67

FORMULATION AND EVALUATION OF A CALIBRATED TORNADO FORECAST GUIDANCE ENSEMBLE

David E. Jahn[1,2], Bryan Smith[2], Israel Jirak[2], Chris Karstens[2], Eric Loken[1,3], Burkely Gallo[1,2], Tim Supinie[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

1. Introduction

Over the past few years, five different calibrated tornado guidance methods have been developed by various researchers associated with the Cooperative Institute for Severe and High-impact Weather Research and Operations (CWIRO) at the University of Oklahoma and/or the Storm Prediction Center (SPC) of the National Weather Service. These methods are based on High Resolution Ensemble Forecast (HREF) data and are either formulated using machine learning or utilize known relationships among specific forecast environmental variables and tornado frequency.

Although having a suite of guidance forecasts for a given day can provide a perspective on the uncertainty and range of possible outcomes, it was proposed that producing an ensemble average or percentile could be convenient for forecasters, but it needed to be established whether such an ensemble product would also provide meaningful forecast skill. Here the skill of a calibrated ensemble product is compared to that of each calibrated method separately for tornadic cases during a one-year period as well as during the five-week 2023 Hazardous Weather Testbed (HWT)/Spring Forecast Experiment (SFE).

The five methods that comprise the ensemble members are described briefly in Table 1. Further information can be found at https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE_2022_Prelim_Findings_FINAL.pdf.

Table 1. Description of five calibrated methods.

Method	Description
STP Cal Circle	HREF grid points where UH > threshold, match STP with tornado event frequency to identify tornado probability.
STP Cal MCS-TF	Similar to STP Cal Circle, but uses tornado historical frequency associated with STP as is appropriate either for MCS or supercells according to diagnosed storm mode.
MLRF	Uses a ML random forest approach based on 137 predictors consisting of HREF ensemble-mean variables.
Nadocast	A ML method based on an ensemble of gradient-boosted decision trees with 10,000+ HREF and SREF features as predictors.
HREF/GEFS	HREF grid points where UH > threshold and GEFS env. variables (STP, MUCAPE, Eff. Shear) > thresholds, combined UH/env. variables values matched with tornado frequency to calculate tornado probability.

Acronyms

HREF = High Resolution Ensemble Forecast
GEFS = Global Ensemble Forecast System
SREF = Short-range Ensemble Forecast
UH = Updraft Helicity
STP = Significant Tornado Parameter
MCS = Mesoscale Convective System
TF = Tornado Frequency
MUCAPE = Most Unstable Convective Available Potential Energy
ML = Machine Learning

2. Method

For a given case, an ensemble of 24-hour tornado probabilities is calculated using HREF data (12 UTC initialization) and based on each of the five calibrated methods. These data, which are generated on the 3-km HREF grid, are recast to a common 80-km grid. A point-by-point ensemble average and 90th percentile are then calculated for each grid point in the domain. Figures 1 and 2 present examples of a suite of tornado forecast

