CORRESPONDENCE

Comments on "Applying a General Analytic Method for Assessing Bias Sensitivity to Bias-Adjusted Threat and Equitable Threat Scores"

ZORAN PASARIĆ AND JOSIP JURAS

Geophysical Department, Faculty of Science, University of Zagreb, Zagreb, Croatia

(Manuscript received 23 June 2010, in final form 12 August 2010)

The recent paper by Brill and Mesinger (2009, hereinafter BM) addresses difficulties encountered in practice when assessing the quality of weather forecasts in the presence of bias. Despite new approaches to forecast verification (e.g., Casati et al. 2008), utilization of contingency tables, which arose more than 100 years ago (Murphy 1996), continues to be an area of intensive research. Even in cases in which forecasts and respective observations are condensed into the simplest possible contingency table (2×2), there is little agreement on how to express forecast quality. Nonetheless, there is a clear need to express forecast quality, if possible, as a single number. This would enable easy comparison among different forecasting systems, as well as continuous monitoring of development of such systems. Numerous scores have been proposed (e.g., Daan 1984; Murphy 1996), and the meteorological community is far from consensus on which of them to use. The key difficulty lies in the well-established fact that forecast verification is a multifaceted problem, and each score measures the various facets of forecasts in its own way. Some scores tolerate or even award bias in special cases, whereas other scores penalize bias. It ultimately appears impossible to condense the many facets of some set of forecasts into a single score, and nearly all known measures of forecast quality are still in everyday use. Furthermore, new measures continue to appear in the scientific literature (e.g., Stephenson 2000; Stephenson et al. 2008; Rodwell et al. 2010).

Which facets of forecasts are essential and which are of secondary importance? Agreement in such considerations is difficult to achieve, and attempts to answer such a question can be bypassed by adopting the Murphy and Winkler (1987) approach, which involves the analysis of the joint distribution of forecasts and observations. Although the joint distribution carries all of the statistical information about a forecasting system and enables the full exploration of that system, it is too complicated to be practical for continual monitoring of forecasts. Murphy (1993; see his Table 2) listed 10 characteristics of forecasts; from this list we take three with respect to their importance: association, bias, and uncertainty. In the case of continuous variables, association is measured by the Pearson correlation coefficient. For binary variables, association is partially assessed by various scores based on the contingency table. Bias is expressed as $B = P_f/P_o$, where P_f and P_o are the frequencies of forecasting and observing the event, respectively. Uncertainty is expressed by P_o (also called the base rate), which in the case of binary variables determines the probability distribution of dichotomized observations.

It is obvious that if a certain measure depends on all three facets then it is of little practical value to either users or the forecaster; BM correctly assess the equitable threat score (ETS) based on that reasoning. The ETS is influenced by all three of the facets listed above (Mason 1989), and it consequently cannot be used by itself to compare different forecasting systems. Furthermore, several papers dealing with 2×2 contingency tables (BM and references therein) attempt to eliminate the influence of bias on the ETS, with the goal of inventing new scores that will describe forecast quality more truly. The so-called dHdA method is proposed by Mesinger (2008). The resulting score (hereinafter referred to as "adjusted ETS") is cited as a good attempt at a more accurate ETS, but the proposed method is relatively complicated, and we do not believe that it will achieve wide application in

Corresponding author address: Dr. Z. Pasarić, Geophysical Department, Faculty of Science, University of Zagreb, Horvatovac 95, 10000 Zagreb, Croatia.
E-mail: pasaric@irb.hr

