170 Developing and Evaluating First-Guess Watch Guidance Within SPC Operations

David R. Harrison *
Cooperative Institute for Severe and High-Impact Weather Research and Operations, The University of Oklahoma, and NOAA/NWS Storm Prediction Center, Norman, OK

Israel L. Jirak and Patrick Marsh
NOAA/NWS Storm Prediction Center, Norman, OK

1. INTRODUCTION

The National Weather Service's Storm Prediction Center (SPC) is responsible for issuing Severe Thunderstorm and Tornado Watch products when conditions become favorable for organized severe thunderstorm development (SPC 2024). In particular, a Tornado Watch is issued when satellite, radar, and environmental trends appear conducive for multiple tornadoes over a focused geographic area, or when a single intense tornado is possible over the next several hours. Similarly, Severe Thunderstorm Watches are used when organized convection is expected to result in at least six severe weather events over a confined geographic region, including severe wind gusts (≥ 58 mph), large hail (≥ 1 in. diameter), and brief or weak tornadoes. As described by SPC (2024), watches are intended to encourage the general public to stay alert to changing weather conditions while providing emergency managers, storm spotters, and broadcast media lead time to prepare for severe weather operations. Additionally, the issuance of watch products has been shown to positively correlate with the quality of NWS warnings (Hales Jr 1990; Krocak and Brooks 2021) and may considerably influence weather awareness among the general public (Gutter et al. 2018).

Forecasters at the SPC issue Severe Thunderstorm Watches with the goal to provide at least 45 minutes of lead time prior to the first severe weather event (SPC 2024). Conversely, Tornado Watches are issued with an intended lead time of 2 hours before the first tornado occurrence and at least 1 hour before non-tornado severe weather hazards (i.e., wind or hail). Convective watches are typically preceded by a mesoscale convective discussion (MCD; SPC 2024) - a combined graphic and text product that conveys a forecaster's thoughts about how convection will evolve over a mesoscale domain during the next 1 to 6 hours. Severe weather MCDs are often used to highlight areas of meteorological interest and indicate the likelihood that a watch will be issued during the next few hours. Per SPC (2024), these products are intended to provide extra lead time ahead of potential severe weather development and serve as advance notice to NWS partners that a watch may be issued in the near future. It is the goal of SPC to publish an MCD at least 1 to 2 hours prior to a watch issuance when workload and predictability allow.

Given the stated lead time goals of MCDs and convective watches, SPC forecasters must begin to plan when and where a watch will be issued several hours before the impacts of severe weather hazards are observed. To aid in this challenge, Harrison et al. (2022) trained a gradient-boosted classifier on the High-Resolution Ensemble Forecasting System (HREF) to predict when and where conditions may warrant a watch within a rolling 3-h forecast window. This dynamic, first-guess watch guidance provides SPC forecasters with both probabilistic and deterministic recommendations for the location and timing of watches through the HREF's full 48-h forecast cycle. Recommended watches are designed to provide 2 to 3 hours of lead time prior to the issuance of storm-based warnings or local storm reports (LSRs), and counties within a first-guess watch are cleared within the hour after the severe weather threat is predicted to end. The first-guess watch guidance was implemented operationally at SPC in March 2023 and has been running

**Corresponding author address*: David R. Harrison, 120 David L. Boren Blvd, Norman, OK 73072; email: david.harrison@noaa.gov

continuously for about 20 months at the time of this writing. Feedback from forecasters has been overwhelmingly positive, with many noting the value of the guidance for prompting the issuance of pre-watch MCDs, increasing situational awareness, and aiding in general shift planning. Additionally, objective verification metrics have shown the machine learning (ML) guidance is skillful at emulating human-issued SPC watches and capturing observed severe weather hazards (Harrison et al. 2022).

The existing first-guess watch guidance aids forecasters in deciding when and where to issue a convective watch product, but it does not provide any assistance in determining whether a Tornado or Severe Thunderstorm Watch is preferred. Krocak and Brooks (2021) found that the type of watch issued prior to severe weather occurrence often has considerable influence on the performance and lead time of warnings issued later in the event. However, the optimal watch type is not always obvious in conditional or rapidly evolving convective environments. To address the challenge, this study attempts to expand upon the existing first-guess watch guidance by training a new ML model to predict whether a Tornado or Severe Thunderstorm watch is recommended.

