47

ANALYSIS OF SUBJECTIVE AND OBJECTIVE EVALUATIONS OF CALIBRATED TORNADO GUIDANCE IN THE HAZARDOUS WEATHER TESTBED

David E. Jahn[1,2], Eric Loken[1,3], Burkely Gallo[1,2], Chris Karstens[2,4], Brian Hempel[1,2]

[1]Cooperative Institute for Severe and High-Impact Weather Research and Operations/Univ. of Oklahoma
[2]Storm Prediction Center/National Weather Service/NOAA
[3]National Severe Storms Laboratory/Office of Oceanic and Atmospheric Research/NOAA
[4]School of Meteorology, University of Oklahoma, Norman, Oklahoma

1. Introduction

During the five-week Hazard Weather Testbed (HWT) Spring Forecasting Experiment (SFE) of 2022 (Clark et al. 2022), 80 individuals representing both operational forecasters and researchers subjectively evaluated a suite of seven calibrated tornado guidance methods against available local storm reports (LSRs) and associated practically perfect hindcasts (PPHs). Evaluations were conducted on the morning following the event for a roughly 300 km x 300 km domain judiciously placed to encompass significant convection for the given day. After the SFE, these same calibrated guidance methods were evaluated objectively as well for the same days and HWT-defined domains, using various commonly invoked methods including performance and receiver operating characteristic (ROC) diagrams as well as Brier skill score reliability component (Brier_rel).

It was found that the performance evaluations of the calibrated guidance methods based on objective versus subjective evaluations did not always concur. This study investigates reasons for incongruence among these evaluation methods. As such, this study ventures into the challenge of what constitutes a “good” forecast as well posed by Murphy (1993). Here the means for analyzing forecast “quality” or “skill”’ per objective metrics is analyzed as compared to forecast “value” as given by the subjective evaluations of SFE participants.

As part of this analysis of evaluation methods, a new objective parameter is formulated using a weighted sum of metrics based on the three objective verification methods (listed above) to provide a verification perspective that is consistent with the subjective evaluation, and to identify the degree to which these objective metrics measure skill in congruence with subjective evaluations.

The specific calibrated methods evaluated are generically represented here (methods A-G) because the purpose of this study is not necessarily to identify the best method, but rather to analyze the differences among subjective and objective evaluation approaches. In general, the calibrated guidance methods fall into two types: machine learning models that use various forecast storm and environmental fields as predictors, and a more traditional approach that is based on an identified correlation among the significant tornado parameter (STP, Gallo et al. 2018, Jahn et al. 2020) and observed tornado frequency (Thompson et al. 2012). An explicit description of the seven calibrated methods can be found at: https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE2022_operations_plan.pdf.

2. Method

HWT participants subjectively scored the calibrated methods on a scale of 1-10 (with 10 being best) comparing the areal coverage of tornado probability at set thresholds (2%, 5%, 10%, 15%) as compared to PPH tornado probabilities at the same thresholds. A PPH represents a density coverage of observed tornadoes and was calculated using a Gaussian filter with $\sigma = 1.5$ (~120 km; Hitchens et al. 2013). Subjective

* *Corresponding author address:* David E. Jahn, CIWRO/Univ. of OK, 120 David L. Boren Blvd., Norman, OK 73072; e-mail: djahn@ou.edu.

scores from HWT evaluators are averaged for each of the seven calibrated methods across 12 tornadic cases.

