7.3A EVALUATION OF FIRST-GUESS WATCH GUIDANCE IN THE 2022 HWT SPRING FORECASTING EXPERIMENT

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

Amy McGovern
The University of Oklahoma, Norman, OK

Christopher D. Karstens
NOAA/NWS Storm Prediction Center and School of Meteorology, University of Oklahoma, Norman, OK

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

1. INTRODUCTION

The Storm Prediction Center (SPC) is responsible for issuing Severe Thunderstorm and Tornado Watch products when conditions become favorable for organized severe thunderstorm development. 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, and 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). Watch products are initially produced by SPC forecasters as parallelograms that define the approximate area of the predicted severe weather threat. Those parallelograms are then converted into official county-based watch products following direct collaboration with affected National Weather Service (NWS) Weather Forecast Offices (WFOs) (NWS 2021). During these collaboration calls, the SPC lead forecaster discusses meteorological conditions and any local considerations that might impact the spatial or temporal scope of the watch with WFO forecasters. The resulting Severe Thunderstorm and Tornado Watches typically range in size from 20,000 to 40,000 square miles and have a duration of 6 to 8 hours. However, a watch may be canceled early or extended in space and/or time by local WFOs as conditions require.

Given the stated lead time goals of Severe Thunderstorm and Tornado 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. This process is further complicated by NWS procedures which transfer ownership of a watch product from SPC to the affected WFOs immediately after issuance (NWS 2021). Any changes to a watch after it is initially transmitted by SPC must be coordinated with and enacted by local WFO forecasters who may be otherwise occupied issuing warnings, communicating with partners, and performing other high-priority severe weather operations. As such, it is important that SPC forecasters correctly estimate the location and time of severe weather development to ensure each watch optimally covers the severe weather threat. A watch that does not adequately define the spatial or temporal domain of the convective weather hazards might require a local extension by WFOs or the issuance of another watch by SPC. These actions ultimately increase workload on both SPC and WFO forecasters and may result in delayed product dissemination, inconsistent messaging, and public confusion.

As with most products in the modern NWS watches, warnings, and advisories (WWA) paradigm, Severe Thunderstorm and Tornado Watches are static products that are cumbersome to modify once issued. Although local WFOs have

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

the authority to cancel or add individual counties to an existing watch, large changes typically require the issuance of a new watch by SPC. These limitations of the current paradigm potentially result in non-uniform lead times across the spatial domains of WWA products (Stumpf and Gerard 2021), where locations on the upstream side of a static product often see impacts from hazardous weather sooner than those farther downstream. This disparity of lead time not only has the potential to be inequitable to populations within WWA products, but it can also lead to undesirable discrepancies in public response (Doswell 1999, Krocak et al. 2019, Hoekstra et al. 2011).

To address these challenges of the WWA system, NOAA's Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) initiative is tasked with exploring innovative methods that can shift NWS services from static, deterministic products to a dynamic, probabilistic paradigm. One such product has been realized in the form of dynamically-updating deterministic warning polygons which automatically track along a severe storm's path in real time. These non-static tornado and severe thunderstorm warnings have been termed "Threats-in-Motion" (TIM; Stumpf et al. 2011; Stumpf and Gerard 2021) and are currently being proposed as a bridge between the deterministic WWA system and the probabilistic FACETs paradigm. Although TIM is primarily concerned with storm-scale warnings, the concept of deriving dynamically-evolving deterministic products from an underlying probabilistic paradigm can be further developed and applied to address the operational challenges and limitations of SPC Severe Thunderstorm and Tornado Watches. To this end, this research presents and demonstrates a prototype system which applies machine learning (ML) techniques to produce a dynamic, first-guess watch product.

2. DATA AND METHODS

Initial collaboration with SPC forecasters and management identified three key objectives for an automated first-guess watch guidance product. First, the product should provide both probabilistic and deterministic outputs in a format comparable to existing watch products (i.e., county-based forecasts). Based on the research conclusions of Gutter et al. (2018) and Krocak et al. (2019), informal feedback from WFO forecasters, and formal operational requirements specified by NWS directives, it was next determined that the first-guess watch products should optimally provide 2-3 hours of lead time prior to the issuance of storm-based warnings or observed local storm reports (LSRs). Finally, SPC forecasters and management suggested that these goals should be achieved using both machine ML and non-ML techniques so that the two approaches might be compared during later evaluations.

