Reply

KEITH F. BRILL

Hydrometeorological Prediction Center, National Centers for Environmental Prediction, Camp Springs, Maryland

FEDOR MESINGER

Environmental Modeling Center, National Centers for Environmental Prediction, Camp Springs, and Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

(Manuscript received and in final form 26 August 2010)

1. Introduction

Pasarić and Juras (2011, hereinafter PJ) pose a persistence-based quantitative precipitation forecast scenario to evaluate the bias-adjusted equitable threat score (ETS) introduced by Mesinger (2008) and examined by Brill and Mesinger (2009, hereinafter BM), who apply the critical performance ratio (CPR) analysis described by Brill (2009). The study by PJ applies the tetrachoric correlation coefficient (TCC) and is of considerable interest. As an aid to understanding, we would appreciate a fourth panel for Fig. 2 of PJ, showing the probability of detection (POD) in the bias–TCC plane. Having considered the discussion proffered by PJ, some important clarifying remarks are in order. They are presented in two sections below. Section 2 clarifies the concept of the frequency bias adjustment for the ETS in terms of the intention and motivation for its development. Section 3 endeavors to elucidate matters concerning the frequency bias sensitivity or lack thereof exhibited by the TCC.

2. Bias-adjusted ETS

While we welcome the interest of PJ in our bias-adjusted ETS and its discussion in BM, we are puzzled by their criticism of its alleged weaknesses. A misunderstanding is clearly in place. Namely, it was not the intention of Mesinger (2008) to ''obtain a score that better assesses forecast quality'' (PJ); the intention was to obtain a score that, better than the ETS does, assesses the accuracy of *placing* the forecast event.

Corresponding author address: Keith F. Brill, NCEP/HPC, W/NP32, NOAA Science Center, Rm. 410B-2, 5200 Auth Rd., Camp Springs, MD 20746-4304.
E-mail: keith.brill@noaa.gov

DOI: 10.1175/2010WAF2222459.1

The two are not the same. Imagine a situation of, say, 10 realizations of an event, such as rain above a chosen threshold, and a forecasting system that predicted only a single realization, but one that ''verified'' as matching the observed conditions. This would not be a very successful forecast; its frequency bias (hereinafter referred to as ''bias'') would be only 0.1, and its ETS would be small, depending on the base rate associated with the event. Yet, with the placing of that single realization being perfect, the intention of the bias adjustment of ETS was to accord such a forecast a perfect bias-adjusted ETS: a value of 1.

The combination of the two performance measures, bias of 0.1 and bias-adjusted ETS of 1 or close to 1, we find describes well the skill of this hypothetical forecast. This would not have been that well achieved with the bias and the unadjusted ETS and would be even worse if done with just the ETS alone. In other words, we agree with PJ that, relative to striving for a single measure that would ''encompass all aspects of forecast quality,'' it is a better approach ''to use scores that focus on a single aspect,'' such as, for example, Mesinger (2008) has done with the recommendation for the use of the bias-adjusted ETS along with bias.

The extent to which the bias adjustment of Mesinger (2008) is successful in achieving its intended objective of assessing the skill of the *placement* of predicted events in comparison with other possible measures, including the TCC, is a question of obvious interest. Note that the *motivation* for developing the bias adjustment scheme of ETS was not the idea that the ETS is perhaps an optimal performance measure in some way but rather was its widespread use. Thus, the objective was to arrive at a score that would have a clear meaning on its own but at the same time in cases of bias equal to 1 would have the same value as the unadjusted ETS, a score in many cases

used so far. Note the very recent words of Hogan et al. (2010): ''ETS is now one of the most widely used verification measures in meteorology.''

This widespread use is typically not a result of the conviction on the part of the users of ETS that its properties are superior to properties of other possible scores but rather is a result of history. There are numerous scores that for the same forecasting outcome would give numerically higher values, such as two of the PJ examples. Given that it is the comparison of scores in relative terms that just about always takes place, this is apparently generally accepted as not really a concern. What we feel is of much greater importance is a clear meaning of the score used and its resistance to hedging so that a ''better'' value is not easily obtained by increasing an error of a forecast. It is along these lines that we feel improvements to the bias-adjusted ETS or other scores are worth researching.

