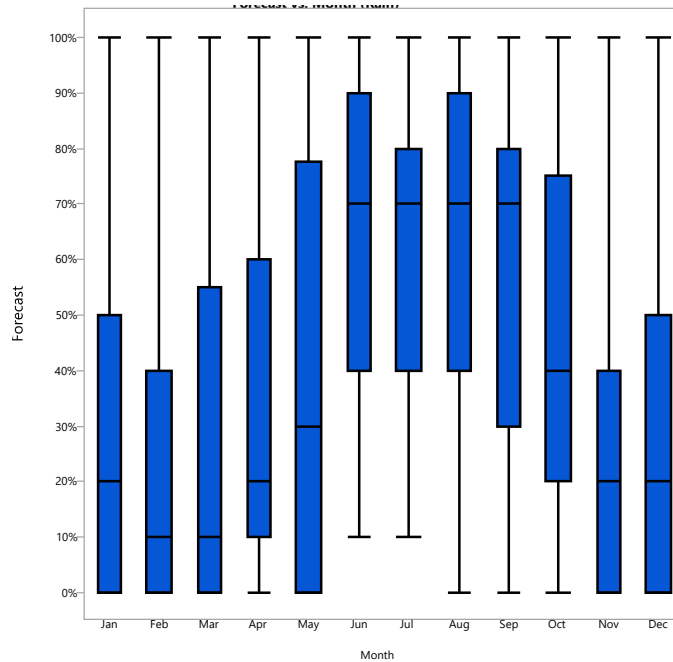


Figure 5. 45 WS (2015-2018) Rain Forecast Distributions by Month

After determining the seasons according to rain forecasts, the data was divided by the respective seasons. The counts of each forecast in the given season are displayed in Tables 16 and 17 and were used to determine the proportion of times a forecast was predicted 24-hours 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. For example, for a 10% over-forecast, 0% is not a possible forecast; therefore 294 forecasts were removed from the total number of rain forecasts in the cold season. The remaining possible proportions were then calculated for forecasts of 10% to 100%.

Table 16. Number of Cold Season Rain Forecasts for 2015-2018

Forecast	Count
0%	294
10%	104
20%	117
30%	95
40%	84
50%	13
60%	68
70%	38
80%	55
90%	24
100%	78

Table 17. Number of Warm Season Rain Forecasts for 2015-2018

Forecast	Count
0%	2
10%	7
20%	26
30%	62
40%	52
50%	12
60%	56
70%	82
80%	71
90%	39
100%	76

Lightning had the same seasonal relationships as found for the rain events for 24-hour forecasts, apparent in Figure 6. Table 18 confirms the seasonal divide with significantly different mean lightning forecasts among the seasons. The multiple month comparisons in Appendix C reveals the mean lightning forecasts for May as not significantly different from either season, but slightly closer to the cold season group mean calculated excluding May. As stated, only the 24-hour forecast was addressed as lightning forecasts for any more days prior to the event had a significantly different distribution. Lightning events were almost never forecasted as likely to occur more than one day prior to the event.

Table 18. Tukey's Comparisons of Lightning and Rain Forecasts by Season

Event	Season	-Season	Difference	Std Error	t Ratio	Prob > t
Lightning	Cold	Warm	-.316495	.0171651	-18.44	<.001
Rain	Cold	Warm	-.408015	.0140548	-29.03	<.001

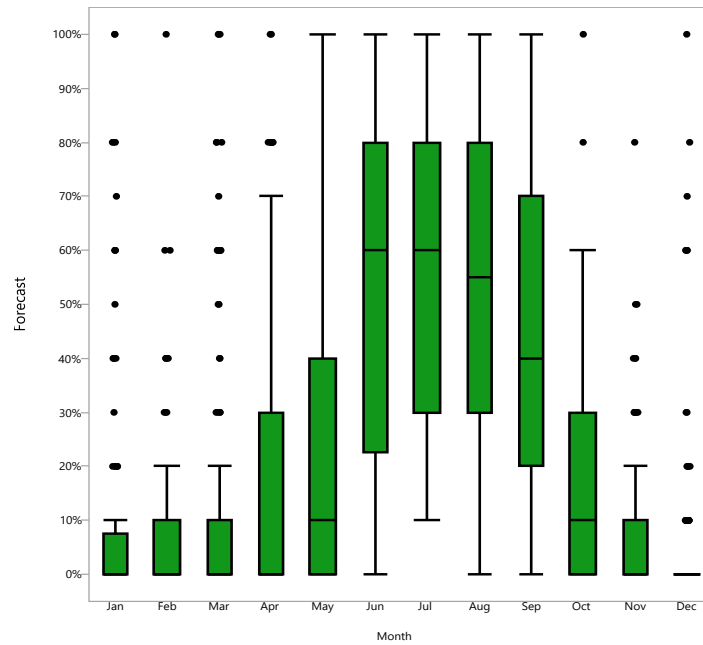


Figure 6. 45 WS (2015-2018) Lightning Forecast Distributions by Month

For both rain and lightning events, forecasts of likely to occur were much higher during the warm season. Separating the data by season was important to determine if the difference in TCT and PCT performance was influenced by predicting more likely to occur events in the warmer season and more not likely to occur events in the colder season.

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.

Tables 19 and 20 display the number of lightning event forecasts for each bin during 2015-2108 cold season and 2015-2018 warm season respectively.

Table 19. Number of Cold Season Lightning Forecasts

Forecast	Count
0%	618
10%	117
20%	58
30%	37
40%	41
50%	9
60%	38
70%	12
80%	22
90%	17
100%	0

Table 20. Number of Warm Season Lightning Forecasts

Forecast	Count
0%	7
10%	50
20%	61
30%	34
40%	59
50%	17
60%	70
70%	56
80%	55
90%	22
100%	54

As an example for a specific forecast, twelve forecasts of 70% were made for lightning during the cold season; therefore 12 was divided by the total count of 969. This quotient was then multiplied by 10,000, resulting in 123.839 simulated 70% forecasts (the number of forecasts were not rounded to maintain the empirical distribution representation). 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. This produced one TCT and one PCT for the unbiased simulation. This was done for all bias levels of over-forecast and under-forecast, 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.

3.4 45 WS Forecasts & Observations

Following the simulations, the data provided by the 45 WS was used to compare the TCT and PCT for both rain events in the colder and warmer seasons, and for lightning events in the colder and warmer seasons. The data divided by seasons was used to determine hits, false alarms, misses, and correct negatives of the TCT and PCT. For the TCT, hits were recorded for an observed “yes” for all 50% or greater forecasts. False alarms were counted for all 50% or greater forecasts observed as “no”. Misses were counted for all forecasts less than 50% observed as “yes”. Correct negatives were determined as all forecasts less than 50% observed as “no”.

For the PCT, the probability, p_i is defined as the forecast and k_i is the binary outcome of one for “yes” observations and zero for “no” observations. Hits were calculated as the dot product of vector, $\mathbf{p}$ and vector, $\mathbf{k}$. False alarms were recorded as the dot product of vector, $\mathbf{p}$ and vector, $\mathbf{1-k}$. Misses were counts as the dot product of vector, $\mathbf{1-p}$ and vector, $\mathbf{k}$. Correct negatives were recorded as the dot product of vector, $\mathbf{1-p}$ and vector, $\mathbf{1-k}$.

After creating TCTs and PCTs for each season and forecasted event, the same seven metrics, POD, POFD, POFA, TS, KSS, HSS, and frequency bias, were calculated. 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.

3.5 Summary

The first step of simulations was conducted with the intent to compare the performance metrics produced from the TCT and PCT at each forecast. It was also used to reveal metric tendencies for varying biases. There were many limitations however in the metric calculations due to the zero values in two of the cells for every TCT. This problem was resolved by summing over the different forecasts, introducing the empirical distributions. Simulations produced controlled results without the consideration of any unknowns. Comparing the data provided by the 45 WS introduced more chances of variability, but the simulations provided general standards and expectations of the real-world data. We next present the results of our analysis.

IV. Analysis and Results

4.1 Chapter Overview

This chapter reveals the trends of the graphs produced from the simulated output as well as from the data provided by the 45 WS. The first section contains the trends for the unweighted simulation for the effect sizes for POD, POFD, POFA, TS, KSS, HSS, and frequency bias for both over biased and under biased forecasts. The frequency bias metric was further examined with forecast by bias level for the TCT and PCT with over-forecasts and under-forecasts. The second section examines the metric effect sizes according to the four empirical distributions from cold and warm season rain and lightning events, separating results by over and under-forecasting. The last section explores the comparison of the TCT and PCT using the 45 WS forecasts and outcomes. The data is divided by the two seasons for the rain and lightning events. The metric comparisons strictly reveal the differences between the TCT and PCT including which metrics are closer to the perfect scores. Both tables are produced from the same data and neither table can be determined as “better” based on the metric values. Section 4.5 addresses BS, frequency bias, and reliability diagrams. The measures are examined with both dichotomized forecasts and probabilistic forecasts to determine the more representative type of table to use for verification.

4.2 Simulation Results and Trends - Unweighted

The effect size for each metric was calculated by subtracting the value produced using the TCT by the value produced using the PCT. The perfect score for each metric is

listed in Table 21. A positive effect size for accuracy, POD, TS, and KSS indicates the PCT has higher values closer to the perfect score, and a positive effect size for POFA and POFD means the TCT has lower values closer to the perfect score. The effect size of the frequency bias metric is not as visually informative. A perfect score for frequency bias is one; so the effect size of the differences from one was calculated. A positive value for this metric indicates the TCT has values closer to the perfect score.

Table 21. Metric Ranges and Associated Perfect Scores

Metric	Range	Perfect Score
Accuracy	0 to 1	1
POD	0 to 1	1
POFD	0 to 1	0
POFA	0 to 1	0
TS	0 to 1	1
KSS	-1 to 1	1
HSS	- ∞ to 1	1
Bias	0 to ∞	1
Absolute Bias	- ∞ to ∞	0
Brier	0 to 1	0

For the unbiased simulation displayed in Figure 7, both the TCT and PCT produce equivalent values of accuracy for a 0% forecast. From the first applicable forecast for the over biased simulations, the TCT has higher accuracy values until the 50% forecast. At 50%, the PCT shows a drastic increase over the TCT, and then decreases in disparity until a 100% forecast. At 100%, the tables produce the same accuracy value. The more biased forecasts have greater effect sizes; meaning for the 20% over bias, the TCT is far higher until the 50% forecast and then the PCT's associated accuracy is higher. The under biased simulations have the opposite effect size pattern in relation to the 50%

forecast. For lower forecasts, values of accuracy are greater utilizing the PCT and for forecasts of 50% and greater, the accuracy metric calculated from the TCT is higher. Equivalent to that of the over-forecasting, the more biased simulations resulted in a higher accuracy effect size.

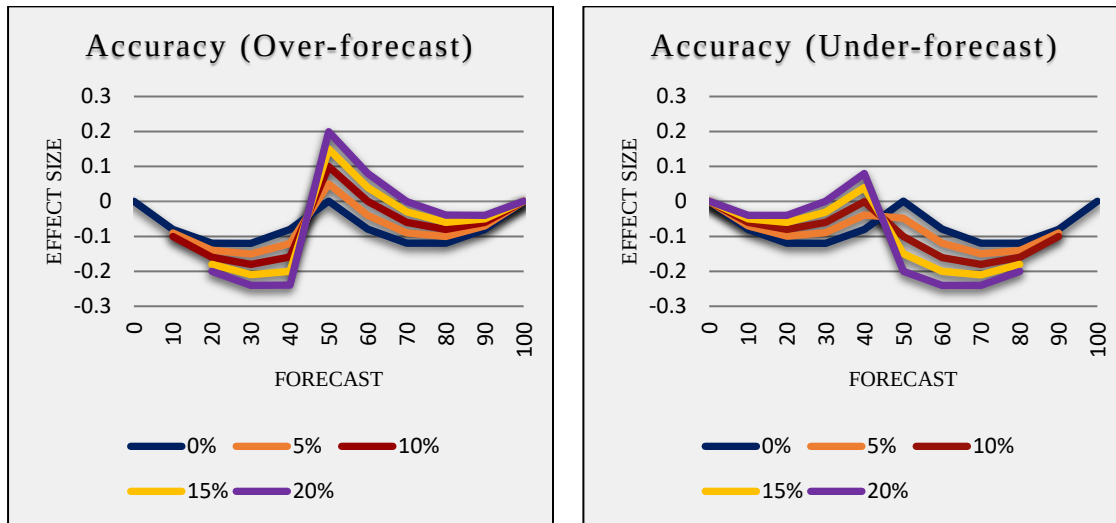


Figure 7. Accuracy Simulation Effect Size

For both the over-forecasts and under-forecasts, the POD is the same for all varying biases. POD produced by the TCT for forecasts under 50% is zero. For forecast over 50%, the TCT POD is one. POD produced by the PCT is the same for the varying biases so the difference for all bias levels is the PCT POD minus zero or one, depending on 50% forecast marker. From the first applicable forecast in the over biased forecasts, POD is higher for the PCT with an increasing disparity until a 50% forecast. The TCT produces a significantly higher POD at the 50% forecast, but the difference shrinks back to zero at 100% forecast. The under biased forecasts produced the same pattern in effect sizes, however all bias levels start at a 0% forecast and stop at the last respective

applicable forecast. The greatest differences in POD produced by the contingency tables occur at the central forecasts, and the smallest disparity occur at the outer extremes. The left extreme has similar undesired small probability and the right extreme has similar desired higher probability. The effect sizes at each forecast bin are graphically depicted in Figure 8.

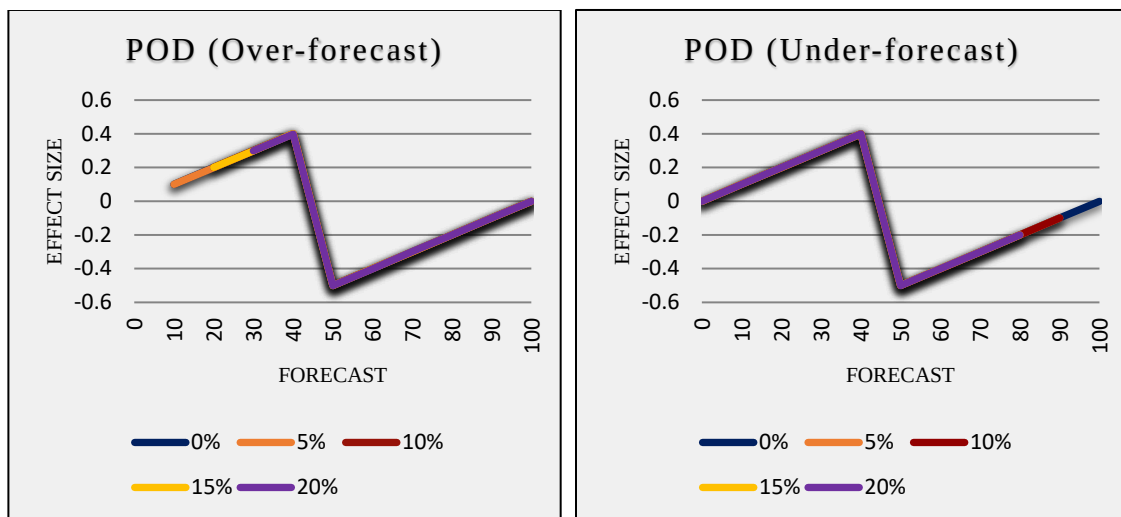


Figure 8. POD Simulation Effect Size

The difference in POFD for the TCT and PCT, shown in Figure 9, has the opposite relationship as POD, but again the effect sizes are equivalent for the varying levels of bias. Zero is the desired value of POFD, so positive effect sizes represent lower values produced by the TCT. For the unbiased forecasts, POFD is equally perfect at 0% forecast. For the over biased forecasts, the TCT has increasingly lower values than the PCT until the 50% forecast, and the PCT has smaller POFD until reaching an equally poor score of one at the 100% forecast. Similar to POD, the greatest differences in the values produced by the contingency tables occurs for the central forecasts. The lower

extreme forecasts have comparable good scores and the higher extreme forecasts have comparable poor scores. Applicable forecasts differed for under biased simulations, starting at 0% for all bias levels and stopping at or prior to the 90% forecast.