2. DATA AND METHODS

The methods and datasets applied in this study were selected in part to remain consistent with those applied by Harrison et al. (2022). The ML-based first-guess watch type model was primarily trained using a combination of prognostic environment and storm-scale attributes derived from the HREFv2.1 and HREFv3 ensembles. Full 48-h 00z and 12z HREFv2.1 forecasts were collected for 10 March 2018 - 10 May 2021, and HREFv3 forecasts were obtained for 11 May 2021 - 31 May 2022 (the same period used by Harrison et al. 2022). Probabilistic surrogate forecasts for tornadoes, severe hail, and damaging wind were derived by calculating the neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017) of the HREF updraft helicity (UH), updraft vertical velocity (UVV), and 10-m wind speed. The NMEP represents the ensemble probability that each storm-scale attribute will exceed a specific threshold within a 40-km neighborhood, and multiple exceedance thresholds were assessed during feature engineering as described in the next section. A detailed depiction of the NMEP data transformation process is provided in Roberts et al. (2019), their Fig. 1.

SPC parallelogram and county-based Tornado and Severe Thunderstorm watches were collected for 10 March 2018 - 31 May 2022 and mapped to the HREF's native 3-km grid. The watches were then aligned temporally with the HREF data, such that each valid hour of a watch was paired with the most recent valid HREF cycle and forecast hour. Examples of each watch type (Tornado and Severe Thunderstorm) were compiled by sampling a subset of all grid points within each watch parallelogram.

2.1 *Feature Engineering*

Before attempting to train an ML model to predict the optimal watch type, it was first necessary to identify and assess how the various HREF prognostic fields vary by Tornado and Severe Thunderstorm Watch environments. In theory, environment and storm-scale attributes with less distribution overlap between the two watch types should be stronger predictors of those classes. To this end, every prognostic field, storm-scale attribute, and derived severe weather surrogate was compared across Tornado and Severe Thunderstorm Watches via a two-sample Kolmogorov-Smirnov test (KS; Hodges 1958). The KS test measures the goodness of fit between two distributions and tests against the null hypothesis that both distributions are identical. The resulting KS score is the maximum absolute difference between the empirical distribution functions of the two samples, where larger KS scores indicate greater separation between the two distributions. For the purposes of this study, a greater KS score indicates the HREF field exhibited greater discrimination between Tornado and Severe Thunderstorm watch environments.

The KS score of each HREF field was calculated over the 2-year period from 10 March 2018 - 1 March 2020 and the highest-scoring fields were identified. The meridional (v) component of the 700-mb wind was found to have the largest KS