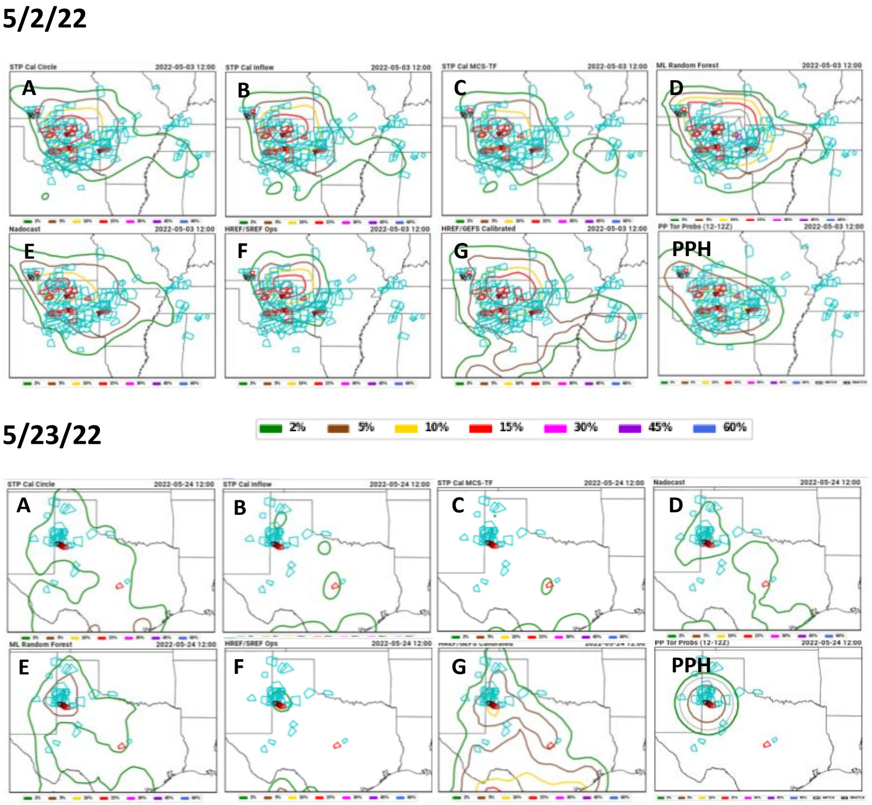

Fig. 1. Contours of tornado probabilities (magnitude by colors in legend) for calibrated guidance methods A-G for 5/2/22 (top plots) and 5/23/22 (bottom plots) with associated PPH.

Only cases with at least one tornado observation were considered (12 of 19 HWT cases) because of inconsistencies among SFE participants regarding their basis for subjective scoring of non-tornadic (null) cases. An example inconsistency is to what degree a guidance product should be penalized for forecasting greater than a 2% probability of tornado occurrence, given that there was no observed tornado but yet the conditions existed such that the possibility of tornadoes was not zero.

The calibrated methods are evaluated objectively using three metrics including the area to the left of the performance curve (Perf_ALC), the ROC area under the curve (ROC_AUC), as well as Brier_rel. To allow for direct comparison of subjective and objective results, the objective values are normalized such that their maximum value across all calibrated guidance products is equal to the maximum mean subjective score for the same case.

To keep the basis of evaluation the same for objective as for subjective evaluations, both considered data restricted to the same HWT domain as selected daily to encompass significant convective activity. Also, objective metrics were calculated using the same threshold levels (2%, 5%, 10%, 15%) as represented in the plot contours of tornado probabilities used for subjective evaluation. Based on comments from a large fraction of evaluators during the HWT, tornado warnings (even if not verified as observations) were heavily considered along with tornado local storm reports (LSRs) when evaluating calibrated methods subjectively. Thus, for consistency, objective metrics were calculated by treating tornado warnings as proxy tornado observations. Only LSRs available by HWT evaluation time were considered.

3. New objective parameter

As a means of investigating the basis for inconsistencies among the subjective and objective evaluations, a new objective parameter is formulated, which judiciously combines the calculated three objective metrics for a result that is more consistent with the subjective evaluation for a given case. For the following system of equations,

$$\begin{bmatrix} A_{11}\ A_{12}\ A_{13} \\ A_{21} A_{22}\ A_{23} \\ A_{31}\ A_{32}\ A_{33} \\ \dots \\ A_{N1}\ A_{N2}\ A_{N3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \dots \\ b_N \end{bmatrix}$$