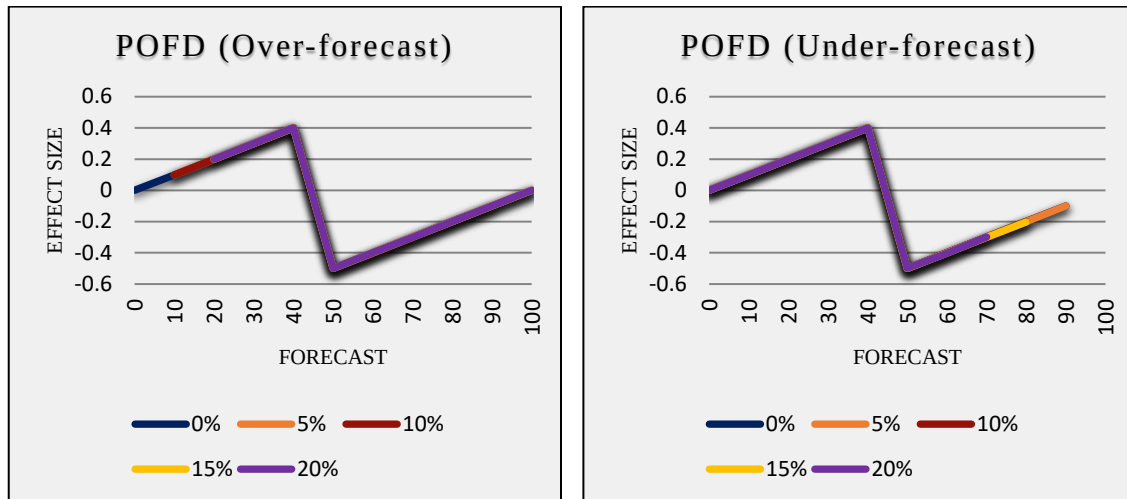


Figure 9. POFD Simulation Effect Size

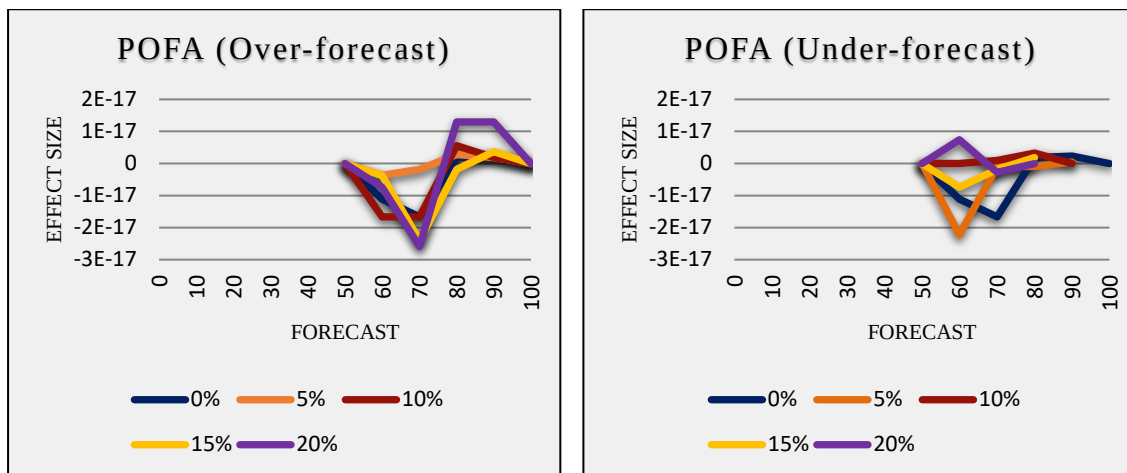


Figure 10. POFA Simulation Effect Size

As the forecast increases, the POFA increases away from the desired probability of zero. The effect sizes between the TCT and PCT for POFA are negligible, apparent by the scale of the y-axis in Figure 10. Effect sizes for forecasts under 50% could not be calculated as the POFA calculation for the TCT results in a division by one and is not feasible for all forecasts under 50%.

Displayed in Figure 11, the TS produced by the TCT and PCT increase as the binned forecasts ascend. The unbiased simulations produce an increasing TS for both the TCT and PCT, achieving a perfect score of one at the 100% forecast. For the over biased forecasts, the PCT has increasing higher scores from the first applicable forecast until the 50% forecast. At 50%, the TCT produces a higher TS value, and the gap decreases until the 100% forecast. The greater over biased simulations have the smallest dissimilarities between the TCT and PCT TS values. The unbiased simulation produce a greater effect size than all levels of over-forecasting. The under biased TS effect sizes start at the forecast of 0% and stop at or prior to 90%. The greater under biased simulations have the largest dissimilarities between the TCT and PCT TS values. The unbiased simulation produce the smallest effect size compared to all levels of under-forecasting. The greatest disparity again occurs for the central forecasts. The lower undesired TS values correspond with lower forecasts, and the higher TS values correspond with the higher forecasts.

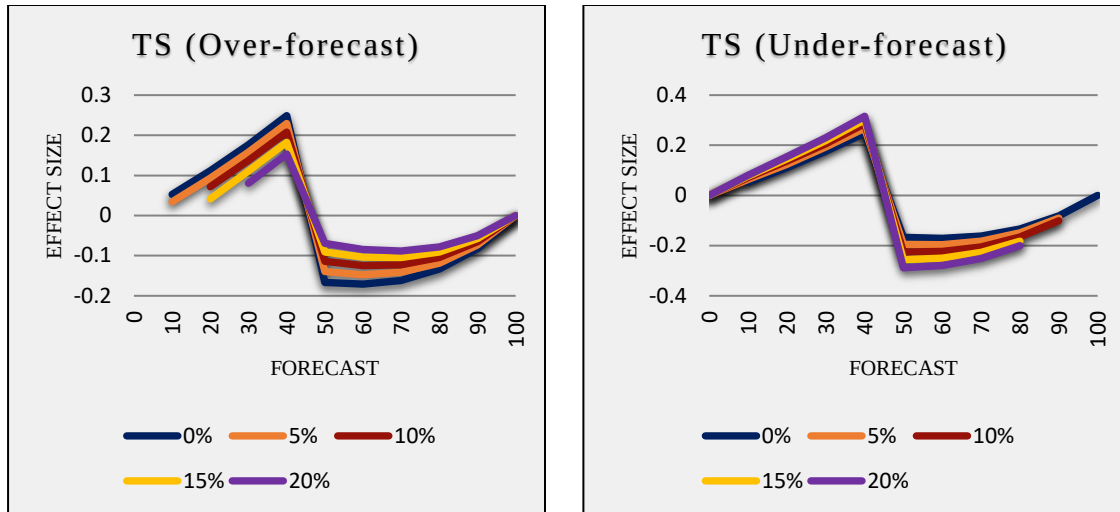


Figure 11. TS Simulation Effect Size

KSS values range from negative one to one, with one being the best score. All of the KSS values produced by the TCT are zero and all KSS values from the PCT are close to zero. The resulting effect size is negligible, shown in Figure 12, only displaying the fluctuation of the PCT KSS values near zero. Like the KSS results, the HSS effect sizes, displayed in Figure 13, for each forecast have negligible differences.

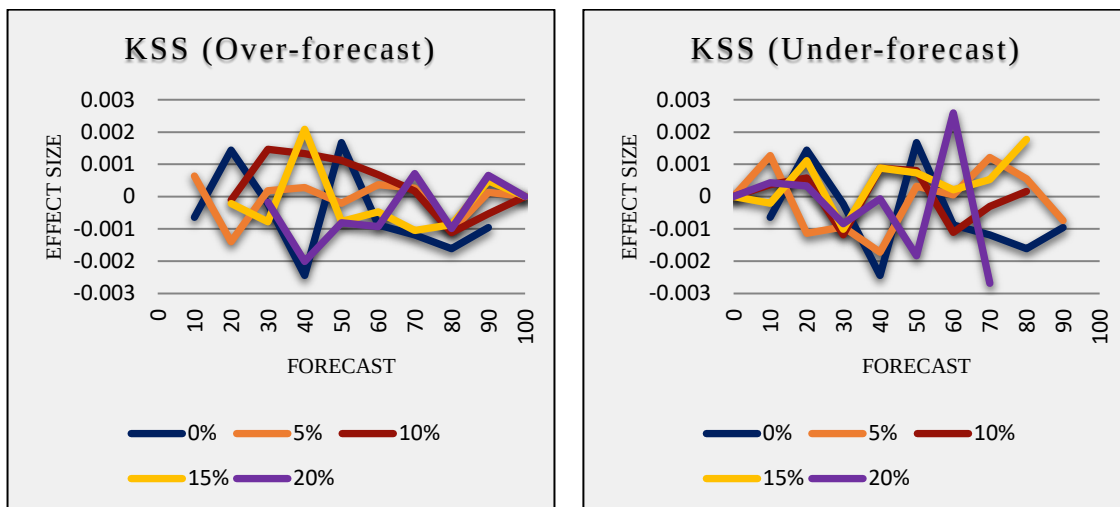


Figure 12. KSS Simulation Effect Size

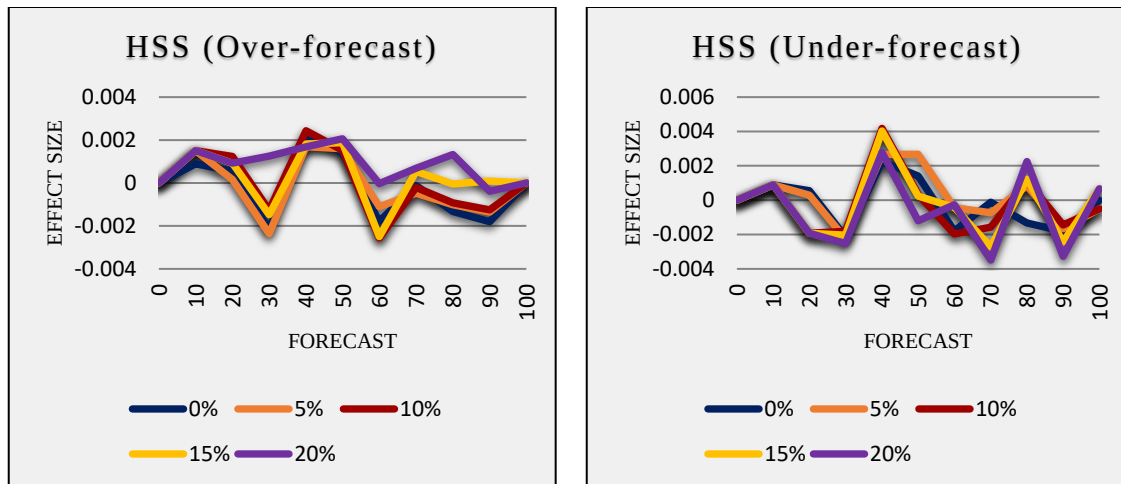


Figure 13. HSS Simulation Effect Size

A frequency bias score of one indicates perfectly unbiased forecasts. The effect sizes considering only the difference in frequency bias are supplemented in this section with the effect sizes of the absolute differences from one. For both the TCT and PCT, graphs of forecasts plotted against frequency bias values for the varying biased simulations can be found in Appendix A. For the unbiased forecast simulations, the TCT produces frequency bias due to the use of a binary cutoff. The PCT produces frequency bias near one for the unbiased forecast simulations. The TCTs for both over-forecasting and under-forecasting produce a frequency bias value of zero until 50% forecast where the bias significantly increases. After the 50% forecast, the frequency bias decreases toward a perfect score of one. The PCTs for over-forecasting produces the highest frequency bias for the first applicable forecast and then decreases toward one as the forecast increases to 100%. The PCTs for under-forecasting have a frequency bias value of zero at 0% forecast and increase toward one until the last applicable forecast. The more biased simulations have a greater distance from one for the PCTs. Considering 50%

and greater forecasts, more biased simulations produce frequency bias values farther from one for the over-forecasted simulations, but values closer to one for the under-forecasted simulations.

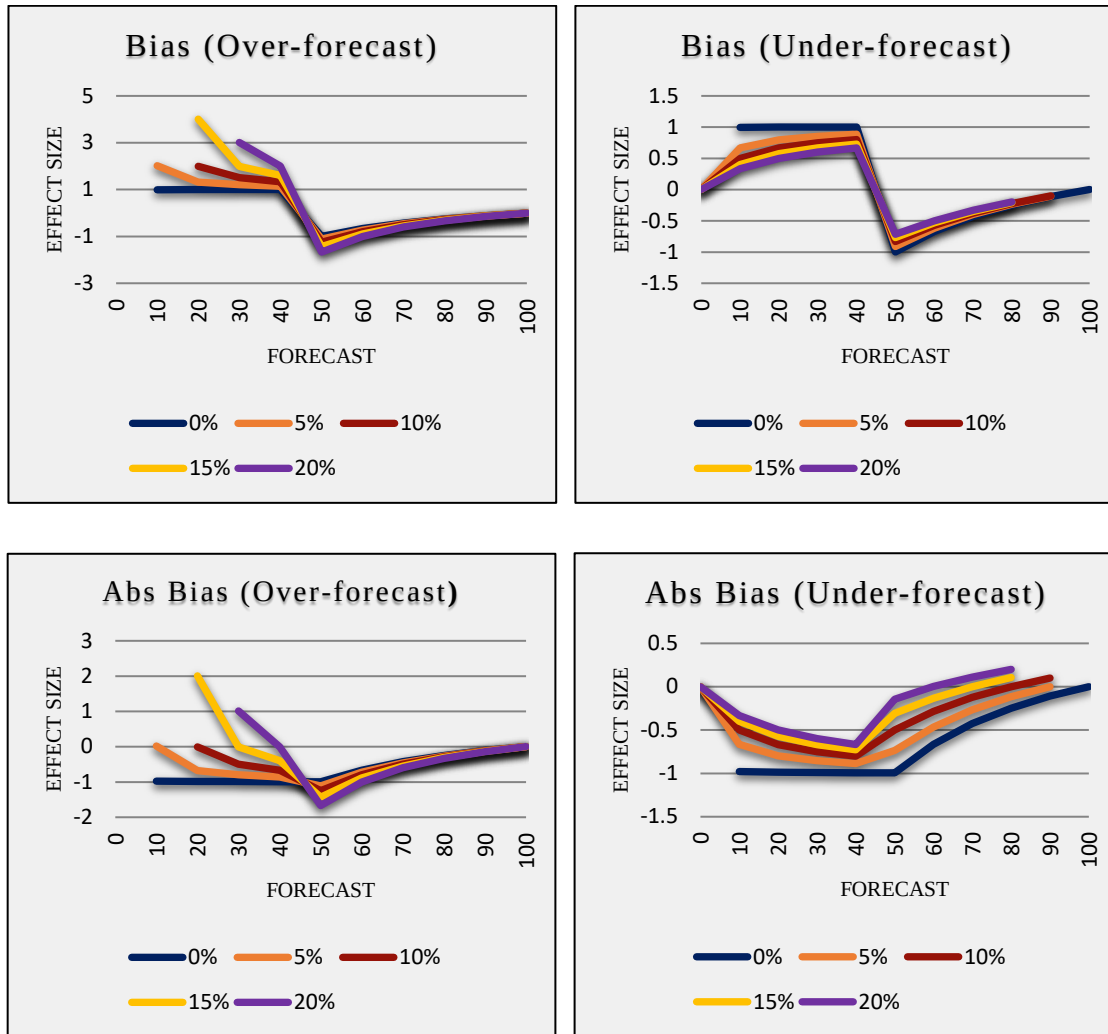


Figure 14. Bias and Absolute Bias Simulation Effect Size

Figure 14 displays the absolute effect sizes, highlighting more biased simulations produce closer values of frequency bias for the under-forecasted TCT and PCT while the less biased simulations produce less disparity for the over-forecasted simulations. Figure 14 also shows the TCT produces frequency bias values closer to one from the first

applicable forecast until 50% forecast for the over-forecasted simulations. PCT produces frequency bias values slightly closer to one for 50% to 90% forecasts. For the under-forecasted simulations, the PCT has values of frequency bias closer to one until the 50% forecast. At 50% and greater, the TCT produces values of frequency bias closer to one for under-forecasting.

4.3 Simulation Results and Trends - Weighted

This section reveals the results and trends from the weighted simulations comparing the metrics calculated using the TCT and PCT. The resulting number of forecasts for each probability bin (sharpness) grouped by season and event type are displayed in Figure 15. The cold season has fewer forecasted likely to occur events and the warm season forecasts have a greater number of likely to occur events. Figures 16, 17, 18, and 19 display the effect sizes of the metrics calculated from the proportionally simulated events. The bias level graphed against metric values for each type of table is displayed in Appendix B.