2.1 *Feature Engineering*

The ML-based first-guess watch guidance was primarily trained using prognostic storm-scale attributes derived from the High-Resolution Ensemble Forecast (HREF) v2.1 and HREFv3 ensembles. For this study, full 48-hour 00z and 12z HREFv2.1 forecasts were obtained for 10 March 2018 - 10 May 2021, and HREFv3 forecasts were collected for 11 May 2021 - 31 May 2022 (the full period available). As detailed by Roberts et al. (2019, 2020), the HREF is an ensemble of opportunity composed of five deterministic CAM configurations and their 12-hour (6-hour for the HRRR) time-lagged cycles. Each CAM configuration is compiled using different combinations of dynamical cores, initial and boundary conditions, microphysics schemes, and PBL schemes which provide greater forecast spread and more effectively samples forecast uncertainty than unified convection-allowing ensembles. This forecast spread is perhaps most apparent in the derived storm-scale attributes produced by each HREF member, as deterministic simulated convection often varies considerably in spatial and temporal placement among the different CAM configurations. However, storm-scale attributes have also been shown to be skillful predictors of severe hazards when smoothed and upscaled to produce probabilistic "surrogate severe" forecasts (Sobash et al. 2011, 2016; Roberts et al. 2019, 2020; Gallo et al. 2021). Such surrogate severe fields were hypothesized to be strong candidates for training an ML-based first-guess watch product as they not only serve as

Field	Threshold	Mask
NMEP 1-h Max UH	99.85%	
NMEP 1-h Max UVV	20 ms^{-1}	
NMEP 1-h Max 10-m Wind Speed	35 kt	$Z_{Comp} > 30\ dBZ$
Mean 1-h Max 10 m – 500 mb Shear		$Z_{Comp} > 30\ dBZ$

Table 1: Derived storm-scale and environmental fields, their optimal exceedance thresholds, and spatial masks that make up the training dataset.

proxies for explicit hazard prediction but also represent localized spatial scales similar to that of operational SPC watches.

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 specified threshold within a 40-km neighborhood, where the best exceedance thresholds for each feature were identified during model tuning as described in the next section. A detailed depiction of the NMEP data transformation process is provided by Roberts et al. (2019), their Fig. 1.

While UVV and 10-m wind speed exceedance thresholds were calculated from fixed physical values, it was necessary to base the UH threshold on climatological percentiles. Research by Potvin el al. (2019) and Gallo et al. (2021) noted that some members of the HREF produce higher UH values on average than others. To facilitate these differences in model climatology, the UH exceedance threshold was tuned based on percentile values specific to each HREF member. It was also necessary to mask the 3-hour maximum 10-m wind field such that only grid points where the 3-hour maximum composite reflectivity (Z_{Comp}) exceeded 30 dBZ were included in the NMEP calculations. This step was required to exclude any non-convective wind forecast by the HREF, particularly in mountainous regions and in the stratiform region of extratropical cyclones during the cool season. Finally, the ensemble-mean 10 m - 500 mb bulk wind shear was identified as another strong indicator of severe weather potential, particularly in cases when the forecast UH signal is limited. A full list of training variables and their tuned exceedance thresholds are provided in Table 1.

Both 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, and each watch was aligned temporally with the most recent valid HREF cycle and forecast hour. Examples of the positive or target class (WATCH) were compiled by sampling a subset of all grid points within each watch parallelogram. Similarly, examples of the negative or null class (NO WATCH) were compiled by sampling grid points within at least an SPC Day 1 (D1) convective outlook marginal (MRGL) risk category but not within a watch parallelogram. Restricting the null class examples to locations within a MRGL ensured that non-severe and pre-severe convective environments were well represented within the dataset and avoided placing too much emphasis on trivial non-convective environments. Examples of the sampling process are shown in Fig. 1.

2.2 *Model Design*

Prior to model development, the dataset was separated into independent training, calibration, and testing sets. Examples from 10 March 2018 - 1 March 2020 were selected for the combined training and validation set, 10 March 2020 - 10 March 2021 was used for model calibration, and the test set contained examples from 20 March 2021 - 31 May 2022. Ten days were withheld between datasets to avoid cross-contamination from temporal autocorrelation within the features.