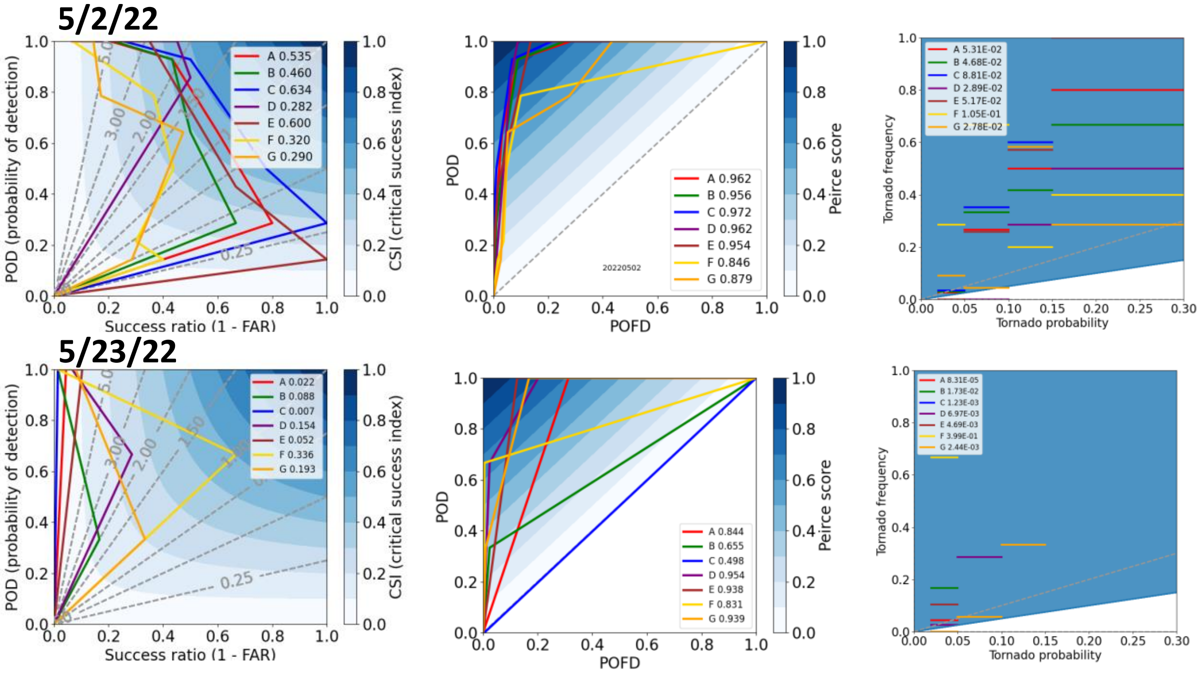

Fig. 2. For 5/2/22 (top row) and 5/23/22 (bottom row), performance (left column) and ROC (middle column) diagrams with associated Perf_ALC and ROC_AUC values as well as Brier_Rel (right column) by calibrated methods A-G and with specific values in the legends

the left matrix consists of variables A_{n1}, A_{n2}, and A_{n3} that represent values respectively of Perf_ALC, ROC_AUC, and Brier_Rel by calibrated method (n=1 through N for the seven methods, A-G) for one case. The variables b_n in the matrix on the right represent for the same case averaged subjective scores by calibrated method. The coefficients x_n are determined by solving $x = A^{-1}b$. Because this is an over-determined system, x_n values are identified as a best-fit, in a least-squares sense, to satisfy the system of equations. The x_n values can be interpreted as relative weights and provide insight as to the degree each of the objective metrics are consistent with the subjective evaluation for a given case. These x_n values are then used to calculate a new objective parameter, Ax.

4. Results for specific cases

Figure 1 shows the suite of tornado probability forecasts based on the seven calibrated methods (A-G) for example high-end (5/2/22) and low-end (5/23/33) HWT tornadic cases. Objective metrics are calculated for both cases to distinguish the performance of the seven different calibrated methods (A-G). Performance curves with the highest Perf_ALC are most influenced by points at relatively high thresholds as seen in the bottom right quadrant of the performance diagram (Fig.2, left column plots). The ROC curve, on the other hand, favors increased probability of detection at low thresholds (points in the upper left corner of the ROC diagram, Fig. 2, center column).

In calculating the new objective parameter for case 5/2/22 (one that concurs with the subjective scores for this case better than any one objective metric alone), a higher weighting coefficient, $x_2 = 0.56$, is calculated for ROC_AUC as compared to $x_1 = 0.38$ for Perf_ALC. This suggests that the performance criteria of ROC_AUC is closer than that of Perf_ALC to the criteria (both explicit and implicit) considered for the subjective evaluation. It could thus be interpreted that emphasis was given by evaluators first to method performance at low (e.g. 2%) thresholds followed by performance at higher thresholds (e.g., 10% and greater).