The cold season distribution of rain forecasts is heavily skewed right with a 0% forecast occurring almost thirty percent of the time. The contingency table metrics which perform better at lower forecasts, therefore have overall stronger metrics. Displayed in Figure 16, the TCT produces values closer to ideal than the PCT in regards to accuracy, POD, POFD, POFA, TS, KSS, and HSS for all forecasts at all levels of bias. The more over biased forecasts, have a greater of a gap in accuracy, POD, TS, KSS, and HSS and less disparity for POFD and POFA. The more under biased forecasts have less of a gap in accuracy, POD, TS, KSS, and HSS and greater disparity for POFA. The differences in

POFD remain consistent as forecasts become more biased. The absolute frequency bias metric is closer to one produced by the PCT for the unbiased and under-forecasted simulations. The TCT has better values of frequency bias for the over-forecasted simulations. Referring back to Figure 14 over-forecasts, the TCT has frequency bias values closer to one for the lower binned forecasts, which is highly favored for the cold season rain distribution of probabilities.

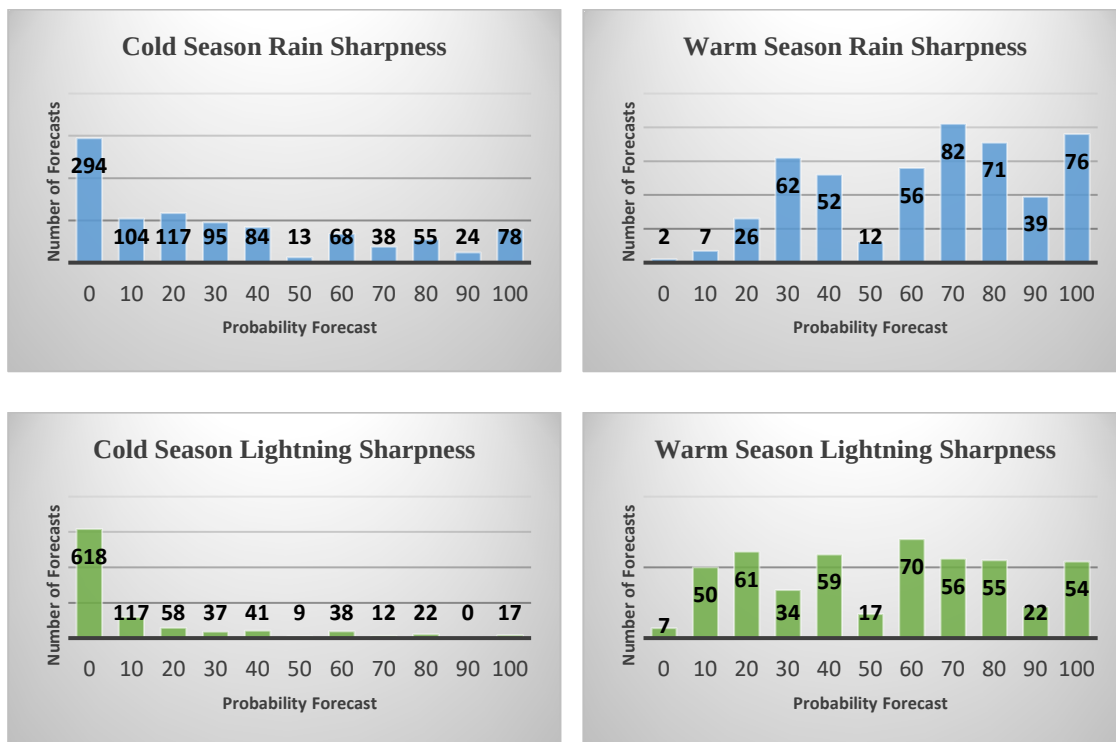


Figure 15. Seasonal Rain and Lightning Distributions

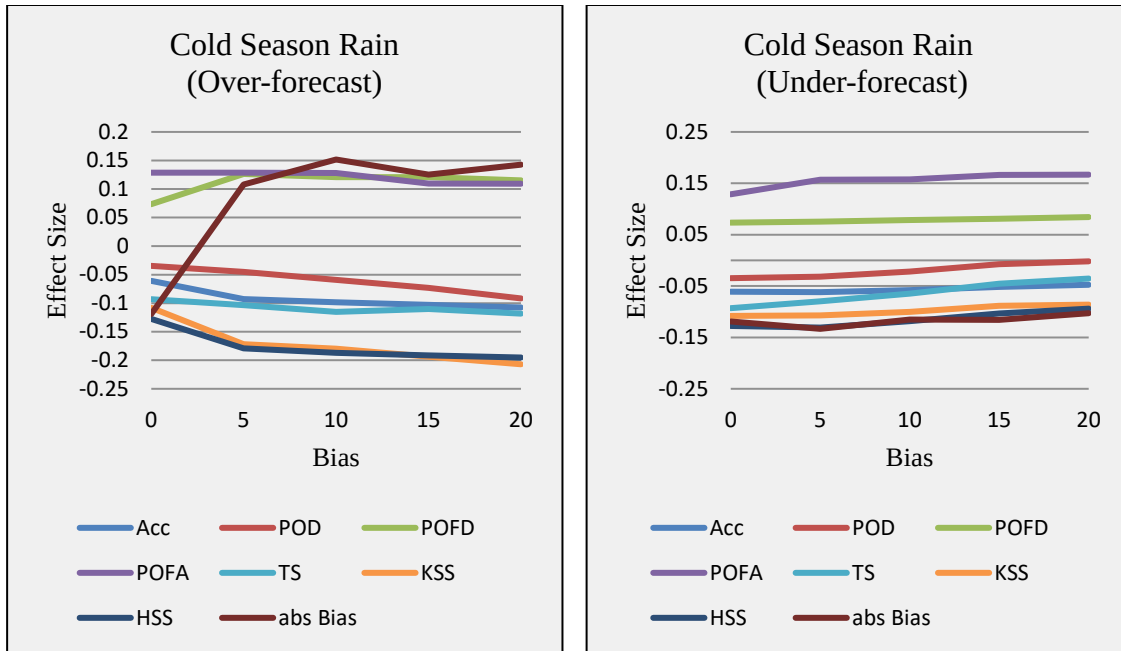


Figure 16. Weighted Cold Season Rain Effect Size

The warm season distribution for rain is somewhat bimodal about 50% with a jump at number of forecasts at 100%. It also contains many more forecasts over 50%. Displayed in Figure 17, the TCT has values closer to the ideal scores in regards to accuracy, POD, POFD, POFA, TS, KSS, and HSS for all varying biased forecasts. The more over biased forecasts have a decreasing gap in values of accuracy, POD, POFD, POFA, TS, KSS, and HSS. The more under biased forecasts have a greater disparity in values of accuracy, POD, POFD, POFA, TS, KSS, and HSS. The PCT had frequency bias closer to one for the unbiased and all over-forecasted simulations. The TCT had better frequency bias for all under-forecasted simulations. Referring back to Figure 14 under-forecasts, the TCT has frequency bias values closer to one for the higher binned forecasts which is highly favored for the warm season rain distribution of probabilities.

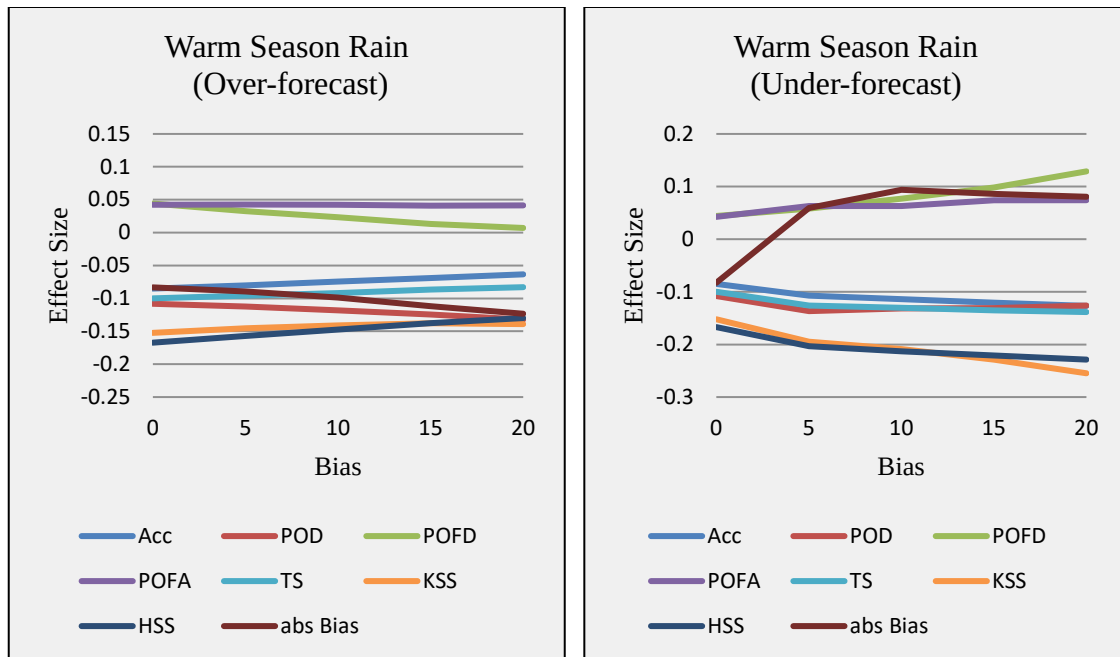


Figure 17. Weighted Warm Season Rain Effect Size

The cold season distribution of lightning forecasts displays the most skewed right distribution with a 0% forecast occurring almost 65% of the time. The lightning cold season distribution is similar to the rain cold season distribution so the comparison of the TCT and PCT, displayed in Figure 18, is comparable with respect to the metrics of accuracy, POD, POFD, POFA, TS, KSS, HSS, and frequency bias.

The warm season distribution of lightning forecasts appear fairly normal, with the exception of very few 50% forecasts. Displayed in Figure 19, accuracy, POD, POFD, POFA, TS, KSS, and HSS have scores closer to the ideal values generated from the TCT for all levels of bias. The more over biased forecast have less disparity in accuracy, POD, POFD, POFA, TS, KSS, and HSS. The more under biased forecasts have a greater gap in accuracy, POD, POFD, POFA, TS, KSS, and HSS. The absolute frequency bias pattern is the same as the results for warm season rain events.

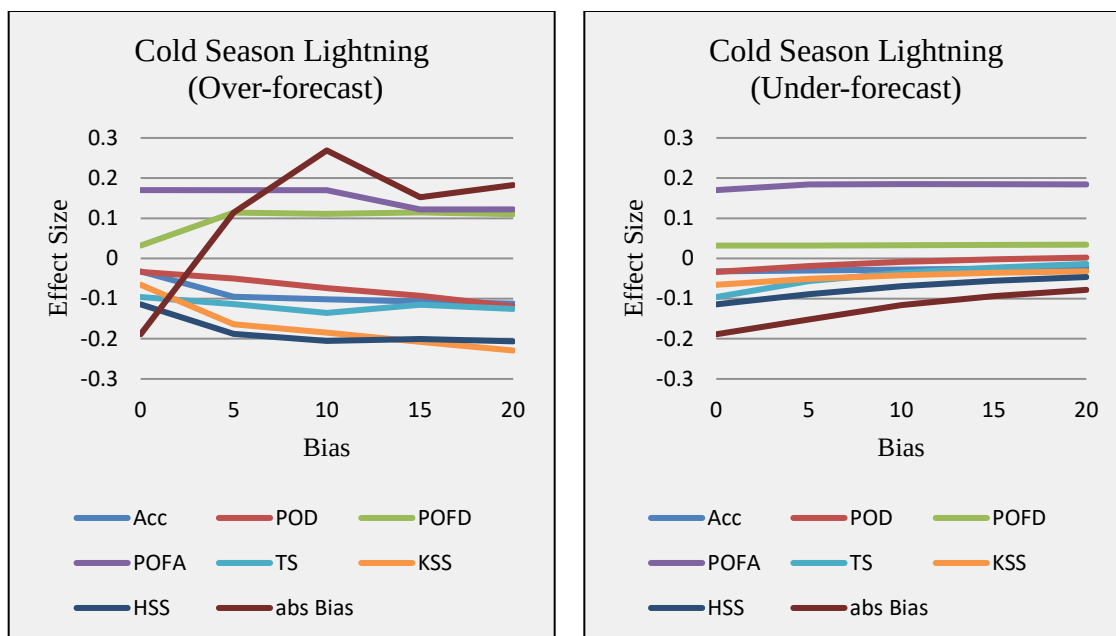


Figure 18. Weighted Cold Season Lightning Effect Size

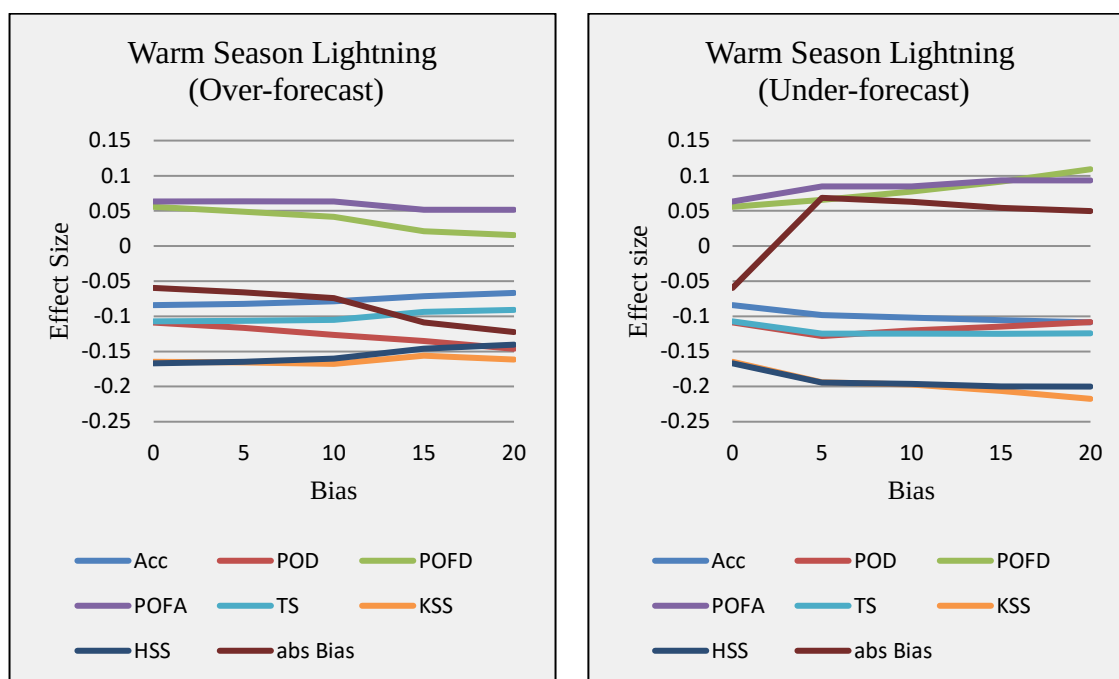


Figure 19. Weighted Warm Season Lightning Effect Size

4.4 45 WS Forecasts and Observations

The final section of analysis includes comparison of metrics calculated from the TCT and PCT for each seasonal event. The cold season is defined as October through May, eight of the twelve months, providing many more observations for the cold season than the warm season. With four years of observations, however, this did not impact the results. Figures 20, 21, 22, and 23 display the TCT and PCT produced from the provided forecasts and outcomes. The figures also display the associated performance metrics, colored in green for the better scores and red for the poor scores. The graphs provide a visual representation of the differences in the metric values. The simulations were each replicated 30 times, however, the variation in metric values was close to zero resulting in population like characteristics of the mean metric values. A practical comparison was conducted with a percent difference between the values from the simulation and real-world data displayed in Tables 23, 25, 27, and 29. Anything under a 5% difference was considered practically insignificant. In the final section, the reliability diagrams as well as BS are addressed for this data to highlight some weaknesses of dichotomizing probabilistic forecasts.

The cold season rain events produced metrics similar to that of the weighted unbiased simulation. Figure 20 displays the TCT and PCT metrics produced from 2015 through 2018 45 WS rain forecasts and observations during October through May. Table 22 displays the unbiased simulation results using the distribution of binned rain forecasts during the cold season. The real-world results and simulation are consistent in regards to the metric values. The PCT KSS values had slightly over a 5% difference, but all other

metrics had under a 5% difference. Consistent with the unbiased seasonal simulation, the TCT produces metrics closer to ideal values compared to the PCT with the exception of frequency bias.