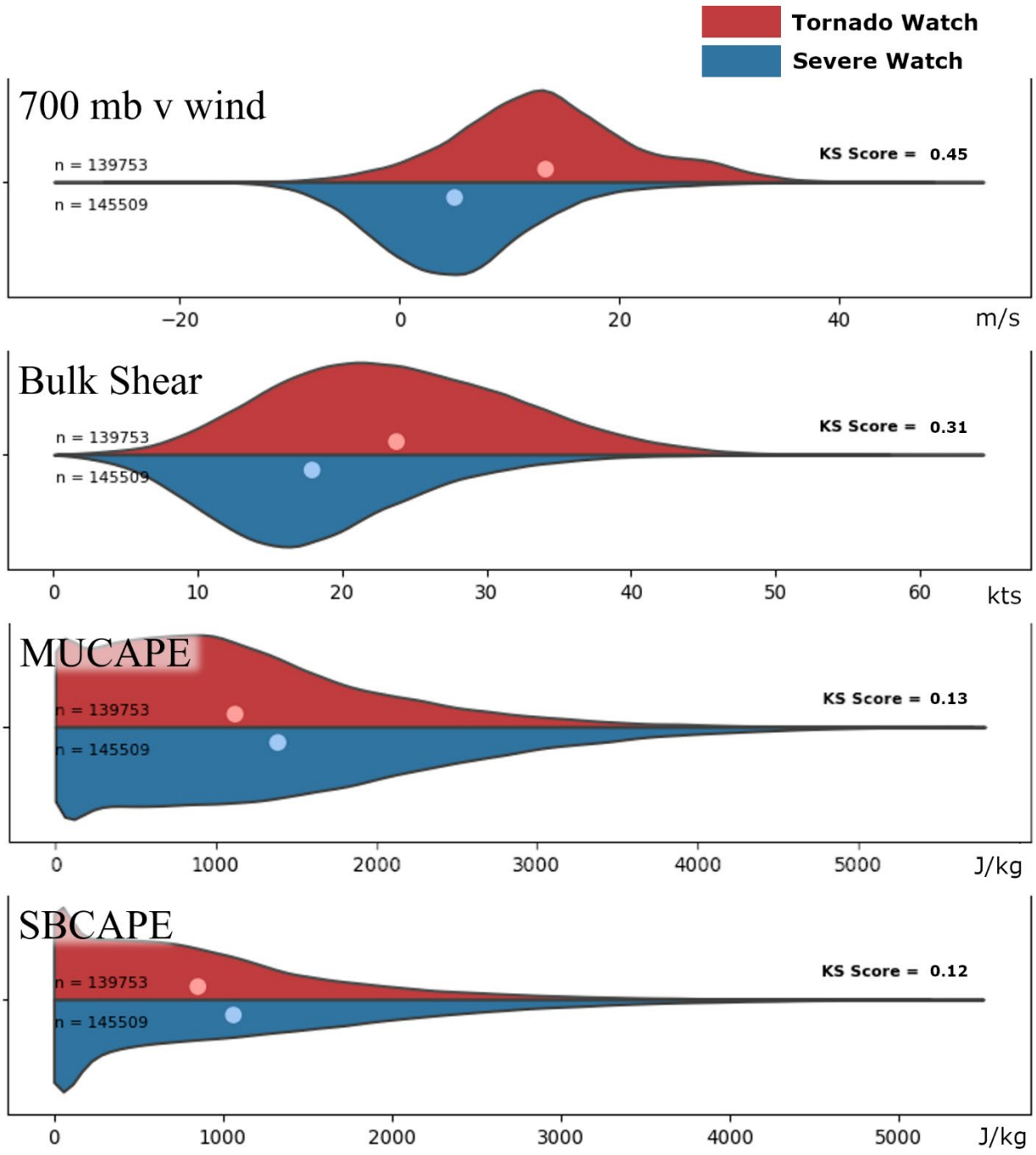

Figure 1: Example distributions of HREF prognostic fields sampled from within Tornado (red) and Severe Thunderstorm (blue) Watches. The dots represent the mean values of each distribution.

Field	KS Score	Field	KS Score
700 mb Wind Speed	0.43	LCL Height	0.25
0-1 km Storm Relative Helicity	0.36	SB CINH	0.22
Sig. Tornado Parameter	0.35	MU CINH	0.20
0-6 km Bulk Shear	0.31	MU CAPE	0.13
0-3 km Storm Relative Helicity	0.29	SB CAPE	0.12
700 mb Wind Direction	0.29	2-m Dewpoint	0.11
Month	0.28	2-m Specific Humidity	0.07

Table 1: HREF prognostic fields selected for training and their corresponding KS scores.

score of 0.45, followed by 0-1 km storm-relative helicity (0.36), the significant tornado parameter (0.35), 0-6 km bulk shear (0.31), and 0-3 km storm relative helicity (0.29). Other high-ranking variables include month of the year, LCL height, surface-based and most unstable CAPE, and 2-m dewpoint. A list of the 14 variables selected to train the ML model and their respective KS scores is provided in Table 1.

Closer inspection of the 700-mb v wind component distribution revealed that environments with 700-mb v wind greater than about 10 m/s were more likely to be associated with Tornado watches than Severe Thunderstorm watches (Fig. 1). Further reconstructing the wind field into speed and direction components revealed this discrimination was largely driven by the speed of the 700-mb flow, though more southerly flow patterns did demonstrate some signal for Tornado watches as well. These results largely align with long-standing operational rules of thumb such as those proposed by Beebe (1956) and are an encouraging start to this ML task. The distributions of the other top-rated variables also behaved within meteorological expectations, but many had considerable overlap between Tornado and Severe Thunderstorm watch environments.