A gradient boosted classifier (GBC) was trained to predict whether each example in the dataset belonged to the WATCH or NO WATCH class. A

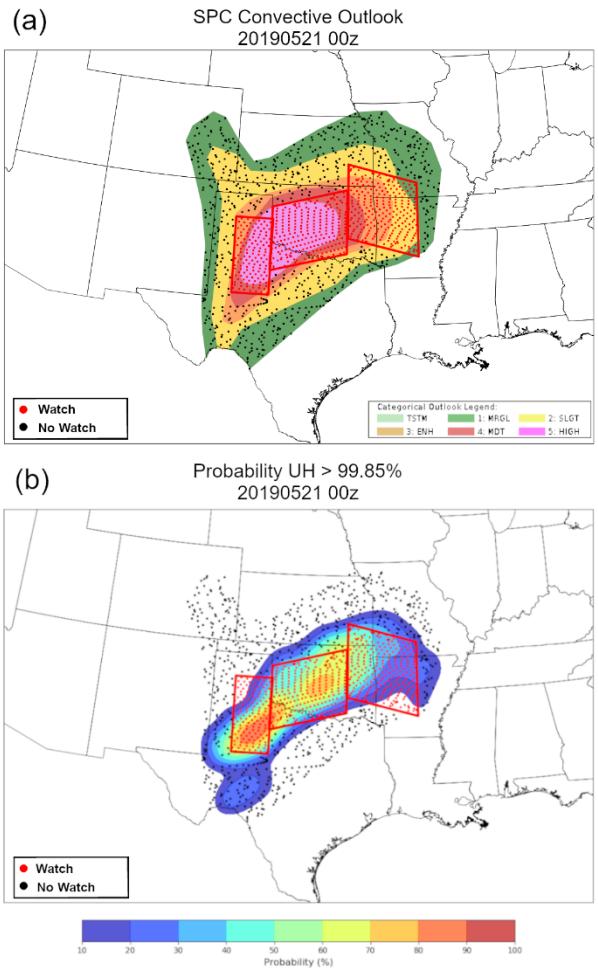

Figure 1: (a) Example of grid point sampling within the SPC D1 MRGL risk on 20 May 2019. Red dots indicate positive class (WATCH) samples while black dots represent negative (NO WATCH) samples. (b) Grid point sampling of NMEP UH values > 99.85% of climatology on 20 May 2019.

randomized grid search with 5-fold cross validation was used to train and tune the model hyperparameters while also identifying the optimal exceedance thresholds for each NMEP feature. The GBC achieved a maximum accuracy of 0.862 (with a bootstrapped 95% confidence interval of 0.859 – 0.865), and the surrogate-severe storm-scale attributes demonstrated the greatest predictive skill when using exceedance thresholds of 35 kts for 10-m wind speed, 20 m/s for UVV, and 99.85% of model climatology for UH.

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 line. To account for this overconfidence, an isotonic regression model was applied by first running the GBC on the calibration 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 accuracy scores; however, the class probabilities produced by the GBC and isotonic regression were found to be less skewed and more statistically reliable with observations. All discussion of the GBC model herein refers to the GBC with the isotonic regression applied.

2.3 *Deriving County-Based Watches*

During the initial design of this research, SPC forecasters indicated the operational desire for model output that is directly comparable to the current watch paradigm. As such, it was not sufficient to provide forecasters with a model that only predicts the probability of a watch at a given location and time; rather, the model should provide derived first-guess county-based watch predictions as well. To achieve this goal, it was first necessary to identify the optimal forecast probability threshold to use when stratifying WATCH and NO WATCH predictions. Full-CONUS probabilistic forecasts were generated by applying the calibrated GBC model to each point within the HREF 3-km grid for each forecast hour in the calibration dataset. The resulting watch probabilities were then interpolated to a 40-km grid and smoothed with a Gaussian kernel (σ = 40 km) to reduce noise and better represent the spatial scales at which SPC watches are typically issued. SPC watches were also mapped to the 40-km grid and temporally aligned with the forecasts as before, and the GBC probabilistic forecasts were compared to the SPC watches on a grid point by grid point basis.