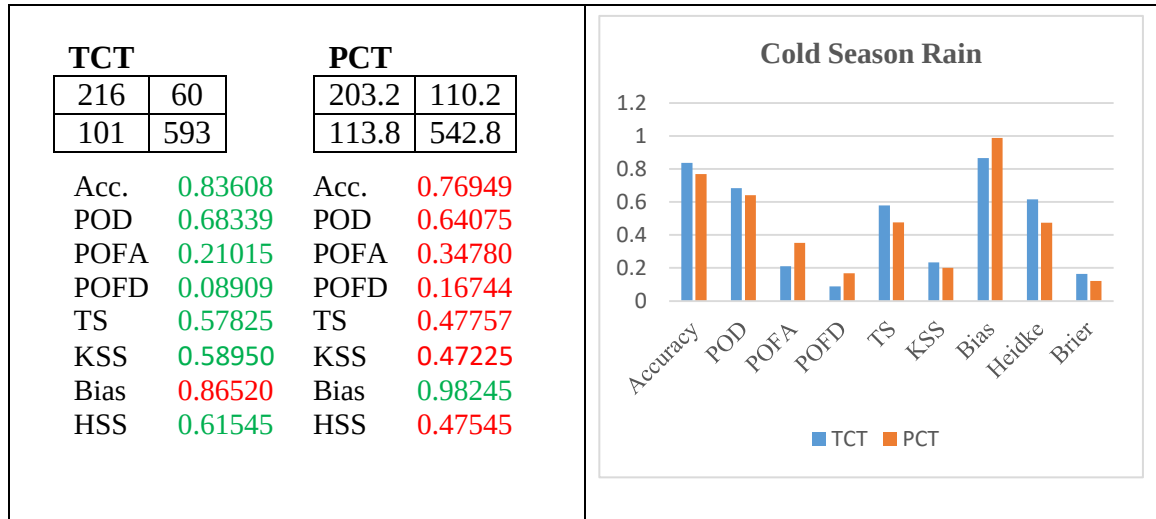


Figure 20. Cold Season Rain 45 WS Metrics

Table 22. Cold Season Rain Unbiased Weighted Simulation Results

TCT		PCT	
Acc.	0.84082	Acc.	0.78002
POD	0.69397	POD	0.65956
POFA	0.21187	POFA	0.34037
POFD	0.08907	POFD	0.16247
TS	0.58486	TS	0.49209
KSS	0.60490	KSS	0.49709
Bias	0.88053	Bias	0.99990
HSS	0.62439	HSS	0.49710

Table 23. Percent Difference in Unbiased Weighted Simulation and 45 WS Data (Cold Season Rain)

Difference	TCT	PCT
Acc.	0.5653%	1.3591%
POD	1.5363%	2.8932%
POFA	0.8151%	2.1594%
POFD	0.0225%	3.0129%
TS	1.1366%	2.9949%
KSS	2.5787%	5.1251%
Bias	1.7563%	1.7605%
HSS	1.4421%	4.4522%

The warm season rain events also produces similar metrics for the TCT and PCT compared to the weighted unbiased warm season rain simulation, evident in Figure 21 and Table 24. The POFA for both the TCT and PCT had slightly over a 5% difference,

but all other metrics were below the practical significance level. Metrics calculated using the TCT are closer to the perfect scores other than the frequency bias metric which is closer to one utilizing the PCT.

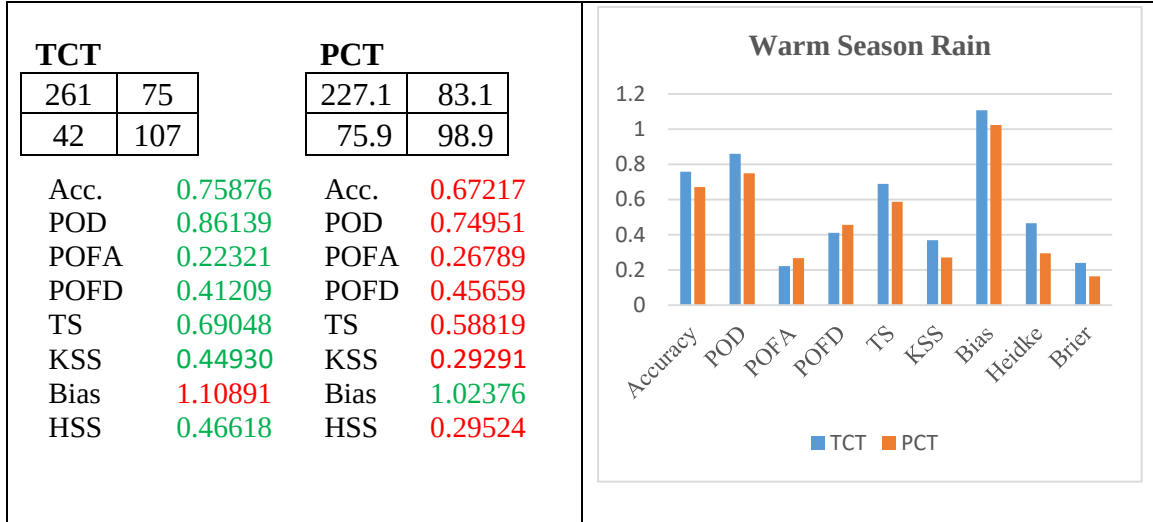


Figure 21. Warm Season Rain 45 WS Metrics

Table 24. Warm Season Rain Unbiased Weighted Simulation Results

TCT		PCT	
Acc.	0.76004	Acc.	0.67493
POD	0.85392	POD	0.74585
POFA	0.21148	POFA	0.25408
POFD	0.40665	POFD	0.45096
TS	0.69480	TS	0.59475
KSS	0.44727	KSS	0.29489
Bias	1.08294	Bias	0.99991
HSS	0.46213	HSS	0.29488

Table 25. Percent Difference in Unbiased Weighted Simulation and 45 WS Data (Warm Season Rain)

Difference	TCT	PCT
Acc.	0.1686%	0.4098%
POD	0.8710%	0.4895%
POFA	5.3969%	5.2915%
POFD	1.3289%	1.2407%
TS	0.6237%	1.1091%
KSS	0.4528%	0.6737%
Bias	2.3697%	2.3571%
HSS	0.8726%	0.1220%

The cold season lightning events had similar metrics from the TCT and PCT compared to the weighted unbiased cold season lightning simulation. Like the cold season rain, there were many correct negative results because of the sparsity of events

during the colder season. Table 26 displays the results from the unbiased cold season lightning referenced for comparison. The POFA and POFD from the TCT had differences around 11%, but all other simulation and real-world differences were below 5%.

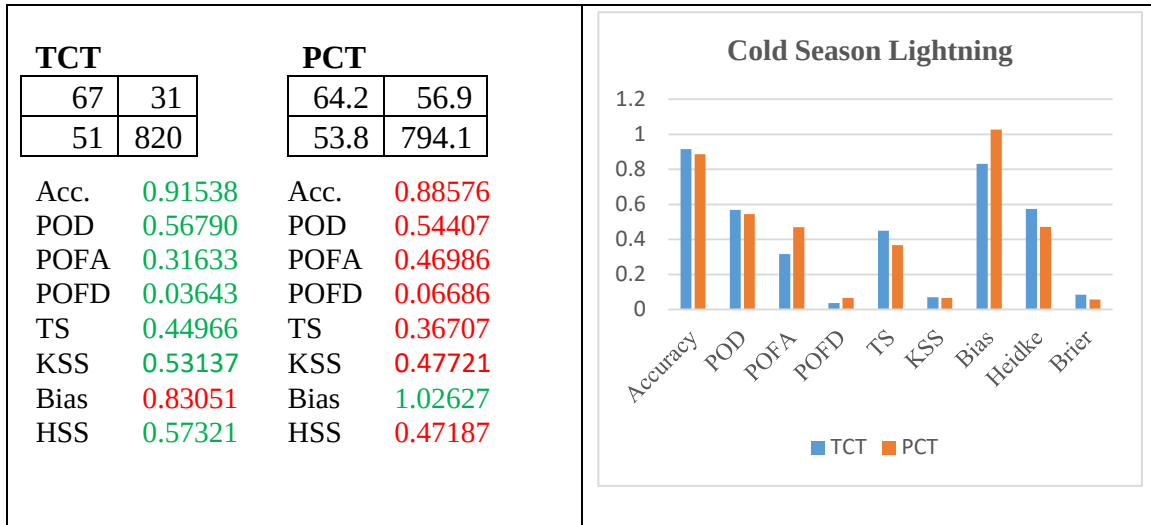


Figure 22. Cold Season Lightning 45 WS Metrics

Table 26. Cold Season Lightning Unbiased Weighted Simulation Results

TCT		PCT	
Acc.	0.91894	Acc.	0.88676
POD	0.58022	POD	0.54691
POFA	0.28234	POFA	0.45262
POFD	0.03264	POFD	0.06465
TS	0.47239	TS	0.37660
KSS	0.54758	KSS	0.48226
Bias	0.80849	Bias	0.99914
HSS	0.59654	HSS	0.48244

Table 27. Percent Difference in Unbiased Weighted Simulation and 45 WS Data (Cold Season Lightning)

Difference	TCT	PCT
Acc.	0.3882%	0.1128%
POD	2.1461%	0.5206%
POFA	11.3552%	3.7378%
POFD	10.9744%	3.3610%
TS	4.9303%	2.5630%
KSS	3.0048%	1.0527%
Bias	2.6870%	2.6790%
HSS	3.9889%	2.2152%

The comparison of the weighted simulation results and the real-world events reveal the 45 WS warm season lightning forecasts were not consistent with the expected simulated results. According to the frequency bias metric alone, the results were more

consistent with under forecasting by 5%, evident in the low percent difference from Table 29. With a 5% under bias, all metrics had superior scores using the TCT. Referring back to Figure 19, at the 5% bias level, POFD, POFA, and absolute frequency bias were above zero indicating better performance of the TCT. Accuracy, POD, TS, and KSS were under zero, indicating again superiority of the TCT. Although the frequency bias metric is consistent with the 5% under-forecast, the other seven tested metrics have greater disparity from the real-world results when referencing the 5% under-forecasted simulation, apparent in Table 28. This is a result of specific unreliability in probability forecast bins, requiring a more thorough examination of reliability diagrams to make necessary adjustments.

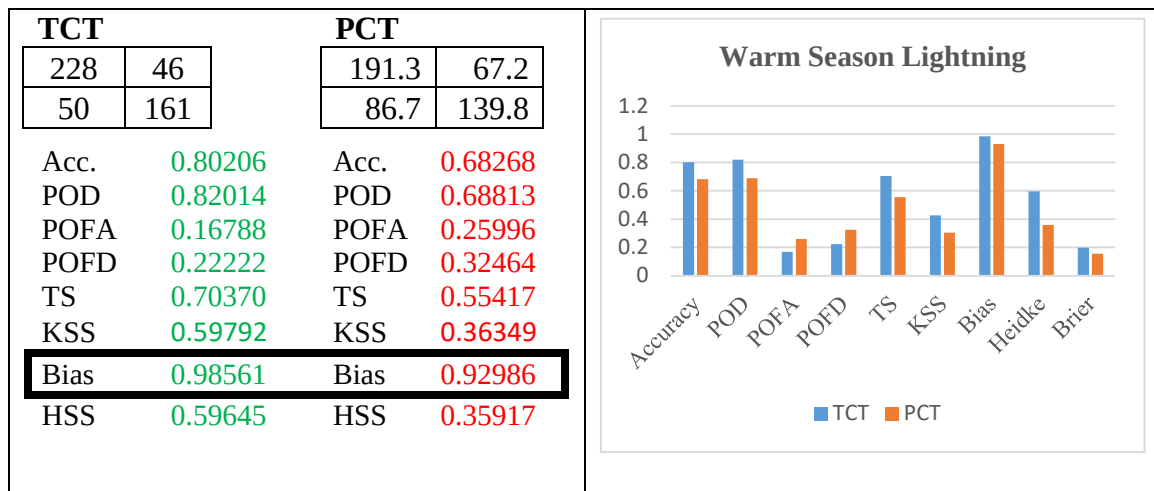


Figure 23. Warm Season Lightning 45 WS Metrics

Table 28. Warm Season Lightning Unbiased & 5% Under-Forecasted Weighted Simulation Results

Unbiased				5% Under-Forecast			
TCT		PCT		TCT		PCT	
Acc.	0.75746	Acc.	0.67356	Acc.	0.72870	Acc.	0.63030
POD	0.80239	POD	0.69374	POD	0.72786	POD	0.59985
POFA	0.24284	POFA	0.30615	POFA	0.25175	POFA	0.33671
POFD	0.29384	POFD	0.34947	POFD	0.27038	POFD	0.33610
TS	0.63816	TS	0.53115	TS	0.58468	TS	0.45983
KSS	0.50855	KSS	0.34427	KSS	0.45748	KSS	0.26375
Bias	1.05974	Bias	0.99984	Bias	0.97275	Bias	0.90435
HSS	0.51072	HSS	0.34426	HSS	0.45683	HSS	0.26245

Table 29. Percent Difference in Unbiased and 5% Under-Forecasted Weighted Simulations and 45 WS Data (Warm Season Lightning)

Difference	TCT	PCT	Difference	TCT	PCT
Acc.	5.7197%	1.3449%	Acc.	9.5848%	7.9788%
POD	2.1879%	0.8119%	POD	11.9225%	13.7083%
POFA	36.5018%	16.3184%	POFA	39.9733%	25.7261%
POFD	27.7565%	7.3668%	POFD	19.5534%	3.4688%
TS	9.7685%	4.2421%	TS	18.4759%	18.6075%
KSS	16.1541%	5.4312%	KSS	26.6136%	31.8028%
Bias	7.2486%	7.2529%	Bias	1.3133%	2.7816%
HSS	15.4863%	4.2392%	HSS	26.5115%	31.1187%

4.5 Probabilistic Evaluation of 45 WS Forecasts and Observations

In further examination of the reliability diagrams, the weaknesses in dichotomizing are highlighted. Figure 24 depicts reliability from dichotomizing the forecast bins of 40% or less as 0% and 50% or greater as 100%. The graphs do not provide enough information to make adjustments to the individual forecasts. Figure 25 depicts the appropriate reliability diagram for each seasonal event from 2015 through 2018. The discrepancy between the simulated warm season lightning events and real-

world results is a consequence of significantly under-forecasting for 0%, 40%, 50%, 60%, and 70% forecasts while simultaneously over-forecasting significantly for 20%, 30%, and 100% forecasts. Probabilistic forecasts provide more detailed information to find room for improvement. If forecasts were made with perfect sharpness in discriminating between binary events, reliability diagrams such as those in Figure 28 would be appropriate, but forecasts contain uncertainty.

The BS containing the reliability, resolution and uncertainty favors maintaining the probabilities rather than dichotomizing. The BS for the real-world data calculated with dichotomization and probabilities is displayed in Table 30.