DOI: 10.1175/2010WAF2222453.1

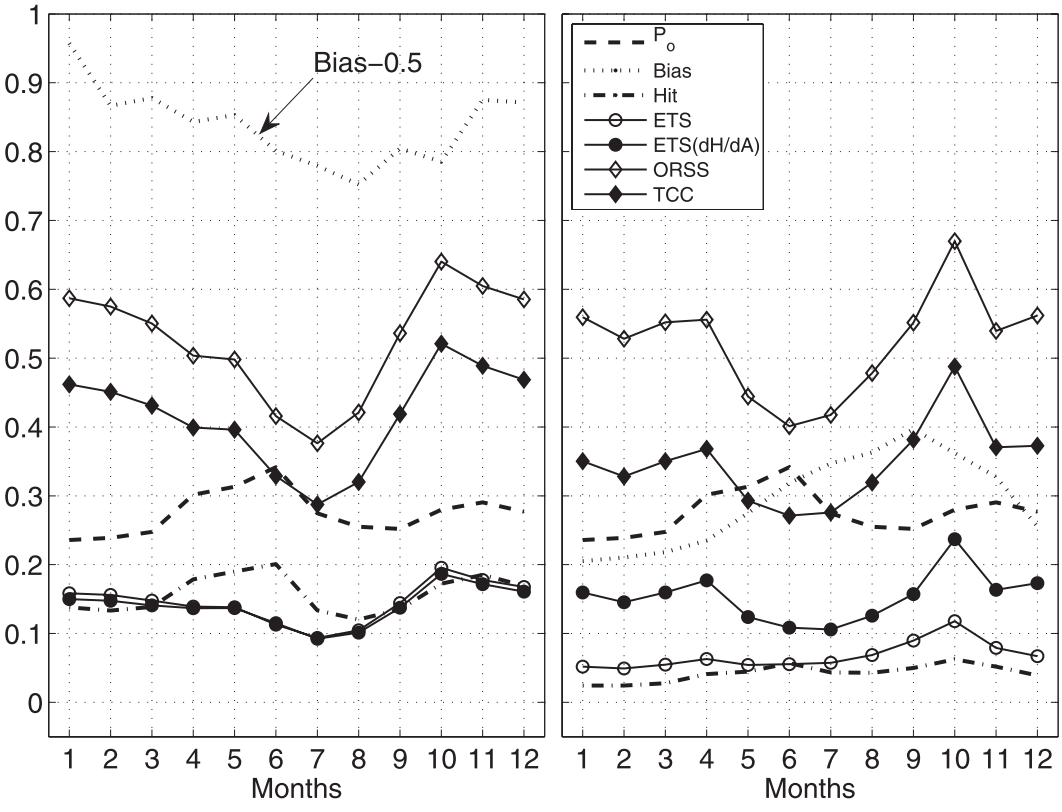

FIG. 1. Verification parameters and some scores for persistence forecasts of daily precipitation greater then 1 mm for Zagreb/Grič during the period 1870–2000. Predictor is the day-before amount of precipitation exceeding (left) 0.1 mm (case 1) or (right) 10 mm (case 2). To facilitate plotting, 0.5 is subtracted from the bias in the left panel.

practice. Moreover, the method possesses certain weaknesses that are best illustrated with an example.

The example is based on the time series of daily precipitation for the Zagreb/Grič meteorological station (45°48′53″N, 15°58′20″E) during the period 1870–2000. The event is defined as a day with total precipitation greater than 1 mm. Two sets of persistence forecasts are constructed separately for each month: the event is forecast if precipitation for the previous day was greater than 0.1 mm (case 1) or greater than 10 mm (case 2). The first set of forecasts carries a bias of $B > 1$, thus simulating overforecasting, and the second one carries a bias of $B < 1$, simulating underforecasting. It is seen (Fig. 1) that the frequency of days with precipitation greater than 1 mm does not vary much throughout the year.

In the first case (Fig. 1, left panel), the same is valid for bias, whereas the hit rate follows the base rate. The tetrachoric correlation coefficient (TCC; Juras and Pasarić 2006) is relatively higher in the colder part of the year, which is characterized by large-scale atmospheric processes. The minimum TCC occurs during summer and is influenced by mesoscale processes of shorter duration. It is interesting that the TCC and the odds-ratio skill score (ORSS; Stephenson 2000) similarly describe the strength of the relationship between wet and dry days in the time series. The ETS also reaches a minimum in summer and a maximum in autumn, but its numerical values are much lower and could be misleading, especially for a layperson. It appears that there is no real reason to prefer the ETS over other scores. In the case presented here, the ETS adjusted for bias is almost equivalent to the original ETS because of its relatively small bias. In fact, the adjusted ETS values are slightly lower than the original values; this point will be explained later.

In the second case (Fig. 1, right panel), the system is substantially underforecast, with bias ranging between 0.2 in January and 0.4 in September. The hit rate is much lower than that of case 1 because of the small number of days with total precipitation greater than 10 mm. The TCC and, in particular, ORSS values do not exhibit a commensurate decrease, indicating that these measures are not strongly influenced by bias. In contrast, ETS values are nearly halved, primarily because of the strong bias. It appears that adjusting the ETS corrects the problem of underforecasting; however, in some