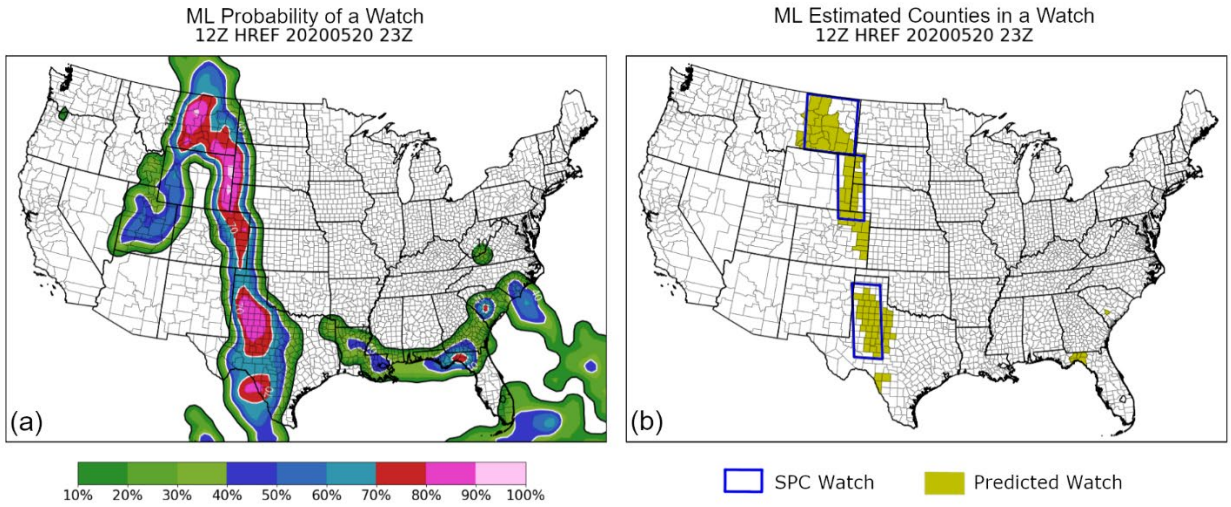

Figure 2: (a) Forecast watch probabilities and (b) derived first-guess county-based watches for 20 May 2020 23:00 UTC. The blue polygons represent operational SPC Severe Thunderstorm Watch parallelograms valid for this hour.

A forecast probability of 70% was identified as the optimal threshold for stratifying WATCH and NO WATCH forecasts, with a mean CSI of 0.24 and a bias of 1.4. However, anecdotal observations and bulk statistics revealed considerable false alarm particularly in the lower forecast probabilities. This was corroborated during initial evaluation by SPC forecasters who noted that much of the false alarm was located in environments supportive of precipitation but generally unfavorable for severe convective storms. To help reduce the model's tendency to overforecast the spatial extent of watch probabilities, the SPC convective outlook was investigated as a potential way to further mask non-severe environments. Model performance notably increased when only considering points within at least a MRGL, with a maximum mean CSI of 0.29 and a bias of 1.05. Additional improvement was observed with the application of a SLGT risk mask, though diminishing returns were noted compared to the MRGL risk mask. Forecasts within at least a SLGT risk area exhibited a maximum average CSI of 0.32 and a bias of 0.95. GBC forecast probabilities also demonstrated improved statistical reliability as the categorical threshold increased, but all models were still found to overforecast on average at most probability bins.

From these results, a set of criteria was proposed to derive deterministic county-based watches from the hourly probabilistic GBC forecasts. 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 a SLGT 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. An example of a probabilistic and deterministic watch forecast for 20 May 2020 23z is provided in Fig. 2.

2.4 *SPC Severe Timing Guidance*

A secondary goal of this research was to derive a first-guess county-based watch product using non-ML techniques. This non-ML watch guidance was produced by leveraging the output from an experimental product known as the SPC Severe Timing Guidance (Jirak et al. 2020). The SPC Severe Timing Guidance is a prototype system that combines explicit convective timing and evolution details from the SREF and HREF with human-issued SPC convective outlooks to provide probabilistic information about how and when the

severe weather threat will evolve during a convective day. This product is represented as rolling 4-hour individual hazard probabilities for tornadoes, severe wind, and hail. Readers are encouraged to refer to Jirak et al. (2020) for details of how the Severe Timing Guidance is derived.

The inputs used to produce the SPC Severe Timing Guidance algorithm exhibit many parallels to those used to train the ML-based first-guess watch product as described previously. For example, both techniques utilize the HREF NMEP UH as a proxy for severe hazards, and both methods temporally aggregate these storm-scale and environmental attributes to produce rolling windows of effective lead time. Because of these similarities, the experimental SPC Severe Timing Guidance was selected as a starting point to derive a non-ML first-guess watch product.

First, the gridded Severe Timing Guidance individual hazard probabilities for a given forecast hour were mapped to their equivalent SPC convective outlook categories. For example, a grid point with a Severe Timing Guidance tornado probability between 5% - 10%, wind probability between 15% - 30%, or hail probability between 15% - 30% was considered equivalent to a SLGT risk. The maximum category between the three hazards was then identified for each grid point. Finally, first-guess county-based watches were derived using similar criteria as that defined for ML-based watches. Specifically, a county was included within a first-guess watch at a given forecast hour if (1) the maximum Severe Timing Guidance equivalent risk category within the county was at least a SLGT, and (2) any part of that county was included within an SPC convective outlook SLGT risk contour. This second condition was primarily included to be consistent with the ML criteria, as the Severe Timing Guidance by definition should never produce probabilities greater than an equivalent SLGT outside of a SLGT risk contour in the official convective outlook.