Based on evaluator comments, forecasts were favored for which the 2% contour encompassed all tornado observations or warnings even when the

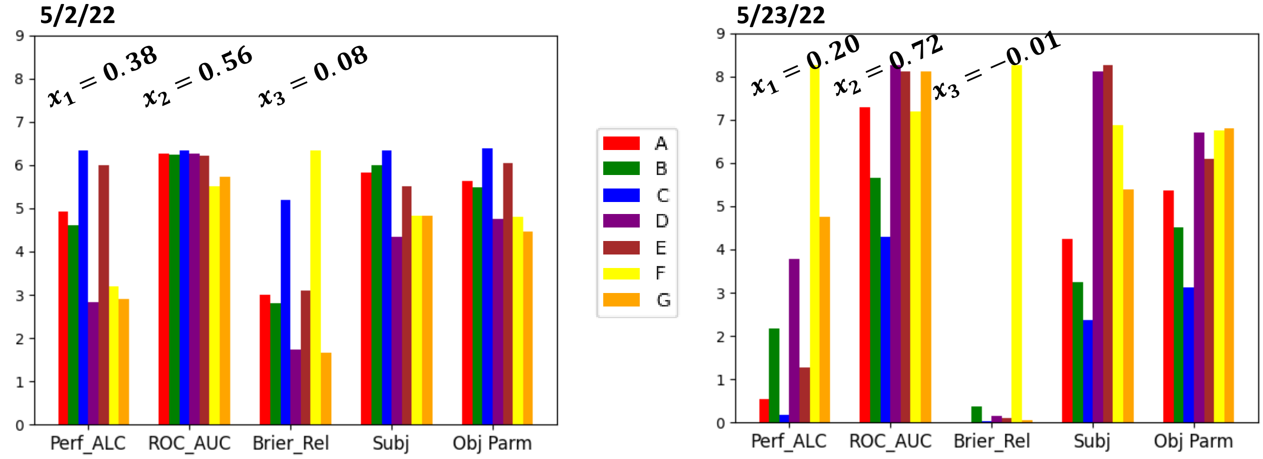

Fig. 3. For cases 5/2/22 (top) and 5/23/22 (bottom), histograms of the objective metrics Perf_ALC, ROC_AUC, and Brier-Rel (first three groups), subjective results (fourth group) and new objective parameter (Ax , fifth group, calculated using x_n as indicated). Calibrated methods A-G denoted by colors in legend.

false alarm ratio (FAR) was relatively large. This perspective is consistent with the results of this same case for methods F and G, which have similar ROC_AUC values (Fig. 2) but 2% contour footprints (FAR values) that are very different (Fig. 1). ROC_AUC favors high performance (high POD) at low threshold levels, but does not strongly penalize for a high FAR.

Turning to the results for 5/23/22, Fig. 1 shows that method A, having a large 2% probability contour that encompasses all observations but does so with a relatively large FAR, exhibits seemingly contradictory objective scores having one of the highest ROC_AUC scores but the lowest Perf_ALC value (Fig. 3). These results suggest that the performance diagram penalizes false detection more severely than the ROC diagram. In the process of generating a new objective parameter to concur with the subjective evaluation, the relatively low 0.20 weighting coefficient, x_1, attributed to Perf_AUC (Fig. 3) suggests that evaluators (as implied also for the 5/2/22 case) are less concerned about a high FAR as compared to achieving a high POD for this low-end case.

It should be noted that for both cases, the weighting coefficients of Brier_rel were relatively very low, indicating that forecast reliability was of less concern for subjective evaluation than POD or false detection, effects of which are innately represented by the other two objective metrics, ROC_AUC and Perf_ALC.

5. Results across full case set

Weighting coefficients, x_n, were calculated separately for each of the 12 tornadic cases (Fig. 4). The ROC_AUC objective parameter, having the highest x_n value in nearly all cases, has the highest overall influence in the formulation of the new object parameter. This is an indication that ROC_AUC conforms more closely than the other two objective metrics to the (implicit or explicit) criteria as used by evaluators to score subjectively the given set of cases. Conversely, except for one case, Brier_rel is consistently the lowest weighted metric thus indicating that reliability was less a factor for the subjective evaluations.