2.2 *Model Design*

Prior to model development, the dataset was separated into independent training, validation, and testing sets. Examples from 10 March 2018 - 1 March 2020 were selected for the training dataset, and 10 March 2020 - 31 May 2022 was used for model calibration, validation, and tuning. Ten days were withheld between datasets to avoid cross-contamination from temporal autocorrelation within the features. Model testing was performed in real-time from 29 April 2024 – 31 May 2024 during the 2024 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE).

A gradient-boosted classifier (GBC) was trained to predict whether each example in the dataset belonged to the Tornado or Severe Thunderstorm class. A randomized grid search with 5-fold cross validation was used to train and tune the model hyperparameters. The GBC achieved a maximum critical success index (CSI) of about 0.65 for both Tornado and Severe Thunderstorm watches on the validation dataset (Fig. 2).

Initial analysis of the GBC performance revealed that the model tended to produce overconfident class predictions, with both positive and negative class probabilities heavily skewed towards 0 or 1. This behavior resulted in probabilistic forecasts that were statistically unreliable with the observed class frequency as indicated by a reliability diagram systemically offset from the one-to-one (not shown). To account for this overconfidence, an isotonic regression model was applied by first running the GBC on the validation dataset and then training the isotonic regression on those predictions. As before, 5-fold cross validation was used to assess the isotonic regression performance and the 95% confidence interval was calculated via 10,000 bootstrapped samples. The resulting calibrated model did not exhibit any notable change in CSI; however, the class probabilities produced by the GBC and isotonic regression were found to be less skewed and more statistically reliable with observations. As a result, the 50% confidence level was found to be the optimal decision threshold when deterministically choosing between Tornado and Severe Thunderstorm Watch predictions.

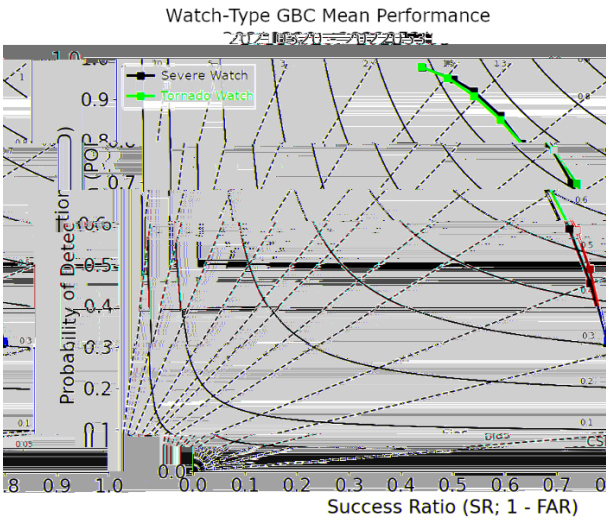

Figure 2: Mean performance of the 12z HREF-based watch-type gradient boosted classifier for 20 March 2021 – 31 May 2022.

2.3 *Application to Existing Guidance*

The watch-type guidance was designed to be an extension of the existing first-guess watch guidance. As described by Harrison et al. (2022), a county is included within a first-guess watch product at a given forecast hour if (1) the mean watch probability of all grid points within the county ≥ 70% and (2) any part of the county falls within at least an SPC-issued Slight risk area. Counties are also removed from the first-guess watch when these criteria are no longer met. These criteria result in an hourly forecast watch product that ideally extends about 3-hours downstream of a predicted severe weather hazard and automatically removes counties for locations where the severe weather threat has passed.

Once the first-guess watch prediction is complete for a given forecast hour, the watch type model is run for all grid points contained within a first-guess watch county. The type of watch for that county is specified as a Tornado Watch if the average model output for all grid points within that county are ≥ 50%. Otherwise, the county is considered part of a Severe Thunderstorm Watch. The resulting watch-type predictions are represented to end users by applying a color fill to each county, where red indicates a Tornado Watch and blue represents a Severe Thunderstorm Watch. An example forecast for 09 May 2024 is provided in Fig. 3.