3. RESULTS AND DISCUSSION

The ML and Severe Timing Guidance first-guess watch products were objectively evaluated on the 14-month independent test set of 20 March 2021 - 31 May 2022. Hourly, full-CONUS, county-based forecasts were generated for each convective day within the evaluation period where the SPC 13z D1 convective outlook contained at least a SLGT risk, and this resulted in a total of 244 days available for verification. During initial collaborations with SPC forecasters and management, two primary goals of the evaluation phase were identified. First, any verification should identify how well the ML and Severe Timing Guidance first-guess watches align spatially and temporally with the official SPC Tornado and Severe Thunderstorm Watches valid during the same period. These metrics are intended to assess whether the forecast products are able to emulate the timing and spatial specificity of human-issued watches. The second goal of the objective evaluation is to determine how well the forecast guidance is able to correctly predict the true severe weather hazard regardless of when or where SPC issued a watch. This question removes the assumption that SPC watches are always optimal and avoids penalizing the guidance for predicting a watch where severe weather occurred but an operational watch was not issued.

3.1 *Comparison to SPC Watches*

The first stated goal of this objective evaluation is to assess how well the forecast guidance emulates the timing and placement of SPC Severe Thunderstorm and Tornado Watches. Deterministic, county-based, first-guess watches produced by the ML and Severe Timing Guidance algorithms were first mapped to a 40-km grid for each forecast hour in the evaluation dataset. SPC watches were also mapped to the same 40-km grid, and the products were compared on a grid point by grid point basis. Contingency table metrics were calculated for the ML and Severe Timing Guidance watches using the operational SPC watches as “true” observations. As such, a predicted first-guess watch at a given grid point verified as a true positive only if there was an SPC-issued Severe Thunderstorm or Tornado Watch valid at that grid point and forecast hour.

The conditional POD, FAR, CSI, and bias were calculated for each predictive model for each convective day, and these values were averaged across the full 14-month evaluation period (Fig. 3). The verification presented herein is for ML

forecasts generated from the 12z HREF and Severe Timing Guidance based on the 12z HREF and 13z D1 convective outlook. Similar results were noted for other forecast cycles. Both the ML and Severe Timing Guidance first-guess watches exhibited a mean CSI of 0.32 (0.27 – 0.38). The ML guidance generally exhibited a higher POD and FAR compared to the Severe Timing Guidance, suggesting that the ML approach produced a greater frequency of positive class predictions than the non-ML approach on average. However, these differences were not found to be statistically significant at the 95% confidence level. Indeed, both products demonstrated little forecast bias, with mean scores of 1.09 (0.99 - 1.23) for the ML-based approach and 0.89 (0.78 - 1.03) for the Severe Timing Guidance. Furthermore, the optimal forecast bias score of 1 was observed to fall within the 95% confidence intervals for each model, indicating that neither product strongly overforecast or underforecast the areal coverage of SPC watches at times when one was in effect.

3.2 *Capturing the Severe Weather Threat*

The second objective of this evaluation was to assess how well the HREF-based ML and SPC Severe Timing Guidance first-guess watch products capture observed severe weather hazards regardless of when or where SPC issued a watch. To accomplish this, LSRs were obtained for each day in the evaluation database and filtered to exclude any reports that fell outside of a 13z D1 SLGT risk area. Filtering reports by the SPC convective outlook avoids penalizing the watch guidance for missing severe weather events in locations where it was systematically precluded from producing a forecast. Mean contingency table metrics were calculated for the ML and Severe Timing Guidance first-guess county-based watch products using a similar method to that described by Anthony and Leftwich (1992).

First, POD was calculated as the percentage of LSRs contained within a first-guess watch at the time of the report. Similarly, the percent verified (PV) was defined as the percentage of counties included in the first-guess watch product that contained an LSR during the watch's valid duration.

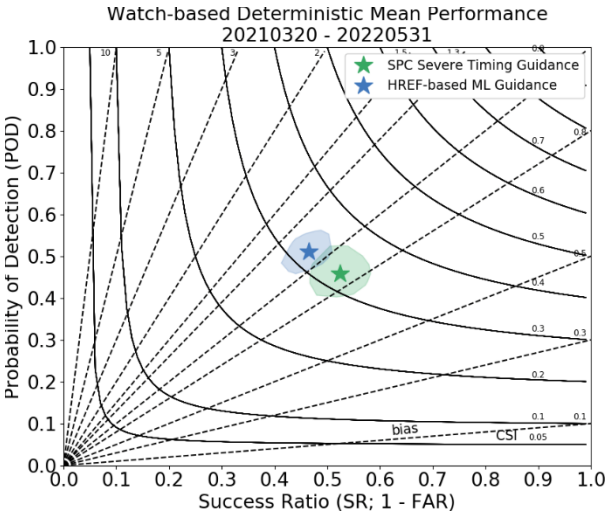

Figure 3: Mean performance of the 12z HREF-based ML and 13z SPC Severe Timing Guidance deterministic, first-guess, county-based watch predictions for 20 March 2021 - 31 May 2022. Shaded regions denote 95% confidence intervals from 10,000 bootstrapped samples.