Table 30. Brier Scores for Dichotomized and Probabilistic 45 WS Forecasts

Brier Score	Dichotomized Forecasts	Probabilistic Forecasts
Cold Season Rain	0.16392	0.12052
Warm Season Rain	0.24124	0.16532
Cold Season Lightning	0.08462	0.05766
Warm Season Lightning	0.19794	0.15412

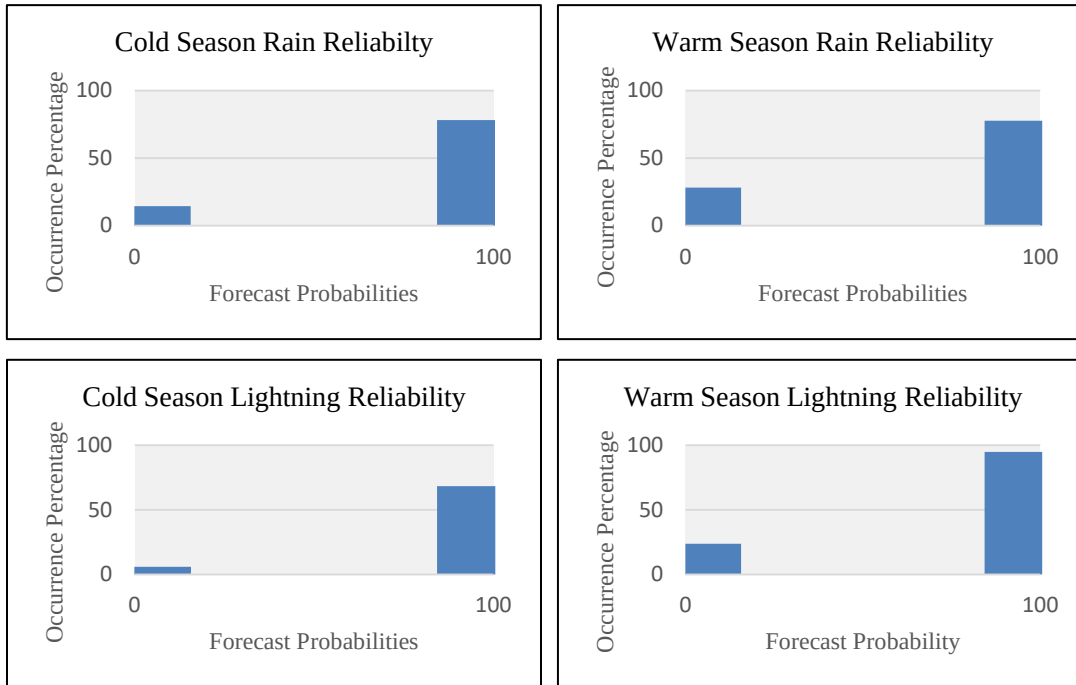


Figure 24. 2015-2018 45 WS Dichotomized Forecast Reliability Diagram

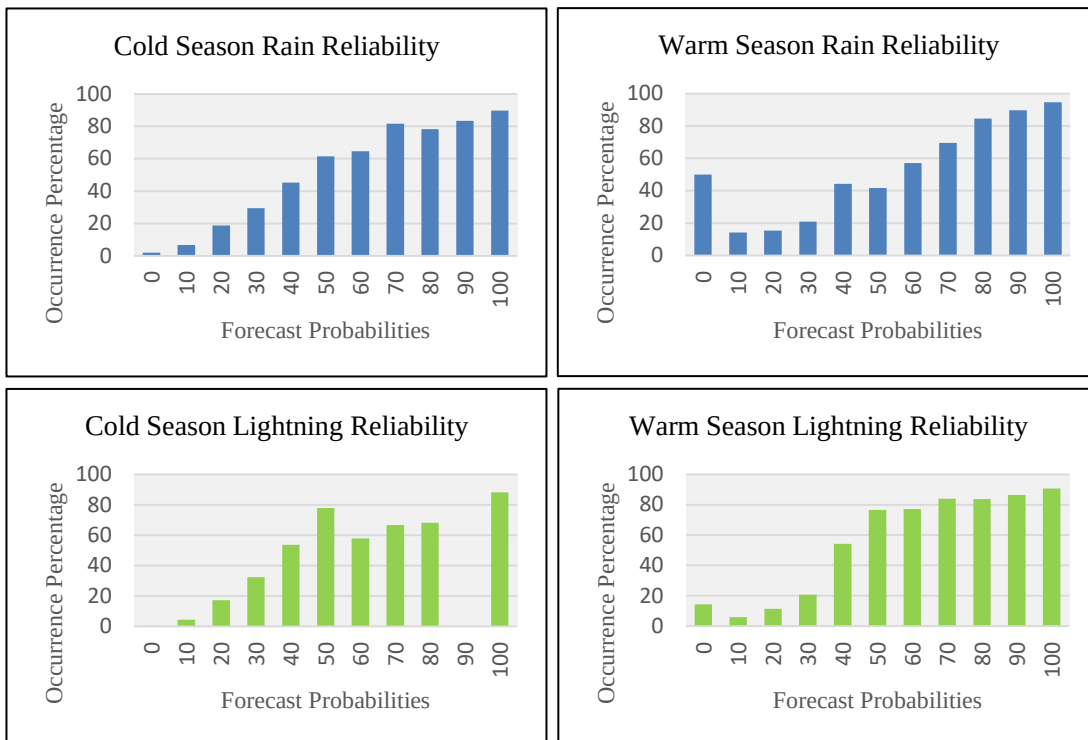


Figure 25. 2015-2018 45 WS Probabilistic Forecast Reliability Diagram

4.6 Summary

The unweighted simulations revealed differences in the metrics from the TCT and PCT by individual forecast bin from 0% to 100%, incremented by 10%. Overall the TCT and PCT had the biggest discrepancy in scores at the central forecasts of 40% and 50%. The table type producing values closer to the perfect score changed at the central forecasts. POFA, KSS, and HSS had negligible differences for all bias levels at every forecast bin. With the binary constraints, POD and POFD values for the individually binned forecast TCTs were either one or zero, lacking descriptive information regarding performance.

For the weighted simulations, the TCT had metric values closer to the perfect scores for all metrics except frequency bias. The PCT showed frequency bias closer to one for all unbiased simulations, warm season over-forecasts, and cold season under-forecasts. The TCT revealed frequency bias values closer to one for the cold season under-forecasts and warm season over-forecasts.

The real-world data table comparison of cold season rain, warm season rain, and cold season lightning forecasts resulted in values closer to perfect scores for the TCT in regards to all metrics except frequency bias. The warm season lightning results were not consistent with the unbiased weighted simulation, eliciting a need for further examination. The TCT metrics were all closer to the perfect scores for this seasonal event, but exploring the reliability, the forecasts of many of the forecast bins were discovered to be biased. Chapter 5 discusses the implications of the results determined from this research.

V. Conclusions and Recommendations

5.1 Chapter Overview

Chapter 5 discusses the implications of poor verification practices utilizing the TCT and reiterates why the 45 WS should instead use the PCT for verification. Section 5.2 summarizes the methodology and results from this research. Section 5.3 addresses recommendations and potential adjustments for any future comparison of the TCT and PCT using 45 WS forecasts. Section 5.5 provides a final justification for avoiding dichotomization of the probabilistic forecasts and need for the PCT.

5.2 Discussion

The BS along with the other metric comparisons for the TCT and PCT reveal an apparent difference between the two tools in analyzing the 45 WS data. The differences however do not directly reveal which tool should be utilized to improve forecasts. The reliability diagrams support the intuitive practice of maintaining probabilities in the verification process as more information is readily apparent. The verification methods should represent the true system, and forecasts are generated with probabilities and not with binary certainty. The simulation frequency bias results reveal the metrics from the TCT do not accurately represent the true forecasts and observations. 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 rain frequency bias from the TCT indicates the 45 WS forecasters should be predicting fewer rain events. Making such an erroneous adjustment could result in a lack of necessary precautions. The metric comparisons revealed the greatest disparity in the TCT and PCT according to forecast bin and bias level.

5.3 Summary

This research compared a traditional 2 x 2 contingency table used for verifying binary weather events to a proposed alternative tool termed a Probabilistic Contingency Table. The comparison was conducted by first simulating uniform weather events and computing the values in each of the tables' four cells for hits, false alarms, misses, and correct negatives. The tables were computed for each forecast, binned by 10% increments from 0% to 100%. This first part of the simulation revealed the greatest discrepancy around the threshold of 50%. It also revealed the TCT frequency bias differed from the expected perfect one. Varying levels of bias were tested to determine the trends and differences in metric values assuming the forecasts were not perfectly reliable. Bias levels did not affect POD, POFD, POFA, KSS, or HSS for the individual forecast simulations.

The next part of the research involved simulating the proportional number of binned forecasts, based on the number of rain or lightning events forecasted by the 45 WS during 2015 through 2018. The averaged metrics revealed values closer to the perfect scores utilizing the TCT for all except frequency bias. Different levels of bias were again

added to the simulation, resulting in increasing or decreasing effect size trends for each over and under-forecasting simulation by season and event type. Following the simulations, the data used to produce the weights was examined for the same metrics in addition to BS and reliability. The values for three of the four seasonal events were consistent with the unbiased weighted simulations. The TCT frequency bias indicated over and under-forecasting for the 45 WS data, however with the weighted simulation as evidence such bias was not present. The BS and reliability diagrams highlighted the prevalent information available from maintaining the probabilistic forecasts.

5.4 Future Research Recommendations

As discussed in Chapter 3, May and October were not statistically different from either of the seasons defined for rain events and May was not statistically different for lightning events. Future research should explore splitting months into biweekly increments to determine seasons. Previous forecasts and observations prior to PCT implementation should also be used to compare forecast improvement. It is not possible to directly attribute forecast adjustments to any one metric, but evaluating changes in performance would reflect the value of the verification process. The data provided by the 45 WS was collected while using the PCT. Comparing the performance to earlier years would help compare the impact of the tables. The simulations were able to provide information on changes made based on bias level, but other real-world adjustments made to the subjective forecasts are not as easily simulated. Due to this limitation, the data prior to PCT implementation would provide the information necessary to compare the

TCT and PCT. A final area for further research should explore an optimal threshold for the TCT to compare again to the PCT.

5.5 Final Remarks

Proper verification techniques should not overestimate or underestimate forecast performance. 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. The metrics addressed in this research all attempt to measure the strength in forecast predictions. Each metric addresses different considerations of performance.

Accuracy reveals the proportion of forecasts correctly predicted. POD considers of the positive event outcomes how many were forecasted as likely to occur. POFD addresses the number of false alarms in proportion to the non-event outcomes. POFA considers the proportion of misses out of forecasts predicted as likely to occur. The TS answers how well the forecasted likely to occur events correspond to the positive event outcomes. KSS compares forecast skill to random chance. HSS compares forecast accuracy to random chance. Frequency bias reveals the proportion of forecasted to occur events to the positive event outcomes. The importance of each metric differs depending on the user's priorities.

From the seasonal forecasts produced by the 45 WS each year, the metrics examined in this thesis reveal a difference when utilizing the TCT and PCT. For both

seasonal events the TCT presented values closer to the perfect scores for accuracy, POD, POFD, POFA, TS, KSS, and HSS. Frequency bias values were closer to a perfect one using the PCT for rain events in both seasons, and lightning events in the cold season. The 45 WS warm season lightning events were slightly under-forecasted, so the TCT produced a frequency bias closer to one. 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 calculated after dichotomizing the forecasts did not converge to the probabilistic results. For the unweighted simulations, the greatest disparity for each metric occurred around the threshold marker of 50%. The TCT is artificially favored for the metrics examined with the exception of frequency bias. The lack of fidelity in the TCT metric scores leaves little confidence for making forecast adjustments.

Dichotomizing is the result of an aversion to probabilities, however it diminishes the available information. 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, supplemented with other measures such as reliability and the BS which also maintain probabilistic forecasts for verification. To promote quality forecasts, the verification process must evaluate the true probabilistic prediction determined by the forecaster through utilization of the PCT and seasonal reliability diagrams.