3. 2024 Spring Forecasting Experiment

Initial testing of the watch-type guidance was performed as part of the 2024 HWT SFE. The 2024 SFE was jointly conducted by SPC and NOAA/OAR/National Severe Storms Laboratory over a five-week period from 29 April - 31 May. This experiment marked the second hybrid SFE in which a combination of in-person and virtual participants simultaneously contribute to activities and evaluations. Participants included over 160 forecasters, researchers, model developers, university faculty, and graduate students from around the world, with 68 participants contributing remotely (Clark et al. 2024). The 2024 SFE also coincided with one of the most active Mays on record, during which all but five days of the experiment had an SPC-issued Enhanced Risk or greater.

The first-guess, county-based watch forecasts were presented to SFE participants during a daily evaluation period via an interactive webpage with two graphic panels as shown in Fig. 3. Hourly forecasts from the ML guidance were presented in the left panel, while the “observed” SPC-issued Severe Thunderstorm and Tornado Watches were provided in the right-most panel. An interactive slider bar at the top of the webpage enabled participants to step through each available forecast hour (12z - 12z; f00 - f24), and overlays of local storm reports, NWS storm-based warnings, and the 13z D1 SPC outlook could be toggled on both panels. The spatial scope of the evaluation was limited to a rectangular domain of 15º longitude × 8.721º latitude, and this domain was set each day by the SFE facilitators to best contain the most significant severe-weather threat for the day. All evaluations of the first-guess watches were performed for the previous day’s severe weather.

The evaluation survey presented to participants consisted of six questions, including three open-response and three multiple-choice questions. The first two questions collected metadata, asking respondents to enter the date of the forecast being evaluated and their unique participant number.

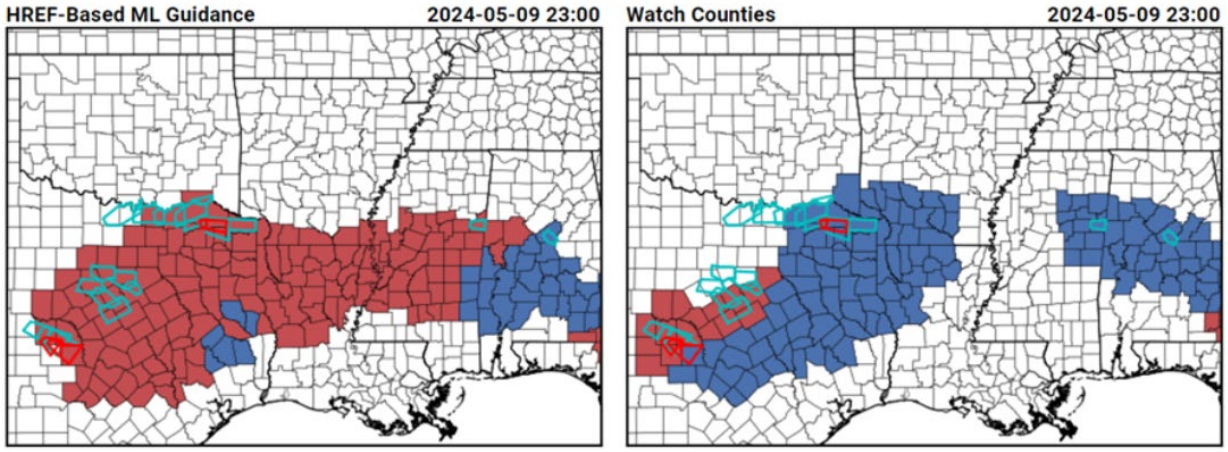

Figure 3: An example of tornado (red) and severe thunderstorm (blue) watches by county predicted by the GBC (left) and issued by SPC (right) valid for 20240509 2300 UTC. Polygons indicate NWS tornado (red) and severe thunderstorm (cyan) warnings.

These questions were included in all evaluation surveys during the 2024 SFE and enabled facilitators to remove results from participants who did not agree to share their responses for scientific study. Question 3 (Q3) asked respondents to subjectively rate how well the first-guess watch guidance captured the timing and location of the severe-weather threat during the forecast period. This question was specifically focused on the performance associated with the location and timing of a predicted watch with no consideration to the recommended watch type. The guidance was assessed on a 10-point Likert scale with values ranging from "Very Poor" to "Very Good." Respondents were instructed to consider the full 24-hour forecast period when determining their responses, and an option of "N/A" was provided if there were no operational watches issued for the event.