Finally, Anthony and Leftwich (1992) proposed a modified calculation of FAR to assess the spatial false alarm of watch products. This modified FAR accounts for the spatial distribution of LSRs within a watch product by first mapping those reports to a 40-km grid. Next, an estimated area of impact is assigned to each LSR by defining a 200 x 200 km (5 x 5 40-km grid blocks) neighborhood centered on the report. Anthony and Leftwich (1992) then define the "good area percentage" (*A*) as the cumulative area of impact contained within a watch divided by the total area of that watch. The modified FAR of the watch guidance for a given convective day is then 1 - *A*, where *A* is spatially summed for all LSRs and predicted watch counties during the convective day. Note that the modified FAR proposed by Anthony and Leftwich (1992) also includes a temporal component which was excluded for this study, as the metrics were calculated on an hourly basis before being aggregated into daily verification scores.

These metrics were computed for both the HREF-based ML and Severe Timing Guidance watch products and averaged across the 14-month evaluation period. Operational SPC Tornado and

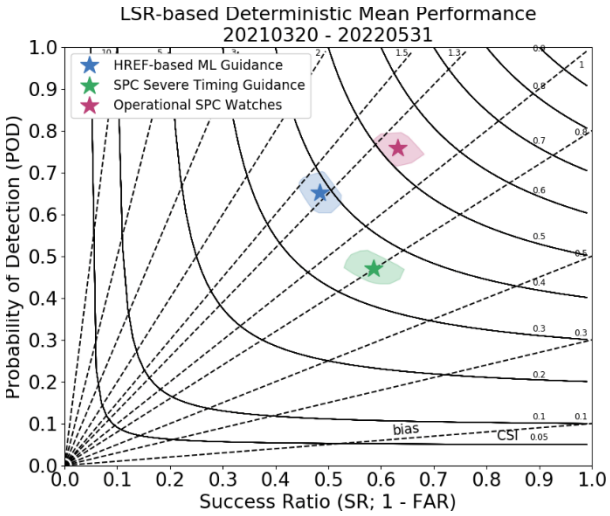

Figure 4: Mean performance of the 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch products evaluated against local storm reports for 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples.

Severe Thunderstorm Watches were also evaluated using this method, and the mean CSI was calculated from the POD and modified FAR (Fig. 4). As before, the ML and Severe Timing Guidance watch products achieved similar mean CSI scores of 0.39 (0.37 - 0.41) and 0.35 (0.33 - 0.38) respectively. However, more notable differences were observed between the products' POD and FAR. The ML watch guidance was found to have a mean POD of 0.65 (0.61 - 0.71), significantly greater than the POD of 0.48 (0.44 - 0.51) exhibited by the Severe Timing Guidance. Conversely, the Severe Timing Guidance demonstrated markedly improved FAR over the ML, with mean scores of about 0.41 (0.36 - 0.48) and 0.52 (0.48 - 0.57) respectively. From these scores, the ML guidance was determined to have an overforecast bias of about 1.3 (1.2 - 1.5) while the Severe Timing Guidance underforecast with a bias of 0.8 (0.7 - 0.9). These results are further supported by the mean PV of each model, with an average 37% of ML-predicted watch counties verifying with an LSR on a given convective day. In contrast, 57% of counties predicted by the Severe Timing Guidance verified with an LSR on average.

Both forecast guidance products exhibited notable skill at predicting observed severe weather hazards; however, the human-issued SPC Tornado and Severe Thunderstorm Watches outperformed the two algorithms by a considerable margin. Operational watches had a mean POD of 0.75 (0.71 - 0.79) and a mean FAR of 0.38 (0.32 - 0.42) over the 14-month evaluation period, resulting in an average CSI of about 0.53 (0.50 - 0.55). SPC watches did exhibit a tendency to overforecast with a mean bias of 1.2 (1.1 - 1.3), and about 51% of operational watch counties verified with an LSR on average. This impressive performance serves as a reminder of the skill and expertise of SPC forecasters and again demonstrates the importance of collaboration to ensure that expert knowledge is incorporated into product design. However, it should also be noted that the ML and Severe Timing Guidance products are forecasts based solely on 12z model output (which isn't available for processing until about 1530z), whereas SPC watches are typically issued based on real-time radar, satellite, and environment trends much closer to the occurrence of severe weather. As such, the difference in issuance time should be considered when comparing the performance of the first-guess and SPC-issued products.