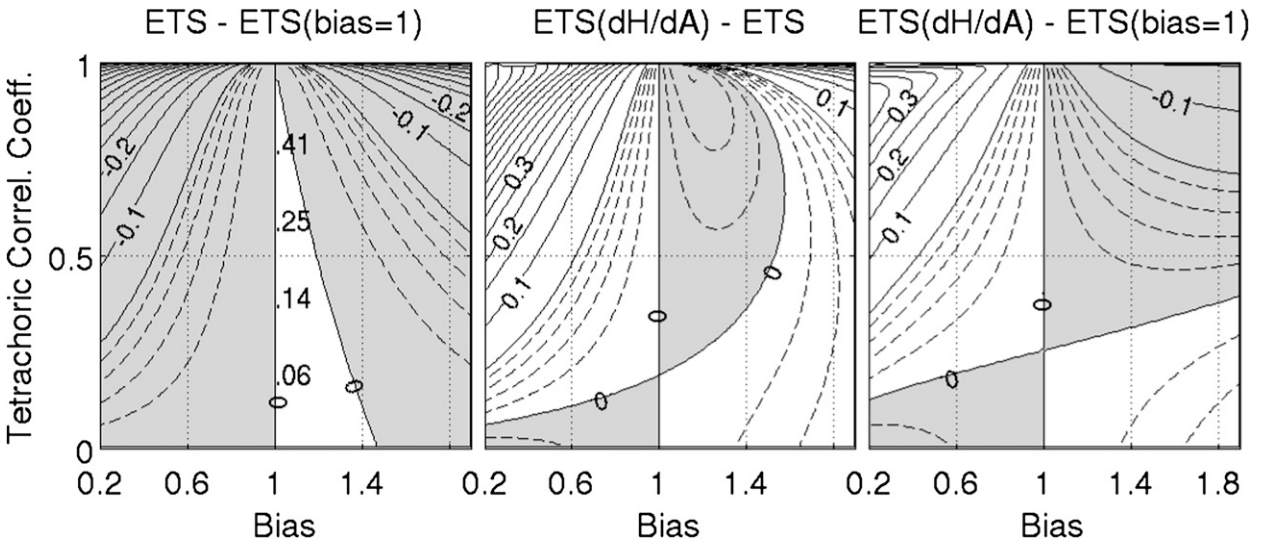

FIG. 2. The difference between (left) ETS and its value when bias = 1, (middle) the ETS adjusted for bias (dHdA method; Mesinger 2008) and the ETS itself, and (right) the adjusted ETS and the ETS itself when bias = 1, as functions of the TCC and bias for the fixed base rate $P_o = 0.3$. The isolines are plotted with steps of 0.05 (full lines) and 0.01 (dashed lines). Negative values are shaded in gray. Numbers along the middle line in the left panel denote the ETS values for respective values of TCC with bias = 1.

cases, the adjusted values are 2 times the original scores. The adjustment of ETS leads to the unexpected conclusion that the relationship between wet and dry days in the time series is stronger in the second forecast scenario than in the first one. The ORSS and TCC values suggest a different, more intuitive, conclusion.

To gain further insight into these matters, it is instructive to analyze more generally how the ETS is influenced by bias and what is the effect of the adjustment proposed by Mesinger (2008) and BM. To that end, we fix the base rate at $P_o = 0.3$ (to be consistent with the previous example) and express various quantities as functions of bias and the tetrachoric correlation, the latter measuring only the association between forecasts and observations. In Fig. 2, the left panel depicts the difference between ETS and its values obtained for the same TCC but with bias equal to 1. The ETS penalizes both underforecasting and overforecasting, although there is a narrow area of positive differences, which indicates that maximum values of the ETS occur where biases are greater than 1. In certain situations, a number of other scores also favor biases that significantly differ from 1 (see Juras and Pasarić 2006). Note also the large area in the bias–TCC plane in which changes in the ETS due to bias do not surpass ± 0.05. The central idea of Mesinger (2008) and BM was to compensate for the penalty due to bias and to obtain a score that better assesses forecast quality. In general, the strong effect of bias on the ETS is successfully corrected using the dHdA method (Fig. 2, middle panel); however, the correction can yield unrealistically high values of adjusted scores when $B < 1$. In contrast, when $B > 1$, corrected values are too low; in some cases (e.g., when $B = 1.4$), adjustment of ETS values can result in an even more pronounced effect of bias. As a consequence, it could be expected that the adjusted ETS will favor more often underforecasting than overforecasting; this effect should be more pronounced with the high values of TCC that are common in today's operational forecasts.

In summary, we used examples to highlight some undesirable properties of the adjusted ETS. Because no single measure can encompass all aspects of forecast quality, a better approach would be to use scores that focus on a single aspect, independent of the others. One possible option is to quantify the association, bias, and uncertainty, which is enough to reconstruct the original 2×2 contingency table. Among them the only facet that is not trivial to assess is association, which we believe can best be expressed in most cases by the tetrachoric correlation coefficient.

Acknowledgments. The work was supported by the Croatian Ministry of Science, Education and Sports (Grants 119-1193086-3085 and 119-1193086-1323).

REFERENCES

Brill, K. F., 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.

Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.,* **15,** 3–18.

Daan, H., 1984: Scoring rules in forecast verification. WMO PSMP Publication Series No. 4, 62 pp.

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

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

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

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Rodwell, M. J., D. S. Richardson, and T. D. Hewson, 2010: A new equitable score suitable for verifying precipitation in NWP. ECMWF Tech. Memo. 615, 35 pp.

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

——, B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.,* **15,** 41–50.