Next, Q4 directed participants to subjectively evaluate how well the recommended watch type (Tornado or Severe Thunderstorm) matched the type of watches issued by the NWS during the forecast period. Again, the guidance was independently assessed via a 10-point Likert scale ranging from "Very Poor" to "Very Good." Q5 asked participants to rate how appropriate the recommended watch type (Tornado or Severe Thunderstorm) was for the observed hazards (based on local storm reports and NWS warnings) on the same 10-point Likert scale as before. This question was included to account for situations where the SPC-issued watch may not have been the most optimal for the observed hazards. Finally, Q6 provided an open response field for participants to describe their thoughts about the guidance's performance for the day.

Participants' feedback was generally favorable through the experiment, with operational forecasters particularly noting the potential benefits that such guidance could provide. Several open-response comments indicated the first-guess watch forecasts could provide a quick summary of relevant HREF prognostic fields, and the location and timing aspects of the guidance may be useful for shift planning and situational awareness. Participants gave the guidance a mean rating of 6.6 for how well the timing and placement of the watches captured the observed severe hazards each day (Fig. 4). Some respondents voiced concerns that the guidance tended to include too many counties in a first-guess watch compared to those issued by the SPC. However, some of this discrepancy may be due in part to the longer lead times targeted by the guidance (3 hours uniform across the watch area) compared to the operational watch products (variable depending on hazard and location).

The new watch-type guidance generally received lower ratings overall than the first-guess watch product, with participants commonly noting a strong bias toward Tornado Watches. The