4. HWT EVALUATION

An additional evaluation of the forecast watch products was included as part of the 2022 Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE), where a combination of operational forecasters, researchers, developers, and students provided their subjective assessment. After viewing first-guess watch forecasts from the ML and Severe Timing Guidance alongside operational SPC watches issued during the previous convective day, participants were presented with a survey with two primary questions. The first primary question asked respondents to subjectively rate how similar the placement and timing of the ML and Severe Timing Guidance watch products were to the operational Tornado and Severe Thunderstorm Watches issued by the SPC. Each product was assessed independently on a 5-point Likert scale with values ranging from "Not at all similar" to "Extremely

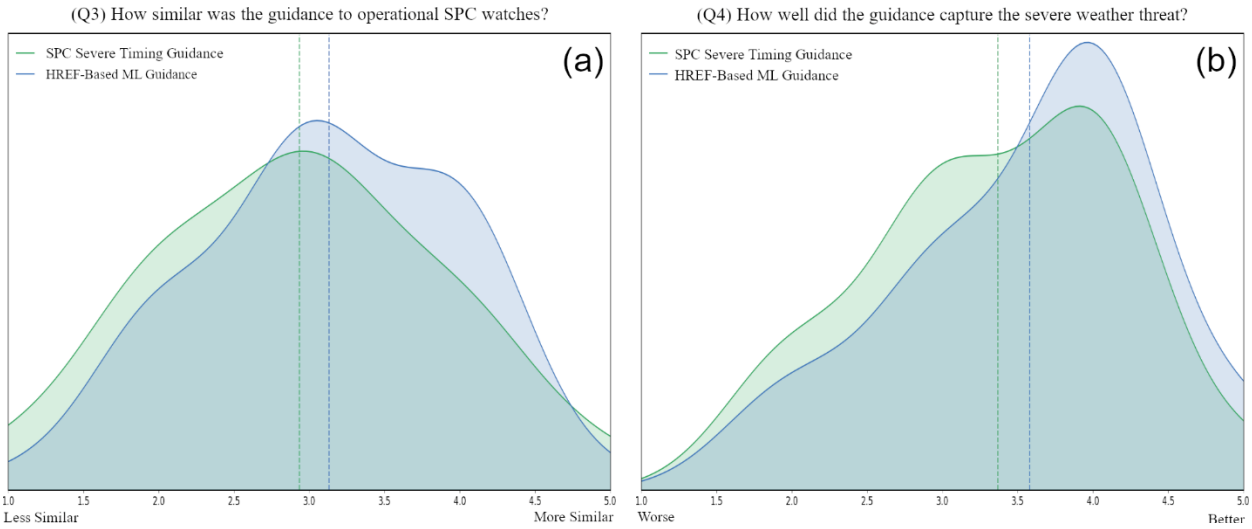

Figure 5: Survey responses approximated as kernel density estimation (KDE) curves. Dashed vertical lines represent the mean score for each guidance product.

similar." Respondents were instructed to consider the full 16-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. Similarly, the second primary question directed participants to subjectively evaluate how well the ML and Severe Timing Guidance watch products captured the location and timing of the severe weather threat during the available 16-hour forecast period. Again, the ML and non-ML products were independently assessed via a 5-point Likert scale ranging from "Terrible" to "Excellent." This evaluation was performed using overlaid LSRs and NWS storm-based warnings to indicate the observed location and time of severe weather occurrence. Additionally, respondents were instructed to only consider reports and warnings that fell within at least a 13z D1 SLGT to avoid penalizing the forecast products for not capturing severe hazards in locations where the guidance was systematically precluded from issuing forecasts.

Respondents were neutral on average when rating how similar the ML and Severe Timing Guidance first-guess watch products were to the SPC-issued Tornado and Severe Thunderstorm watches (Fig. 5a). The ML guidance received a bootstrapped mean score of 3.13 with a standard deviation of 0.82. Similarly, the non-ML guidance was given a mean rating of 2.93 with a standard deviation of 0.93. Differences between the two products were small and ultimately not statistically significant at the 95% confidence level; however, the distribution of survey responses does at least indicate a slight trend in favor of the ML-derived first-guess watch products. Approximately 77% of survey responses indicated the ML guidance was at least "moderately" similar to the SPC watches, and 36% of responses found it to be "very" or "extremely" similar. Conversely, the Severe Timing Guidance was at least "moderately" similar in 67% of responses and "very" or "extremely" similar in only 28% of the results.