Appendix A

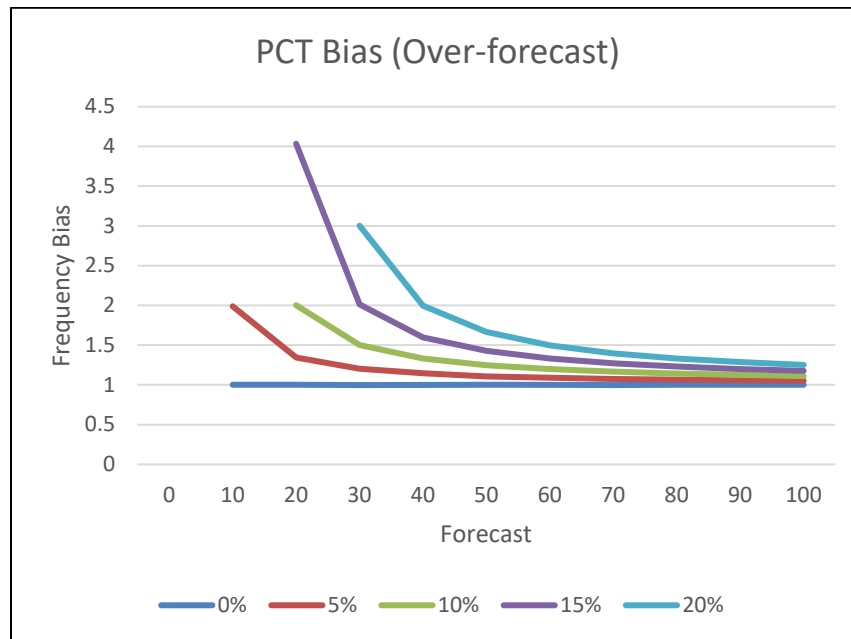


Figure A. 1

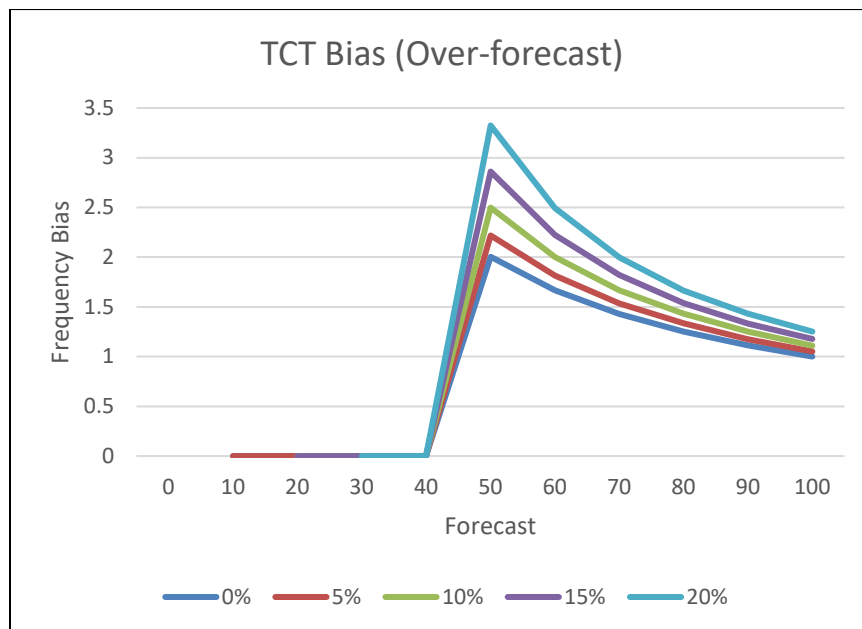


Figure A. 2

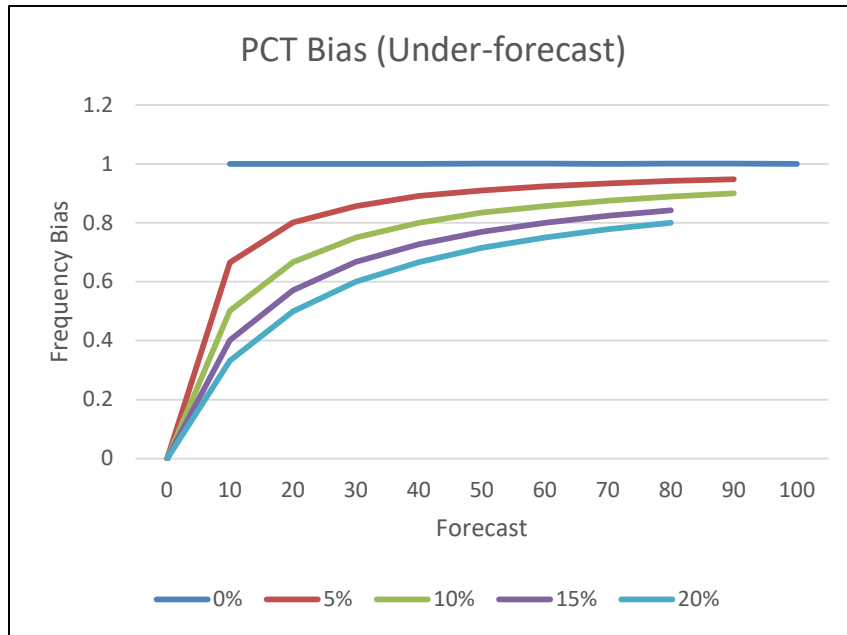


Figure A. 3

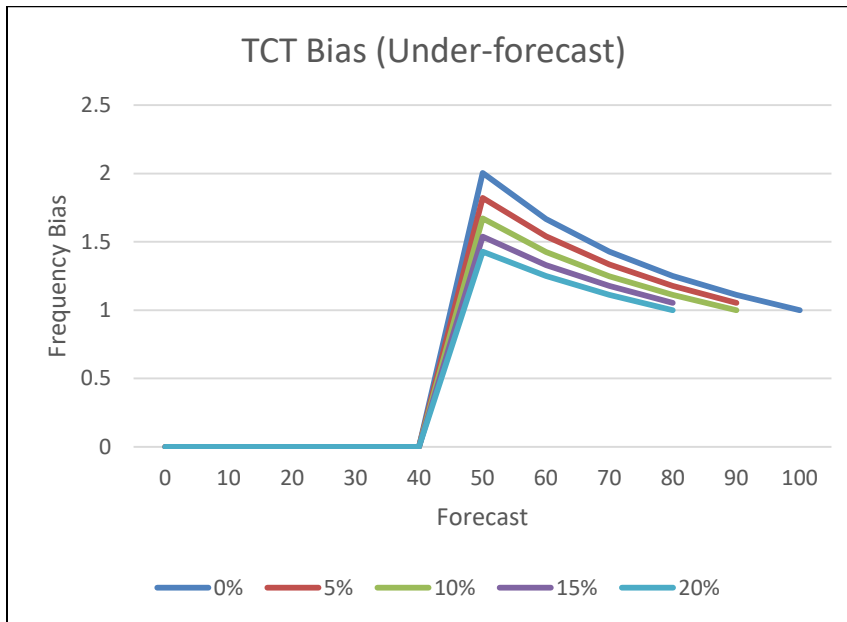


Figure A. 4

Appendix B

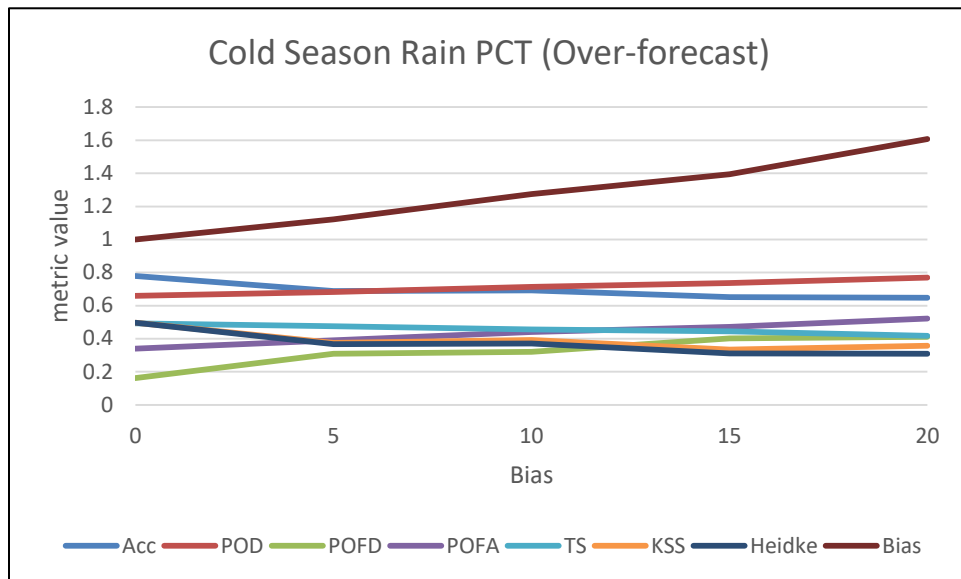


Figure B. 1

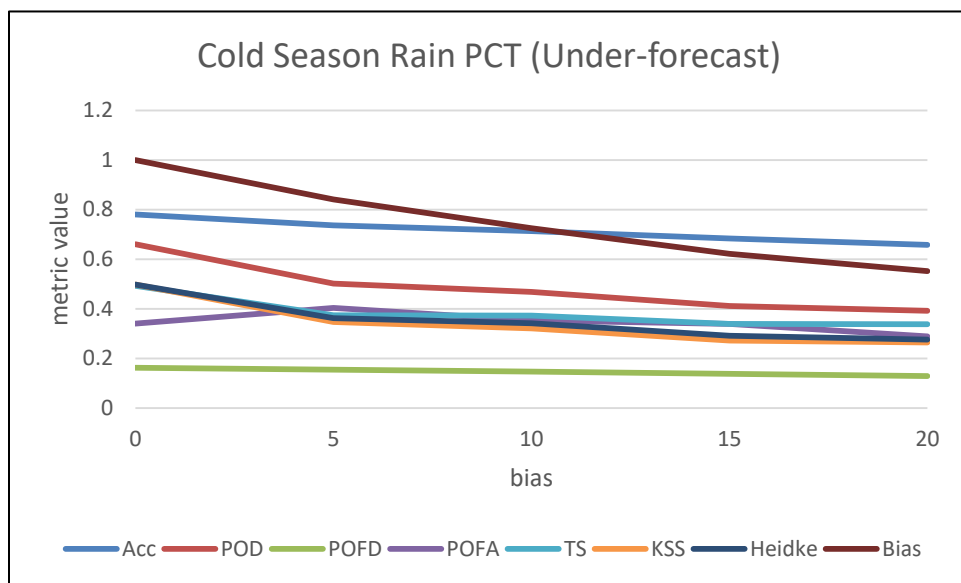


Figure B. 2

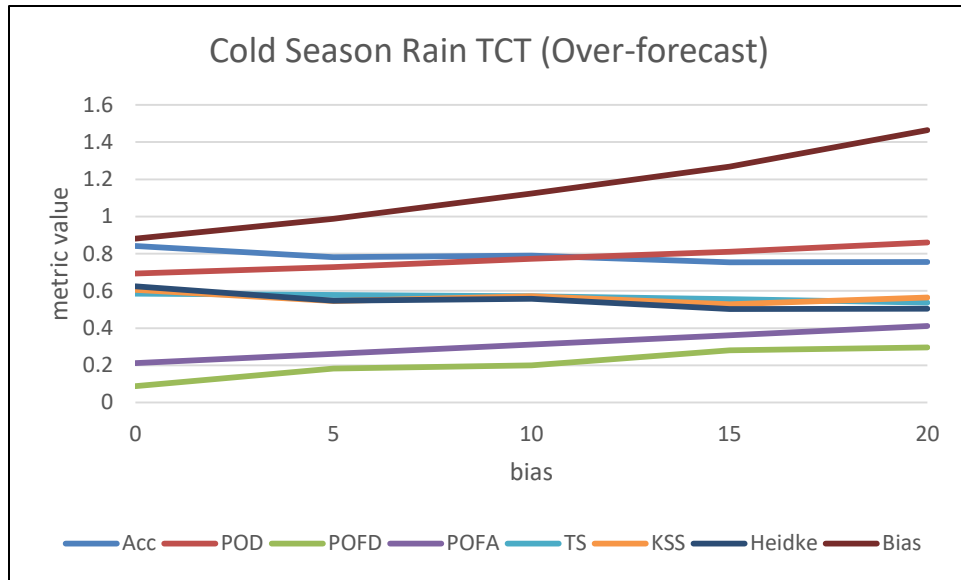


Figure B. 3

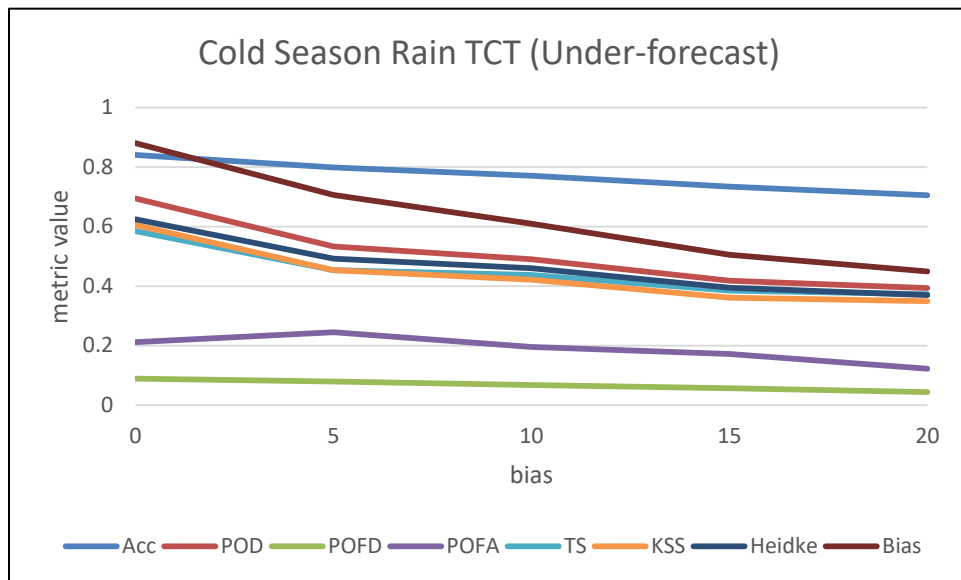


Figure B. 4

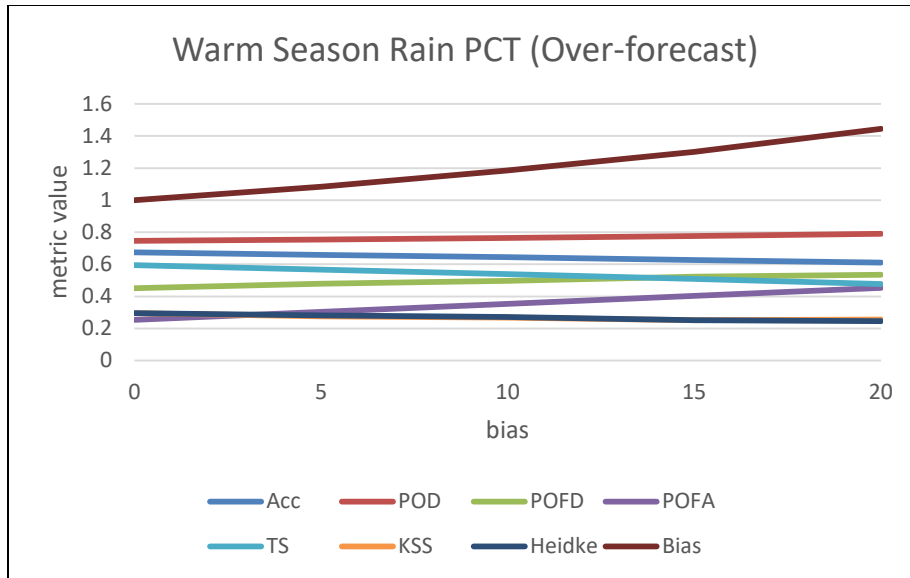


Figure B. 5

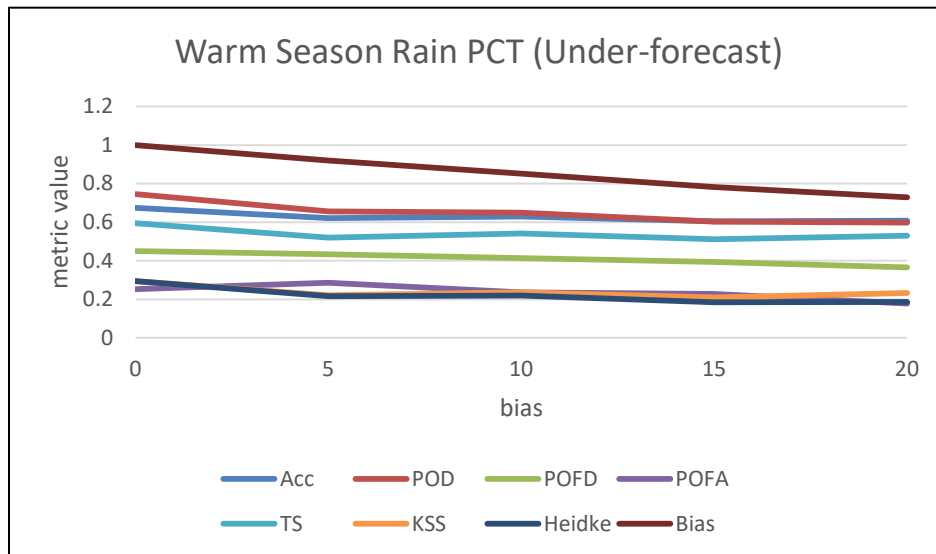


Figure B. 6

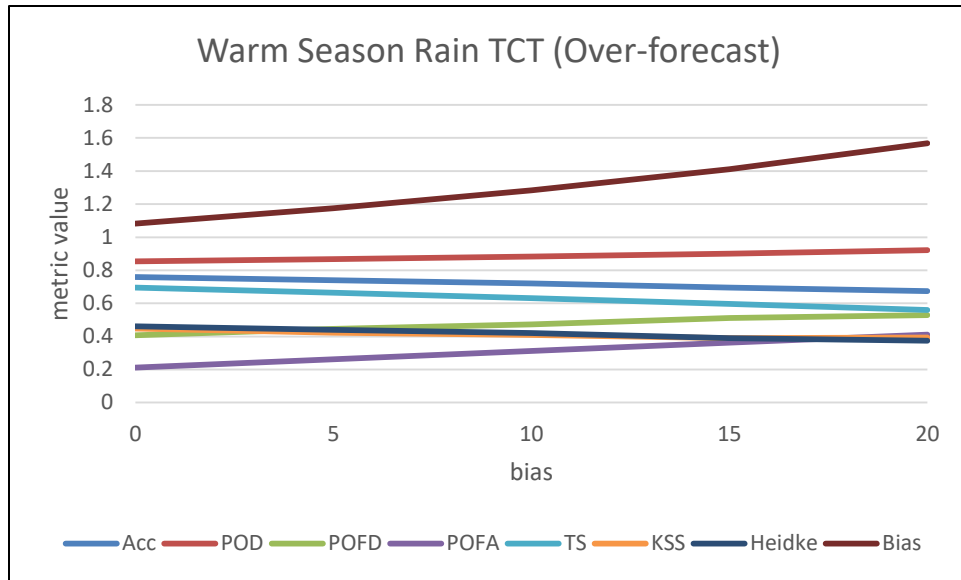


Figure B. 7

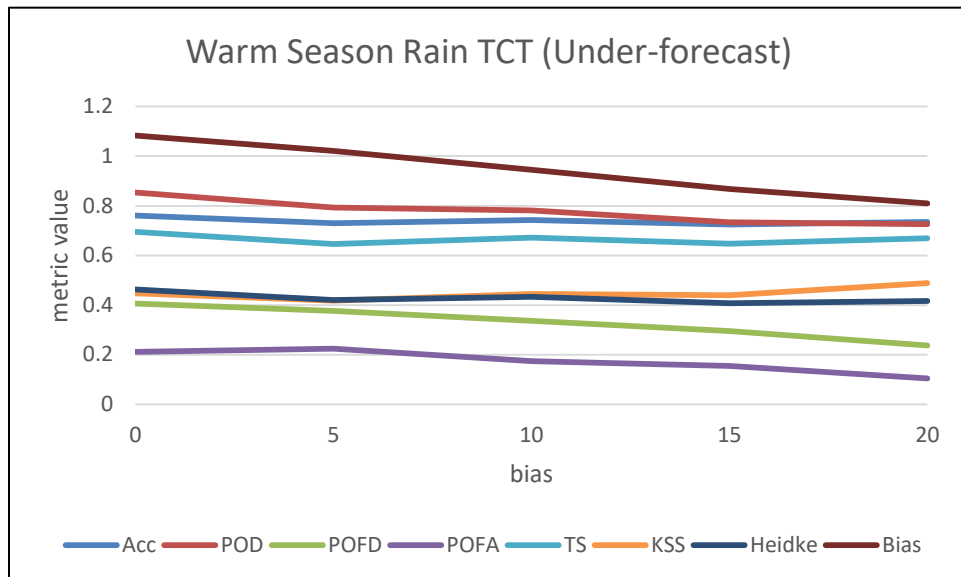


Figure B. 8

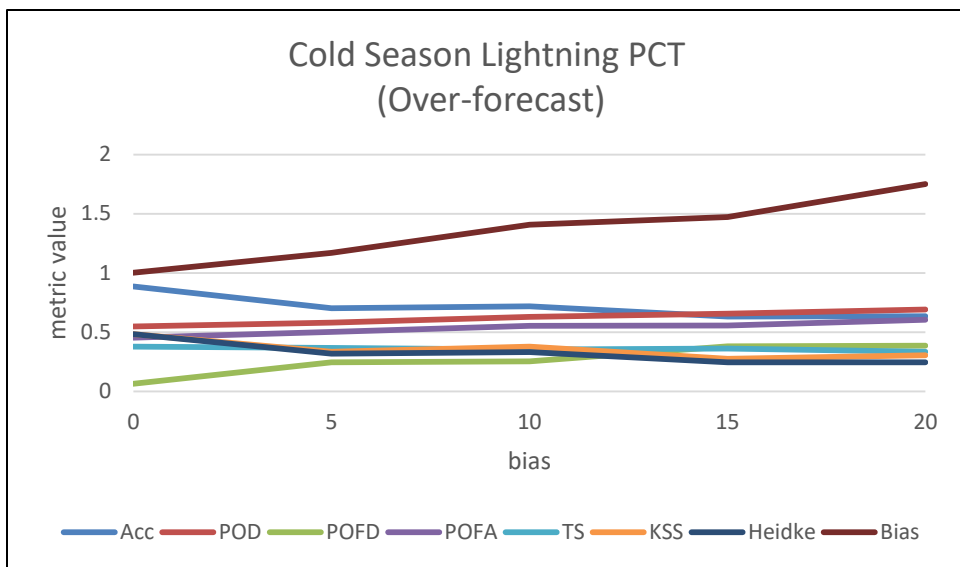


Figure B. 9

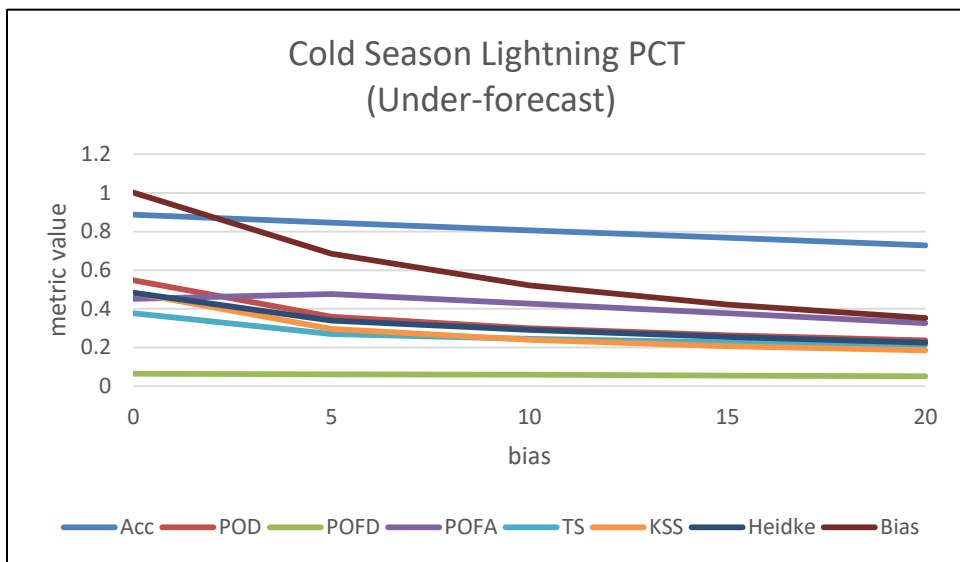


Figure B. 10

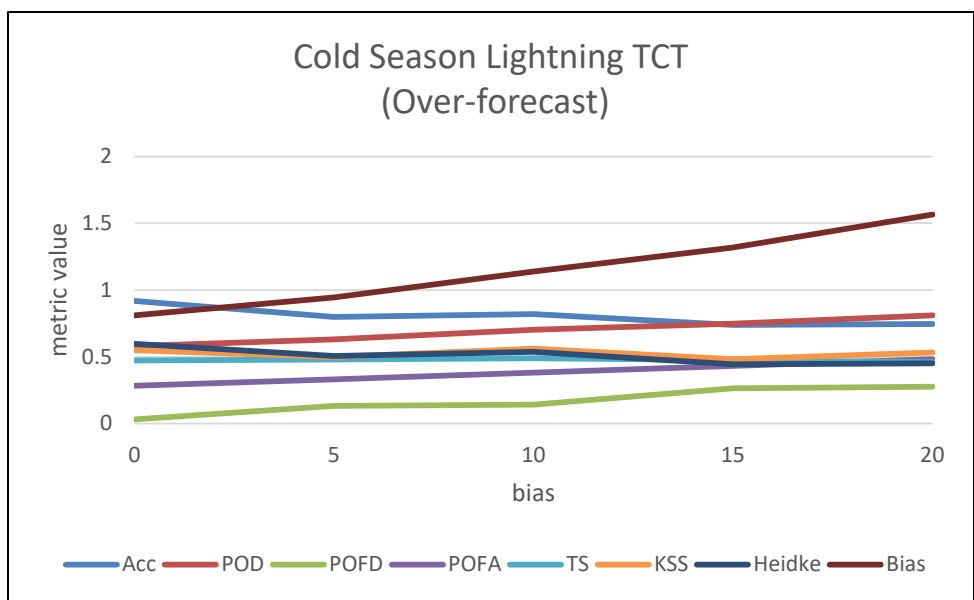


Figure B. 11

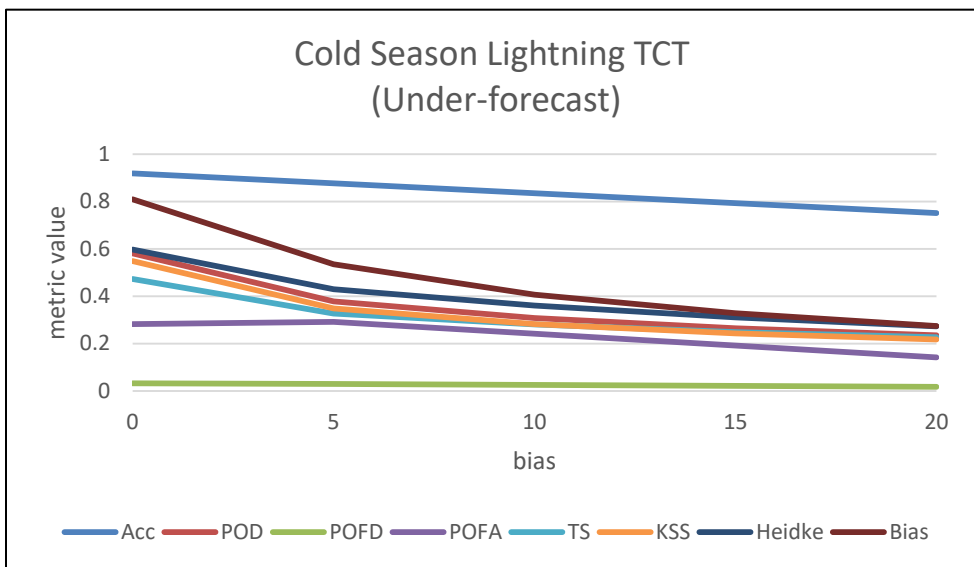


Figure B. 12

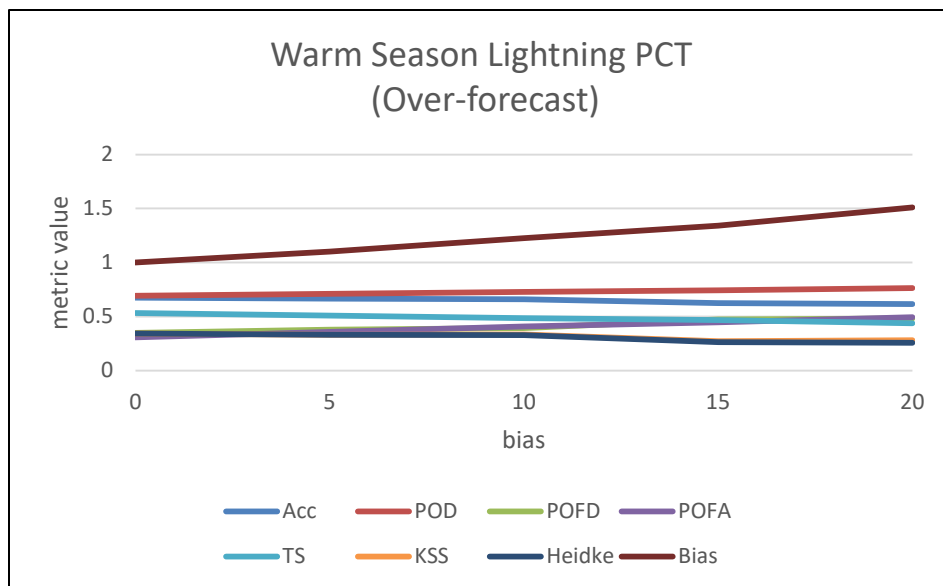


Figure B. 13

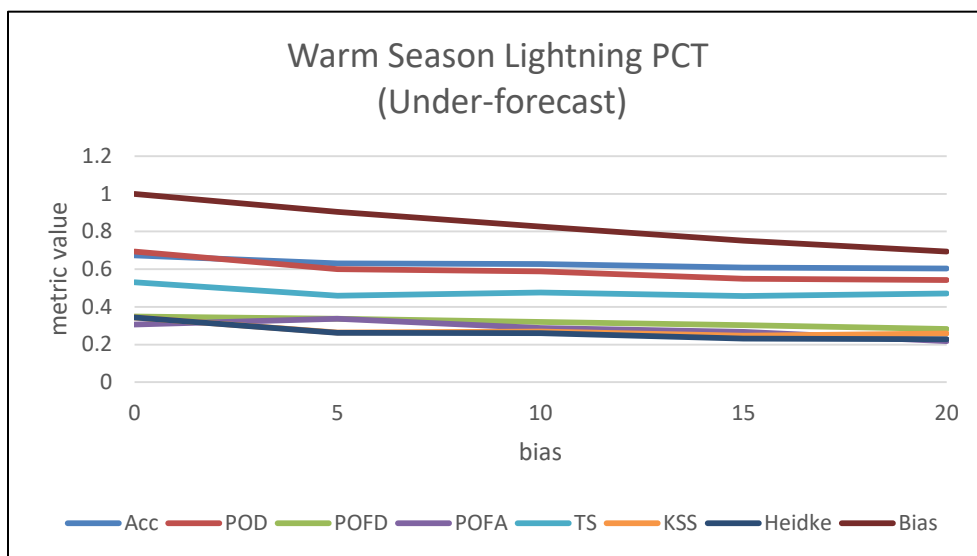


Figure B. 14

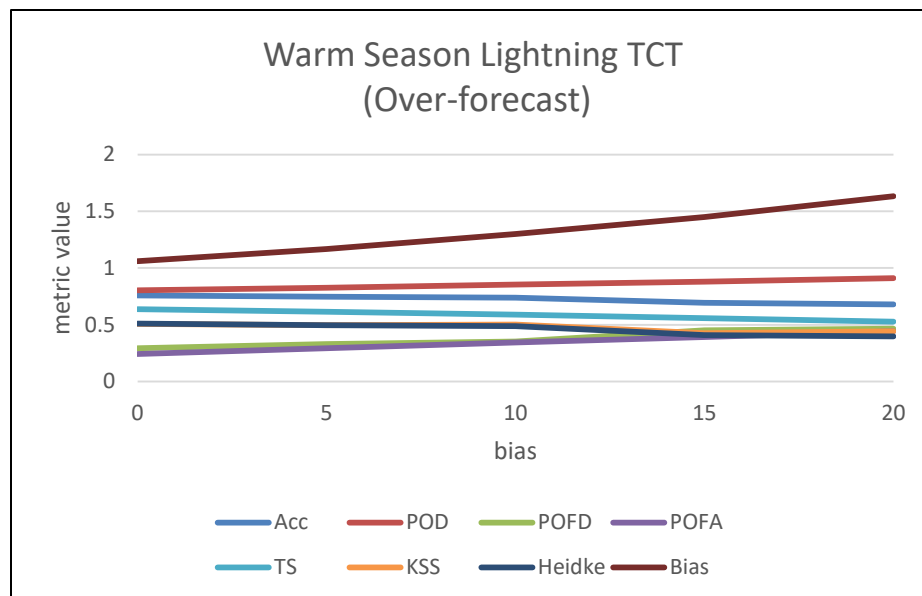


Figure B. 15

Appendix C

All Pairwise Differences							
Month	-Month	Difference	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Jan	Feb	0.023715	0.0320790	0.74	0.9999	-0.081293	0.128724
Jan	Mar	-0.029032	0.0313257	-0.93	0.9989	-0.131575	0.073510
Jan	Apr	-0.093091	0.0315857	-2.95	0.1257	-0.196485	0.010302
Jan	May	-0.156452	0.0313257	-4.99	<.0001 *	-0.258994	-0.053909
Jan	Jun	-0.473091	0.0315857	-14.98	<.0001 *	-0.576485	-0.369698
Jan	Jul	-0.484677	0.0313257	-15.47	<.0001 *	-0.587220	-0.382135
Jan	Aug	-0.442742	0.0313257	-14.13	<.0001 *	-0.545284	-0.340200
Jan	Sep	-0.356831	0.0317908	-11.22	<.0001 *	-0.460896	-0.252766
Jan	Oct	-0.054366	0.0315193	-1.72	0.8568	-0.157542	0.048811
Jan	Nov	0.027742	0.0315857	0.88	0.9993	-0.075651	0.131135
Jan	Dec	0.028514	0.0313893	0.91	0.9990	-0.074236	0.131265
Feb	Mar	-0.052748	0.0320790	-1.64	0.8923	-0.157756	0.052261
Feb	Apr	-0.116807	0.0323329	-3.61	0.0164 *	-0.222646	-0.010967
Feb	May	-0.180167	0.0320790	-5.62	<.0001 *	-0.285175	-0.075159
Feb	Jun	-0.496807	0.0323329	-15.37	<.0001 *	-0.602646	-0.390967
Feb	Jul	-0.508393	0.0320790	-15.85	<.0001 *	-0.613401	-0.403385
Feb	Aug	-0.466457	0.0320790	-14.54	<.0001 *	-0.571466	-0.361449
Feb	Sep	-0.380546	0.0325333	-11.70	<.0001 *	-0.487042	-0.274051
Feb	Oct	-0.078081	0.0322681	-2.42	0.3931	-0.183708	0.027546
Feb	Nov	0.004027	0.0323329	0.12	1.0000	-0.101813	0.109866
Feb	Dec	0.004799	0.0321411	0.15	1.0000	-0.100413	0.110011
Mar	Apr	-0.064059	0.0315857	-2.03	0.6741	-0.167453	0.039334
Mar	May	-0.127419	0.0313257	-4.07	0.0029 *	-0.229962	-0.024877