The weighting coefficients, x_{1-3}, are used to calculate new objective parameters by case, the mean values of which are given in Fig. 5 for each of the calibrated methods. The relatively narrow spread in the mean subjective scores across all calibrated methods is most consistent with the spread in mean values of ROC_AUC as well as the new objective parameter as compared to the relatively large spread in mean values of both Perf_ALC and Brier_rel. This result emphasizes once again that ROC_AUC is the dominant metric

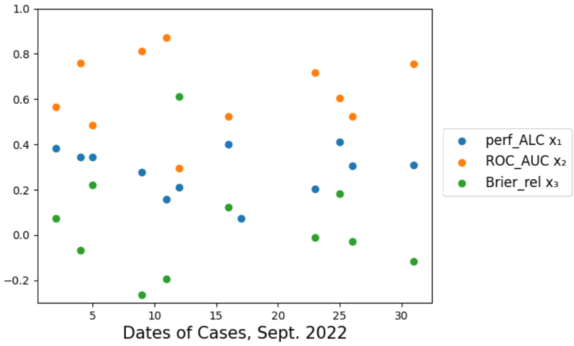

Fig. 4. Weighting coefficients for individual cases by date.

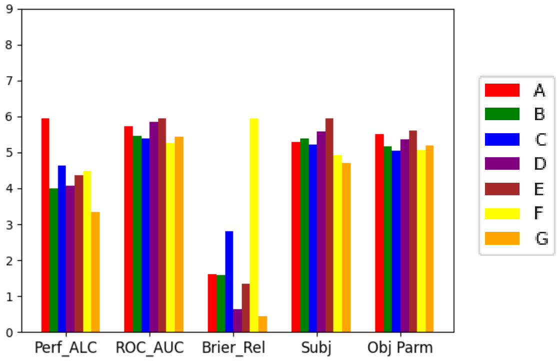

Fig. 5. Same as Fig. 3, but showing mean values across the 12 HWT tornadic cases.

influencing the new objective parameter and it emulates to a larger degree subjective evaluation criteria than do the other two objective metrics for the cases of this study.

6. Summary

As part of the post-HWT analysis, it was found that no one objective metric adequately reflects the collective subjective evaluation as given by participants for a given case. A new objective parameter is proposed here that incorporates a weighted aggregate measure of skill from three different objective metrics: Perf_ALC, ROC_AUC, and Brier_rel. The ROC_AUC is weighted higher because of its strong dependence on POD at relatively low thresholds and minimized penalty for FAR, criteria that are similar to those considered by HWT participants in their subjective evaluations. Method performance at higher thresholds (as favored by Perf_ALC) is still important, but has a secondary relevance. The influence of the Brier reliability component is minimized suggesting that forecast reliability is less a factor for subjective evaluation, as might be expected when participants consider only a few cases during their time in the SFE.

Future work will involve expanding the number of cases, such as from SFE 2021, for which subjective evaluations are available and data are accessible to calculate objective metrics. Consideration will also be given to formulate a new objective parameter based directly on the fundamental objective metrics of POD and FAR rather than on aggregate parameters Perf_ALC, ROC_AUC, and Brier_rel.

Acknowledgements

Funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA21OAR4320204, U.S. Dept. of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

References

Clark, A. and co-authors, 2022: The 3rd real-time, virtual spring forecast experiment to advance severe weather prediction capabilities. *Bull. Amer. Meteor. Soc.* Published on-line. https://doi.org/10.1175/BAMS-D-22-0213.1

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL-WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443-459.

Hitchens, N. M., H. E. Brooks and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525-534.

Jahn, D. E., B. T. Gallo, C. Broyles, B. T. Smith, I. Jirak, J. Milne, 2020: Refining CAM-based tornado probability forecasts using storm-inflow and storm-attribute information. 26th Conf. on Numerical Weather Pred., Boston, MA, Amer. Meteor. Soc.

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

Thompson, R. L., B. T. Smith, J. S. Grams, A. R. Dean, J. C. Picca, A. E. Cohen, E. M. Leitman, A. M. Gleason, and P. T. Marsh, 2017: Tornado damage rating probabilities derived from WSR-88D data. *Wea. Forecasting*, **32**, 1509-1528.