ML and Severe Timing Guidance first-guess watch products were generally rated more favorably in regard to how well they captured the spatial and temporal domains of the true severe weather hazards, with bootstrapped mean scores of 3.58 and 3.37 respectively (Fig. 5b). Additionally, respondent agreement was nearly identical for both products as indicated by a standard deviation of 0.82 for the ML and 0.81 for the non-ML products. As before, these minute differences between the product ratings were not found to be statistically significant at the 95% confidence level, but the response distribution of the ML guidance again trended towards somewhat higher ratings than that of the Severe Timing Guidance. About 86% of responses stated that the ML first-guess watches captured the timing and spatial coverage of the

observed NWS warnings and LSRs with at least "average" skill, and 61% said the model performance was "good" or "excellent." In comparison, the Severe Timing Guidance performance was rated as "average" or better in 83% of responses and "good" or "excellent" in 48% of the results.

5. CONCLUSION

Both the ML and SPC Severe Timing Guidance first-guess watch products were found to be skillful at emulating human-issued SPC watches and capturing observed severe weather hazards. The ML-based approach tends to overforecast in both instances, with increased POD and FAR over the non-ML algorithm. Conversely, the Severe Timing Guidance demonstrated an underforecast of both SPC watches and observed severe weather hazards, resulting in decreased POD and FAR. These results are very encouraging, particularly for the ML guidance, but also suggest potential room for additional improvement. To further assess how the products perform in real time severe weather scenarios, future iterations of the first-guess watch guidance are expected to be run in real-time within SPC operations and presented to SPC forecasters via an experimental web interface. Frequent face-to-face collaborative discussions are also planned, and forecasters have already begun offering ideas for improvements and expansions of the current research.

6. ACKNOWLEDGMENTS

This extended abstract was modified from the doctoral dissertation Harrison (2022) and 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.

7. REFERENCES

Anthony, R. W. and P. W. Leftwich Jr, 1992: Trends in severe local storm watch verification at the national severe storms forecast center. *Wea. Forecasting*, **7**, 613–622, https://doi.org/10.1175/1520-0434(1992)007<0613:TISLSW>2.0.CO;2.

Doswell, C., 1999: Are warning lead times the most important issue in tornado events. *Weatherzine*, **17**, 3–3.

Gallo, B. T., J. K. Wolff, A. J. Clark, I. Jirak, L. R. Blank, B. Roberts, Y. Wang, C. Zhang, M. Xue, T. Supinie, et al., 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the finite-volume cubed-sphere (FV3) model core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

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.

Harrison, D. R., 2022: Machine Learning Co-Production in Operational Meteorology. Doctoral Dissertation. University of Oklahoma, https://hdl.handle.net/11244/335971.

Hoekstra, S., K. Klockow, R. Riley, J. Brotzge, H. Brooks, and S. Erickson, 2011: A preliminary look at the social perspective of warn-on-forecast: Preferred tornado warning lead time and the general public's perceptions of weather risks. *Wea. Climate Soc.,* **3**, 128–140, https://doi.org/10.1175/2011WCAS1076.1

Jirak, I. L., M. S. Elliott, C. D. Karstens, R. S. Schneider, P. T. Marsh, and W. F. Bunting, 2020: Generating probabilistic severe timing information from SPC outlooks using the HREF. *Severe Local Storms Symposium*, Amer. Meteor. Soc., Boston, MA, 3.1.

Krocak, M. J., J. T. Ripberger, H. Jenkins-Smith, and C. Silva, 2019: The impact of hours of advance notice on protective action in response to tornadoes. *Wea. Climate Soc.*, **11**, 881–888, https://doi.org/10.1175/WCAS-D-19-0023.1.

NWS, 2021: National severe weather products specification. National Weather Service Instruction 10-512.

Potvin, C. K., J. R. Carley, A. J. Clark, L. J. Wicker, P. S. Skinner, A. E. Reinhart, B. T. Gallo, J. S. Kain, G. S. Romine, E. A. Aligo, et al., 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT spring forecasting experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? Wea. Forecasting, 35, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

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.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: Facets: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

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.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

Stumpf, G. J. and A. E. Gerard, 2021: National weather service severe weather warnings as threats-in-motion. *Wea. Forecasting*, **36**, 627–643, https://doi.org/10.1175/WAF-D-20-0159.1.

Stumpf, G. J., S. Stough, and S. T. M, 2011: Examining potential improvements to severe weather warnings from a geospatial verification perspective. *First Conf. on Weather Warnings and Communication*, Amer. Meteor. Soc., Oklahoma City, OK, P1.4.