Mar	Jun	-0.444059	0.0315857	-14.06	<.0001 *	-0.547453	-0.340666
Mar	Jul	-0.455645	0.0313257	-14.55	<.0001 *	-0.558188	-0.353103
Mar	Aug	-0.413710	0.0313257	-13.21	<.0001 *	-0.516252	-0.311167
Mar	Sep	-0.327798	0.0317908	-10.31	<.0001 *	-0.431863	-0.223734
Mar	Oct	-0.025333	0.0315193	-0.80	0.9997	-0.128509	0.077843
Mar	Nov	0.056774	0.0315857	1.80	0.8195	-0.046619	0.160168
Mar	Dec	0.057547	0.0313893	1.83	0.7994	-0.045204	0.160297
Apr	May	-0.063360	0.0315857	-2.01	0.6894	-0.166754	0.040033
Apr	Jun	-0.380000	0.0318435	-11.93	<.0001 *	-0.484237	-0.275763
Apr	Jul	-0.391586	0.0315857	-12.40	<.0001 *	-0.494979	-0.288193
Apr	Aug	-0.349651	0.0315857	-11.07	<.0001 *	-0.453044	-0.246257
Apr	Sep	-0.263739	0.0320470	-8.23	<.0001 *	-0.368643	-0.158836
Apr	Oct	0.038726	0.0317777	1.22	0.9875	-0.065296	0.142748
Apr	Nov	0.120833	0.0318435	3.79	0.0085 *	0.016596	0.225071
Apr	Dec	0.121606	0.0316488	3.84	0.0071 *	0.018006	0.225206
May	Jun	-0.316640	0.0315857	-10.02	<.0001 *	-0.420033	-0.213246
May	Jul	-0.328226	0.0313257	-10.48	<.0001 *	-0.430768	-0.225683
May	Aug	-0.286290	0.0313257	-9.14	<.0001 *	-0.388833	-0.183748
May	Sep	-0.200379	0.0317908	-6.30	<.0001 *	-0.304444	-0.096314
May	Oct	0.102086	0.0315193	3.24	0.0555	-0.001090	0.205262
May	Nov	0.184194	0.0315857	5.83	<.0001 *	0.080800	0.287587
May	Dec	0.184966	0.0313893	5.89	<.0001 *	0.082215	0.287717
Jun	Jul	-0.011586	0.0315857	-0.37	1.0000	-0.114979	0.091807
Jun	Aug	0.030349	0.0315857	0.96	0.9984	-0.073044	0.133743
Jun	Sep	0.116261	0.0320470	3.63	0.0155 *	0.011357	0.221164
Jun	Oct	0.418726	0.0317777	13.18	<.0001 *	0.314704	0.522748
Jun	Nov	0.500833	0.0318435	15.73	<.0001 *	0.396596	0.605071
Jun	Dec	0.501606	0.0316488	15.85	<.0001 *	0.398006	0.605206
Jul	Aug	0.041935	0.0313257	1.34	0.9739	-0.060607	0.144478
Jul	Sep	0.127847	0.0317908	4.02	0.0035 *	0.023782	0.231912
Jul	Oct	0.430312	0.0315193	13.65	<.0001 *	0.327136	0.533488
Jul	Nov	0.512419	0.0315857	16.22	<.0001 *	0.409026	0.615813
Jul	Dec	0.513192	0.0313893	16.35	<.0001 *	0.410441	0.615942
Aug	Sep	0.085911	0.0317908	2.70	0.2259	-0.018154	0.189976
Aug	Oct	0.388376	0.0315193	12.32	<.0001 *	0.285200	0.491552
Aug	Nov	0.470484	0.0315857	14.90	<.0001 *	0.367090	0.573877
Aug	Dec	0.471256	0.0313893	15.01	<.0001 *	0.368506	0.574007
Sep	Oct	0.302465	0.0319816	9.46	<.0001 *	0.197776	0.407155
Sep	Nov	0.384573	0.0320470	12.00	<.0001 *	0.279669	0.489476
Sep	Dec	0.385345	0.0318535	12.10	<.0001 *	0.281075	0.489615
Oct	Nov	0.082107	0.0317777	2.58	0.2898	-0.021914	0.186129
Oct	Dec	0.082880	0.0315825	2.62	0.2669	-0.020503	0.186263
Nov	Dec	0.000772	0.0316488	0.02	1.0000	-0.102828	0.104372