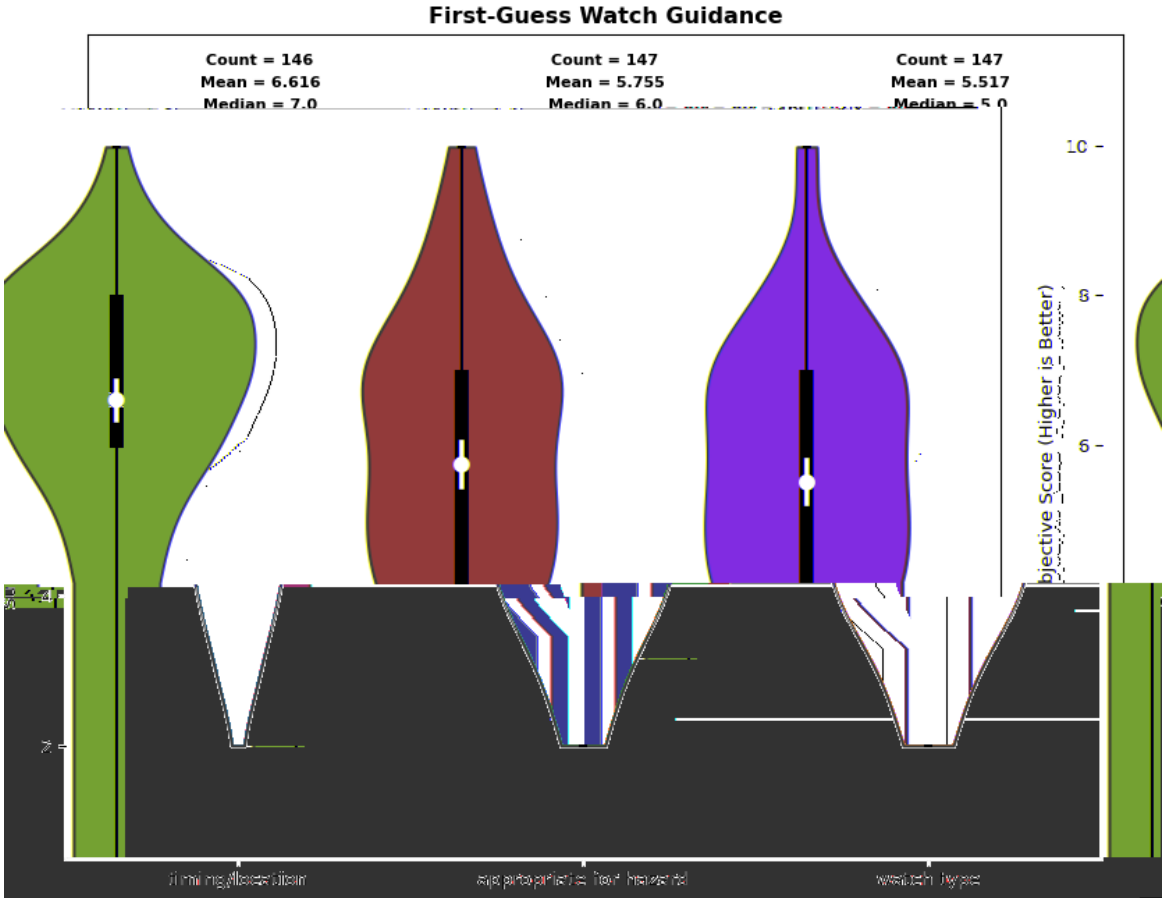

Figure 4: Violin plots showing participant ratings of the watch guidance in consideration of: the recommended timing and location (blue), how appropriate the recommended type was for the observed hazards (green), and how closely the guidance matched the SPC-issued watch type (yellow).

guidance was given a mean rating of 5.5 when assessing how similar the recommended watch type was to that issued by the SPC. Interestingly, participants did give the guidance a slightly higher mean rating of 5.7 when evaluating how appropriate the watch type was for the observed hazards. This may suggest there were times when the guidance differed from the operational watch type but more accurately predicted the hazards. However, differences between the two scores fell well within the 95% confidence interval and thus are not statistically significant.

4. CONCLUSION

To summarize, the new watch-type guidance was found to have skill when discriminating between Tornado and Severe Thunderstorm watches based on HREF prognostic fields. However, SFE participants identified potential biases in the forecast guidance and suggested additional improvement may be necessary before the watch-type model is ready for operational application. Future research will work to investigate and calibrate the reported bias towards Tornado Watches, and more complex ML techniques may be applied to better emulate the forecaster thought process when choosing a watch type. This and other future development will continue to be run in real-time within SPC operations and presented to SPC forecasters via an experimental web interface to encourage frequent feedback and co-development. Ideally, this increased collaboration will continue to allow rapid development of the first-guess watch products.

5. ACKNOWLEDGMENTS

This extended abstract was prepared by David Harrison with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA21OAR4320204, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

6. REFERENCES

Beebe, R. G., 1956: Tornado composite charts. *Mon. Wea. Rev.*, **84**, 127 – 142, https://doi.org/10.1175/1520-0493(1956)084%3C0127:TCC%3E2.0.CO;2.

Clark, A. J., and Coauthors, 2024: Advancing hazardous weather prediction in the 2024 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **105**, E2180 – E2183, https://doi.org/10.1175/BAMS-D-24-0249.1.

Gutter, B. F., K. Sherman-Morris, and M. E. Brown, 2018: Severe weather watches and risk perception in a hypothetical decision experiment. *Wea. Climate Soc.*, **10**, 613–623, https://doi.org/10.1175/WCAS-D-18-0001.1.

Hales Jr, J. E., 1990: The crucial role of tornado watches in the issuance of warnings for significant tornadoes. *Natl. Wea. Dig.,* **15**, 30 – 36.

Harrison, D., A. McGovern, C. Karstens, I. Jirak, and P. Marsh, 2022: Evaluation of first-guess watch guidance in the 2022 HWT Spring Forecasting Experiment. *30th Conf. on Severe Local Storms*, Santa Fe, NM, AMer. Meteor. Soc., 7.3A, https://ams.confex.com/ams/30SLS/meetingapp.cgi/Paper/407263.

Hodges Jr, J. L., 1958: The significance of the Smirnov two-sample test. *Arkiv for matematik*, **3**, 469 – 486.

Krocak, M. J. and H. E. Brooks, 2021: The influence of weather watch type on the quality of tornado warnings and its implications for future forecasting systems. *Wea. Forecasting*, **36**, 1675 – 1680, https://doi.org/10.1175/WAF-D-21-0052.1.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

Schwartz, C. S. and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Review*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

SPC, 2024: SPC products. Accessed 19 November 2024. https://www.spc.noaa.gov/misc/about.html.