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

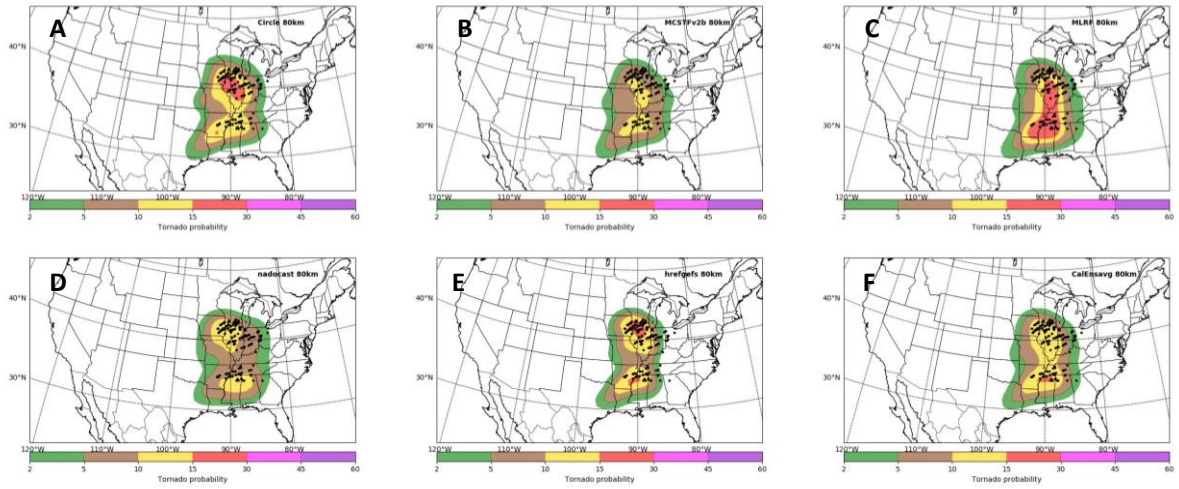

Figure 1. Tornado probabilities over 24-hr period from 12 UTC for 31 March 2023 based on five calibrated methods (A-E respectively: STP Cal Circle, STP MCS-TF, MLRF, Nadocast, and HREF/GEFS) and the ensemble mean of these five methods (F) with tornado reports (black dots).

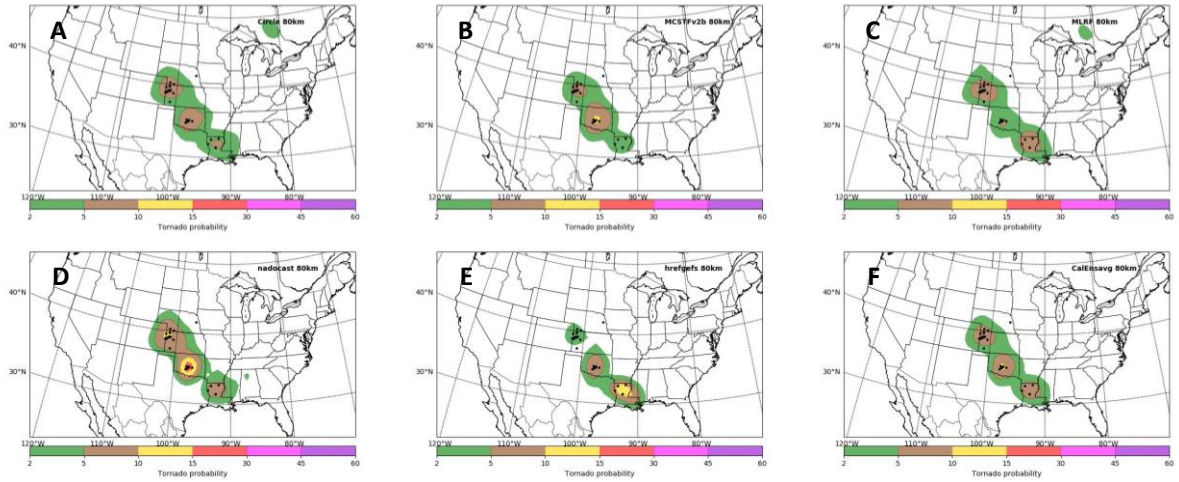

Figure 2. Same as Fig. 1 but over 24-hr period from 12 UTC for 11 May 2023.

guidance products (five calibrated methods and the ensemble mean) for two tornado outbreak events. Also generated are products showing the probability across ensemble members of tornado probabilities above 10% (Fig. 3).

These calibrated guidance products were evaluated subjectively for 19 days during the HWT/SFE 2023 period (1 May-2 June 2023). The day following an event, participants scored each tornado probability product on a scale of 1 to 10 in consideration of tornado occurrence as evidenced by official local storm reports. A score of 10 denoted an excellent forecast.