Figure C. 1

All Pairwise Differences							
Month	-Month	Difference	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Jan	Feb	0.081851	0.0397078	2.06	0.6506	-0.048129	0.211832
Jan	Mar	0.037097	0.0387754	0.96	0.9985	-0.089831	0.164025
Jan	Apr	-0.034624	0.0390972	-0.89	0.9993	-0.162605	0.093358
Jan	May	-0.075806	0.0387754	-1.96	0.7239	-0.202735	0.051122
Jan	Jun	-0.347124	0.0390972	-8.88	<.0001*	-0.475105	-0.219142
Jan	Jul	-0.335484	0.0387754	-8.65	<.0001*	-0.462412	-0.208556
Jan	Aug	-0.333871	0.0387754	-8.61	<.0001*	-0.460799	-0.206943
Jan	Sep	-0.285436	0.0393511	-7.25	<.0001*	-0.414248	-0.156623
Jan	Oct	-0.122654	0.0390150	-3.14	0.0735	-0.250367	0.005059
Jan	Nov	0.025376	0.0390972	0.65	1.0000	-0.102605	0.153358
Jan	Dec	0.019355	0.0387754	0.50	1.0000	-0.107573	0.146283
Feb	Mar	-0.044754	0.0397078	-1.13	0.9935	-0.174735	0.085226
Feb	Apr	-0.116475	0.0400221	-2.91	0.1382	-0.247484	0.014534
Feb	May	-0.157658	0.0397078	-3.97	0.0043*	-0.287638	-0.027677
Feb	Jun	-0.428975	0.0400221	-10.72	<.0001*	-0.559984	-0.297966
Feb	Jul	-0.417335	0.0397078	-10.51	<.0001*	-0.547316	-0.287355
Feb	Aug	-0.415722	0.0397078	-10.47	<.0001*	-0.545703	-0.285742
Feb	Sep	-0.367287	0.0402702	-9.12	<.0001*	-0.499108	-0.235466
Feb	Oct	-0.204505	0.0399418	-5.12	<.0001*	-0.335252	-0.073759
Feb	Nov	-0.056475	0.0400221	-1.41	0.9615	-0.187484	0.074534
Feb	Dec	-0.062496	0.0397078	-1.57	0.9184	-0.192477	0.067484
Mar	Apr	-0.071720	0.0390972	-1.83	0.7988	-0.199702	0.056261
Mar	May	-0.112903	0.0387754	-2.91	0.1377	-0.239831	0.014025
Mar	Jun	-0.384220	0.0390972	-9.83	<.0001*	-0.512202	-0.256239
Mar	Jul	-0.372581	0.0387754	-9.61	<.0001*	-0.499509	-0.245652
Mar	Aug	-0.370968	0.0387754	-9.57	<.0001*	-0.497896	-0.244040
Mar	Sep	-0.322532	0.0393511	-8.20	<.0001*	-0.451345	-0.193720
Mar	Oct	-0.159751	0.0390150	-4.09	0.0026*	-0.287463	-0.032038
Mar	Nov	-0.011720	0.0390972	-0.30	1.0000	-0.139702	0.116261
Mar	Dec	-0.017742	0.0387754	-0.46	1.0000	-0.144670	0.109186
Apr	May	-0.041183	0.0390972	-1.05	0.9964	-0.169164	0.086799
Apr	Jun	-0.312500	0.0394164	-7.93	<.0001*	-0.441526	-0.183474
Apr	Jul	-0.300860	0.0390972	-7.70	<.0001*	-0.428842	-0.172879
Apr	Aug	-0.299247	0.0390972	-7.65	<.0001*	-0.427229	-0.171266
Apr	Sep	-0.250812	0.0396682	-6.32	<.0001*	-0.380663	-0.120961
Apr	Oct	-0.088030	0.0393348	-2.24	0.5220	-0.216790	0.040729
Apr	Nov	0.060000	0.0394164	1.52	0.9345	-0.069026	0.189026
Apr	Dec	0.053978	0.0390972	1.38	0.9672	-0.074003	0.181960
May	Jun	-0.271317	0.0390972	-6.94	<.0001*	-0.399299	-0.143336
May	Jul	-0.259677	0.0387754	-6.70	<.0001*	-0.386606	-0.132749
May	Aug	-0.258065	0.0387754	-6.66	<.0001*	-0.384993	-0.131136
May	Sep	-0.209629	0.0393511	-5.33	<.0001*	-0.338442	-0.080816
May	Oct	-0.046848	0.0390150	-1.20	0.9889	-0.174560	0.080865
May	Nov	0.101183	0.0390972	2.59	0.2874	-0.026799	0.229164
May	Dec	0.095161	0.0387754	2.45	0.3702	-0.031767	0.222090
Jun	Jul	0.011640	0.0390972	0.30	1.0000	-0.116342	0.139621
Jun	Aug	0.013253	0.0390972	0.34	1.0000	-0.114729	0.141234
Jun	Sep	0.061688	0.0396682	1.56	0.9245	-0.068163	0.191539
Jun	Oct	0.224470	0.0393348	5.71	<.0001*	0.095710	0.353229
Jun	Nov	0.372500	0.0394164	9.45	<.0001*	0.243474	0.501526
Jun	Dec	0.366478	0.0390972	9.37	<.0001*	0.238497	0.494460
Jul	Aug	0.001613	0.0387754	0.04	1.0000	-0.125315	0.128541
Jul	Sep	0.050048	0.0393511	1.27	0.9825	-0.078765	0.178861
Jul	Oct	0.212830	0.0390150	5.46	<.0001*	0.085117	0.340542
Jul	Nov	0.360860	0.0390972	9.23	<.0001*	0.232879	0.488842
Jul	Dec	0.354839	0.0387754	9.15	<.0001*	0.227910	0.481767
Aug	Sep	0.048435	0.0393511	1.23	0.9865	-0.080377	0.177248
Aug	Oct	0.211217	0.0390150	5.41	<.0001*	0.083504	0.338930
Aug	Nov	0.359247	0.0390972	9.19	<.0001*	0.231266	0.487229
Aug	Dec	0.353226	0.0387754	9.11	<.0001*	0.226298	0.480154
Sep	Oct	0.162782	0.0395872	4.11	0.0024*	0.033196	0.292367
Sep	Nov	0.310812	0.0396682	7.84	<.0001*	0.180961	0.440663
Sep	Dec	0.304790	0.0393511	7.75	<.0001*	0.175978	0.433603
Oct	Nov	0.148030	0.0393348	3.76	0.0095*	0.019271	0.276790
Oct	Dec	0.142009	0.0390150	3.64	0.0149*	0.014296	0.269721
Nov	Dec	0.006000	0.0390972	0.15	1.0000	-0.134000	0.131000

Figure C. 2

Level					Mean
Jun	A				0.66083333
Jul	A				0.64919355
Aug	A				0.64758065
Sep	A				0.59914530
Oct		B			0.43636364
May		B	C		0.38951613
Apr		B	C	D	0.34833333
Jan		B	C	D	0.31370968
Dec			C	D	0.29435484
Nov			C	D	0.28833333
Mar			C	D	0.27661290
Feb				D	0.23185841

Level						Mean
Jul	A					0.57741935
Jun	A					0.56583333
Aug	A	B				0.53548387
Sep		B				0.44957265
May			C			0.24919355
Apr			C	D		0.18583333
Oct			C	D	E	0.14710744
Mar				D	E	0.12177419
Jan				D	E	0.09274194
Feb					E	0.06902655
Nov					E	0.06500000
Dec					E	0.06422764

Figure C. 3

Bibliography

- Abramson, B., & Clemen, R. (1995). Probability Forecasting. *International Journal of Forecasting*, 11(1), 1–4. [https://doi.org/10.1016/0169-2070\(94\)02000-F](https://doi.org/10.1016/0169-2070(94)02000-F)
- Agresti, A. (2010). Categorical Data Analysis. (3rd ed.). Hoboken, NJ: John Wiley & Son Inc. <https://doi.org/10.1016/B978-0-08-044894-7.01311-7>
- Altman, D. G., & Royston, P. (2006). The Cost of Dichotomising Continuous Variables. *BMJ*, 332(7549), 1080. <https://doi.org/10.1136/bmj.332.7549.1080>
- AMS (American Meteorological Society). (2008). Enhancing Weather Information with Probability Forecasts. *Bulletin of the American Meteorological Society*, 83, 450–452.
- Anderson, J. M., Heaton, P., & Carroll, S. J. (2010). *The U.S. Experience with No-Fault Automobile Insurance*. Santa Monica, CA: RAND Corporation. Retrieved from <https://www.rand.org/pubs/monographs/MG860.html>
- Brown, B. G., Hamill, A. T., Wilson, L., & Fowler, T. (2015). Verification of Ensembles [PowerPoint Slides]. Retrieved from https://dtcenter.org/met/users/docs/presentations/MET_Tutorial_20150202/12_Ensemble_Feb15.pdf
- Burns, N., Kinder, D. R., Rosenstone, S. J., Sapiro, V., & Inter-university Consortium for Political and Social Research. (2016). ANES 2000 Time Series Study. Ann Arbor, MI. <https://doi.org/10.388>
- Dawson, N. V., & Weiss, R. (2012). Dichotomizing Continuous Variables in Statistical Analysis: A Practice to Avoid. *Medical Decision Making*, 32(2), 225–226. <https://doi.org/10.1177/0272989X12437605>
- Ferro, C. A. T., & Fricker, T. E. (2012). A Bias-Corrected Decomposition of the Brier Score. *Quarterly Journal of the Royal Meteorological Society*, 138(668), 1954–1960. <https://doi.org/10.1002/qj.1924>
- Fowler, T. L., Jensen, T. L., & Brown, B. G. (2012). Introduction to Forecast Verification. [PowerPoint Slides]. Retrieved from https://dtcenter.org/met/users/docs/presentations/WRF_Users_2012.pdf
- Harms, D. E., Guiffrida, A. A., Boyd, B. F., Gross, L. H., Strohm, G. D., Lucci, R. M., Weems, J. W., Priselac, E. D., Lammers, K., Herring, H. C., & Merceret, F. J. (1999). The Many Lives of a Meteorologist in Support of Space Launch, *8th Conference On Aviation, Range, and Aerospace Meteorology*, 10-15 Jan 99, Dallas, TX, 5-9.

- Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. (2nd ed.). Chichester, West Sussex: John Wiley & Son Ltd. <https://doi.org/10.1016/j.ijforecast.2005.11.002>
- Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A Misleading Measure of the Performance of Predictive Distribution Models. *Global Ecology and Biogeography*, 17(2), 145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Mason, I. (1979). On Reducing Probability Forecasts to Yes/No Forecasts. *Monthly Weather Review*, 107(2), 107-108. [https://doi.org/10.1175/1520-0493\(1979\)107<0207:ORPFTY>2.0.CO;2](https://doi.org/10.1175/1520-0493(1979)107<0207:ORPFTY>2.0.CO;2)
- Olson, R. H. (1965). On the use of Bayes' Theorem in Estimating False Alarm Rates. *Monthly Weather Review*, 93(9), 557–558. [https://doi.org/10.1175/1520-0493\(1965\)093<0557:OTUOBT>2.3.CO;2](https://doi.org/10.1175/1520-0493(1965)093<0557:OTUOBT>2.3.CO;2)
- Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The Precision-Recall Curve Overcame the Optimism of the Receiver Operating Characteristic Curve in Rare Diseases. *Journal of Clinical Epidemiology*, 68(8), 855–859. <https://doi.org/10.1016/j.jclinepi.2015.02.010>
- Roeder, W. (2017). Lightning Safety: What You Need to Know. Retrieved from <https://www.patrick.af.mil/News/Article-Display/Article/1208197/lightning-safety-what-you-need-to-know/>
- Roeder, W., Huddleston, L., Bauman, W. & Doser, K. (2014). Weather Research Requirements to Improve Space Launch from Cape Canaveral Air Force Station and NASA Kennedy Space Center. *Space In Traffic Management Conference*. Daytona Beach, FL
- Saito, T., & Rehmsmeier, M. (2017). Precrec: Fast and Accurate Precision-Recall and ROC Curve Calculations in R. *Bioinformatics*, 33(1), 145–147. <https://doi.org/10.1093/bioinformatics/btw570>
- Schwartz, C. S. (2016). A Comparison of Methods Used to Populate Neighborhood-Based Contingency Tables for High-Resolution Forecast Verification. *Weather and Forecasting*, 32(2), 733-741. <https://doi.org/10.1175/WAF-D-16-0187.1>
- Tamilvanan, B., & Bhaskaran, V. M. (2017). A New Feature Selection Techniques Using Genetics Search and Random Search Approaches For Breast Cancer. *Biosciences Biotechnology Research Asia* 14(1), 409–414. <http://dx.doi.org/10.13005/bbra/2459>

- Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining*. (2nd ed.). New York City, NY: Pearson
- Vêncio, R. Z. N., & Shmulevich, I. (2007). ProbCD: Enrichment Analysis Accounting for Categorization Uncertainty. *BMC Bioinformatics*, 8, 383.
<https://doi.org/10.1186/1471-2105-8-383>
- Wilks, D. S. (2011). *Statistical Methods in Atmospheric Sciences*. (3rd ed.). Amsterdam: Elsevier
- WMO. (2014). *Forecast Verification for the African Severe Weather Forecasting Demonstration Projects*. Geneva, Switzerland: World Meteorological Organization
- Youden, W. J. (1950). Index for Rating Diagnostic Tests. *Cancer*, 3(1), 32–35.
[https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)
- Zhang, H., & Casey, T. (2000). Verification of Categorical Probability Forecasts. *Weather and Forecasting*, 15(1), 80–89. [https://doi.org/10.1175/1520-0434\(2000\)015<0080:VOCPF>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0080:VOCPF>2.0.CO;2)

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 21-03-2019		2. REPORT TYPE Master's Thesis		3. DATES COVERED (From - To) Oct 2017 - Mar 2019	
4. TITLE AND SUBTITLE Effect of Using Probabilistic Contingency Tables to Modify Forecast Predictions				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB, OH 45433-7765				8. PERFORMING ORGANIZATION REPORT NUMBER AFIT-ENS-MS-19-M-117	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Mr. William Roeder 45 WS/SYS 1201 Edward H. White II St. MS 7302 Patrick AFB, FL 32925-3238 DSN 467-8410 Email: William.roeder@us.af.mil				10. SPONSOR/MONITOR'S ACRONYM(S) 45 WS	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.					
13. SUPPLEMENTARY NOTES					
14. 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 implemented verification process. For deterministic outcomes, weather forecast analysis typically utilizes a Tradition 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 2015-2018 45 WS data were utilized to compare forecast performance metrics produced from the TCT and PCT to determine which verification tool better supports producing the greatest quality forecasts. Comparisons of frequency bias, reliability, and Brier Score (BS) computed from both dichotomized and continuous forecasts revealed misrepresentative performance metrics from the TCT as well as a loss of information necessary for verification. Exploration of the 45 WS data with a probabilistic verification process revealed a need to verify seasonally as well as slightly unreliable warm season lightning event forecasts.					
15. SUBJECT TERMS Contingency Table, Weather Verification, Dichotomization, Probabilistic					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 97	19a. NAME OF RESPONSIBLE PERSON Dr. Edward White, AFIT/ENC
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code) (937) 255-3636 x4540; Edward.white@afit.edu