3. Frequency bias sensitivity

The work of Brill (2009) shows that performance metrics computed using the entries in a 2×2 contingency table are sensitive to arbitrary changes in bias, such as might happen when a human forecaster hedges in hopes of tilting some score in a favorable direction (such changes associated with hedging are usually not random). From the Brill (2009) analysis, one can infer the possibility that, with the proper tandem changes in POD and bias, a score could maintain a constant value, if one assumes a constant base rate. For example, changes following a typical relative operating characteristic (ROC) curve would likely preserve the value of the odds-ratio skill score (ORSS), making it appear to be insensitive to bias (see Stephenson 2000). An arbitrary change in bias could shift the 2×2 contingency table to a different ROC curve, altering the ORSS and revealing it to have sensitivity to bias.

Because the TCC is computed from the 2×2 contingency table, one would expect it to have sensitivity to arbitrary changes in bias. In the case of the persistence forecast method used in the analysis of PJ, the change of threshold required to obtain the much underbiased forecast would be expected to preserve the TCC, which attempts to estimate the coefficient of linear correlation between assumed latent continuous variables representing the forecast and observed values associated with the 2×2 contingency table (see Juras and Pasarić 2006). When a persistence forecast is utilized, the latent variables actually exist and are the observations lagged by the forecast projection time and the observations serving as the forecast and observed variables, respectively. For the first set of forecasts of PJ, which are overbiased, the TCC seeks to estimate the 1-day lagged correlation in the observed time series with 0.1 mm subtracted from the values at the lag time. For the second, much-underbiased, set of forecasts, the TCC seeks to estimate the 1-day lagged correlation in the observed time series with 10 mm subtracted from the values at the lag time. Because the two latent variables representing the forecast values in this experiment are offset from each other by a constant, the coefficient of linear correlation with the observations should be the same. The constant offset induces a marked change in frequency bias for the 2×2 contingency table without altering the linear correlation, which is estimated by the TCC. So, the fact that the TCC is little changed in this experiment does not really testify to its bias insensitivity, but it does indicate the robustness of the Pearson (1900) correlation estimation technique. Expressed a different way, this is another case for which appropriate tandem changes in POD and bias approximately preserve the value of a score.

A general attempt to reveal the nature of the bias sensitivity of the TCC would require calculation of the CPR values defined by Brill (2009). The CPR is the change in POD per change in bias that keeps a score constant. In other words, the CPR quantifies the tandem change required to preserve the value of a score, given a constant base rate. If the CPR is exceeded because of a change in POD and inflated bias, then the score will indicate an improved forecast. If bias is decreased, then the change in POD per change in bias must remain below the CPR value for the score to indicate an improved forecast (see Brill 2009). For scores that are analytic functions of the elements of the 2×2 contingency table, the CPR is the ratio of two partial derivatives that are also analytic functions. For the TCC, the necessary partial derivatives may need to be approximated by finite-difference calculations using the form of the contingency table given by Table 1B of Brill (2009). This computationally intensive exercise is likely to yield interesting results, and PJ are encouraged to pursue such a study in future work. Furthermore, the placement of the TCC CPR values relative to benchmarks for the CPR (see Brill and Pyle 2010) would be of great interest as well. It may be important to note that the CPR analysis makes no assumptions about the existence of any latent variables underlying the 2×2 contingency table. It assumes only that the 2×2 contingency table elements vary continuously from one realization of the table to another, much as temperature and pressure are assumed to vary continuously from one thermodynamic state to another in classical theory.

REFERENCES

Brill, K. F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting,* **24,** 307–318.

——, and F. Mesinger, 2009: Applying a general analytic method for assessing bias sensitivity to bias-adjusted threat and equitable threat scores. *Wea. Forecasting,* **24,** 1748–1754.

——, and M. Pyle, 2010: The response of performance metrics for binary forecasts to hedging that approaches random change. *Wea. Forecasting,* **25,** 1307–1314.

Hogan, R. J., A. T. F. Christopher, I. T. Jolliffe, and D. B. Stephenson, 2010: Equitability revisited: Why the ''equitable threat score'' is not equitable. *Wea. Forecasting,* **25,** 710–726.

Juras, J., and Z. Pasarić, 2006: Application of tetrachoric and polychoric correlation coefficients to forecast verification. *Geofizika,* **23,** 59–82.

Mesinger, F., 2008: Bias adjusted precipitation threat scores. *Adv. Geosci.,* **16,** 137–142.

Pasarić, Z., and J. Juras, 2011: Comments on ''Applying a general analytic method for assessing bias sensitivity to bias-adjusted threat and equitable threat scores.'' *Wea. Forecasting,* **26,** 122–125.

Pearson, K., 1900: Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. Roy. Soc. London,* **A195,** 1–47.

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.