For objective verification purposes, tornado observations were mapped to the same 80-km grid as used for the calibrated tornado probabilities. A positive observation value indicated one or more observed tornadoes within 40-km of a grid point. Standard skill metrics (ROC curves, performance diagrams, reliability diagrams) were generated based on a standard 2x2 contingency table for

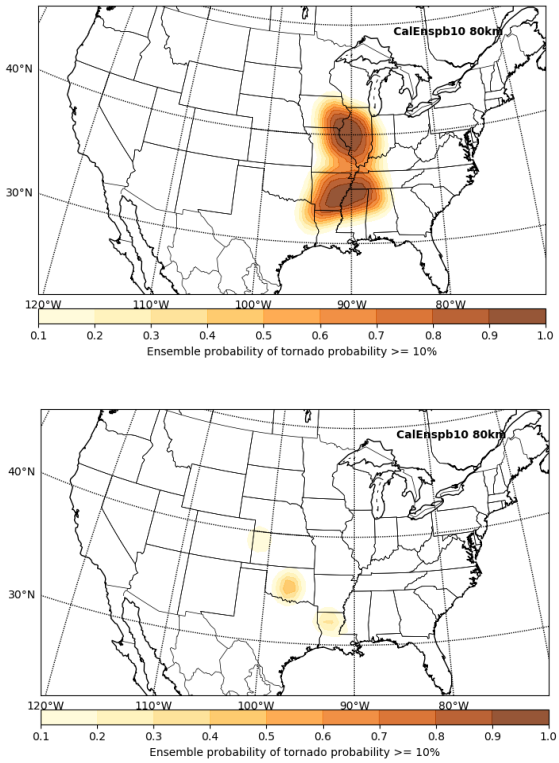

Figure 3. Ensemble probability of forecast tornado probabilities at and above 10% for the cases in Fig. 1 (top) and Fig. 2 (bottom).

which a hit was defined for a point with both a positive forecast and positive observation.

3. Results: Subjective evaluation

Subjective evaluation of the calibrated products during the SFE period is represented in Fig. 4 which shows mean scores by method as recorded by nearly 100 participants and across 19 cases. Results are also shown for subsets of days either with or without tornadoes as well as days with non-weak tornadoes (i.e., filtering out tornadoes with 'land spout' or 'weak' in the storm report text).

During the SFE, both Nadocast and Cal Ens Mean were evaluated by participants as the best performing methods. Whether over the entire 19-day period, only days with tornado observations, or only days without observations, both of these methods register the highest mean scores, and are consistently evaluated with average scores within 3% of each other. These results suggest that the ensemble mean has forecast skill nearly equal the registered skill of the best ensemble member (Nadocast).

It is worth noting that the mean values of subjective scores across all five calibrated methods (boxplots outlined in brown in Fig. 4) are less than the average scores of the Cal Ens Mean product (boxplots outlined in green in Fig. 4) as evaluated by SFE participants. This result suggests that there is forecast skill by the ensemble product itself that is slightly higher than the combined average skill of each separate calibrated method.

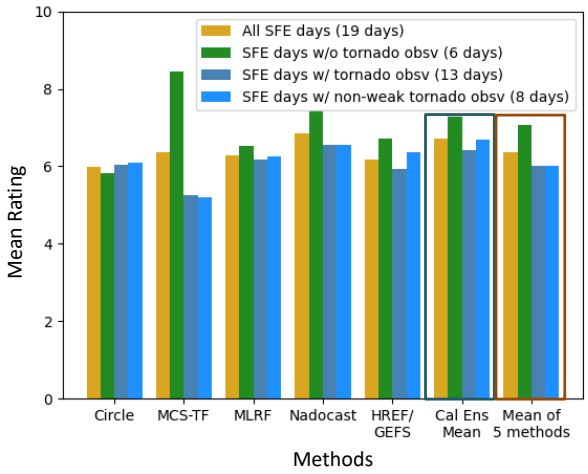

Figure 4. Mean scores by calibrated method from SFE 2023 subjective evaluations. Subsets of days (legend) defined in the text. Subjective evaluation of the Cal Ens Mean product (green box outline) and the mean of subjective scores across all five calibrated methods (brown box outline).

4. Results: Objective evaluation

In addition to subjective evaluations during the SFE, the five calibrated methods and their ensemble were evaluated objectively over a course of a year (1 May 2022 – 30 April 2023) on days for which there was at least one observed tornado in the continental US (130 days). As described in the Methods section, commonly used objective metrics were produced to compare the performance of the suite of methods over this time period (Fig. 5). Consistent with the SFE results, Nadocast registered the highest skill with Cal Ens Mean a close second. Specifically, the ROC area-under-the-curve (ROC-AUC) values were the highest for

Nadocast and the Cal Ens Mean at 0.929 and 0.912 respectively, and the areas to the left (ALC) of the performance curve were highest as well for these same methods at 0.161 and 0.138 respectively. Their reliabilities were also similar; both methods only slightly over-forecast for tornado probabilities at or below 10%.

The ensemble 90th percentile was also evaluated. Although its ROC-AUC registered the highest skill (0.934 as compared to 0.929 and 0.912 for Nadocast and the Cal Ens Mean respectively), its ALC of 0.131 from the performance diagram was less than 0.161 and 0.138 as registered for the top two methods. In addition, from the reliability diagram, the 90th percentile substantially over-forecast tornado probabilities as compared to Nadocast and the Cal Ens Mean.

5. Summary

Based on one-year of tornado cases and results from SFE 2023, the calibrated ensemble mean on average demonstrates better forecast performance than all but one of its ensemble members, namely, Nadocast. Further analysis, such as calculating daily fractional skill scores, could help to identify types of cases for which the calibrated ensemble product outperforms all ensemble members. Beyond assessing overall performance, an additional value of the ensemble is that it provides a statistical consistency of forecast probabilities among ensemble members above a given threshold and thus also a degree of forecast confidence. Lastly, it was discovered that using the ensemble mean registers a higher overall forecast performance as compared to the ensemble 90th percentile, especially considering false alarms and reliability.

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.

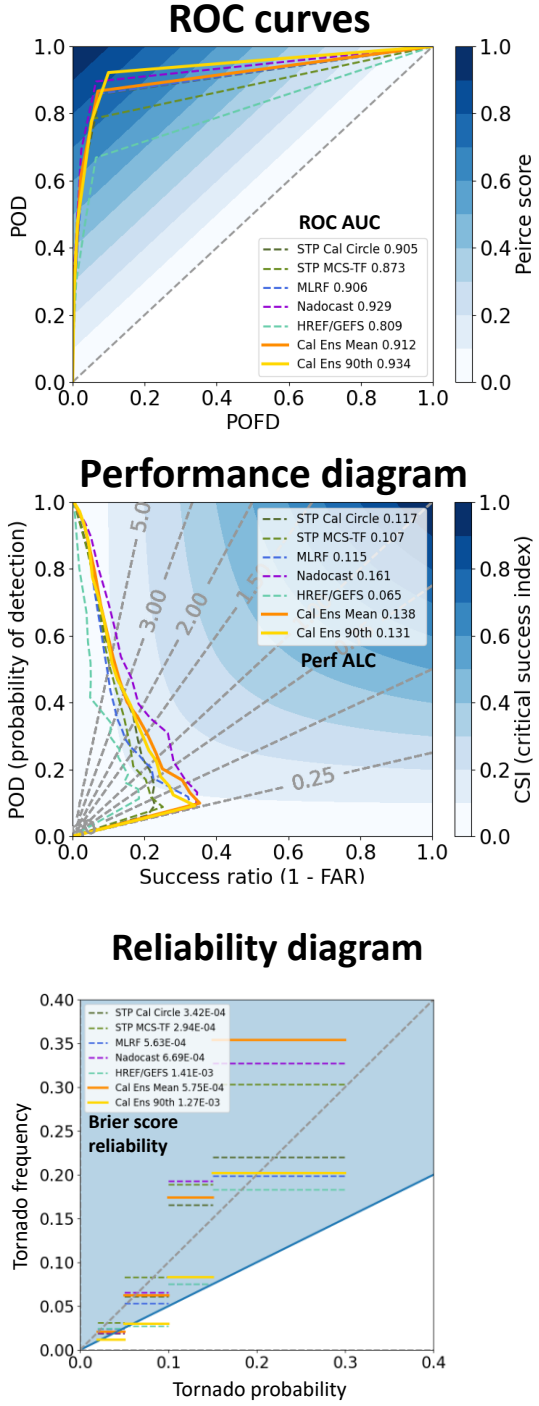

Figure 5. ROC, performance, and reliability diagrams as described in the text providing an objective evaluation of the five calibrated guidance methods along with their ensemble mean and 90th percentile.