17.3B WARN-ON-FORECAST SYSTEM OUTPUT AS A VERIFICATION TOOL FOR SEVERE WIND EVENTS

Nathan A. Dahl*
University of Oklahoma/CIWRO and NOAA/NWS/Storm Prediction Center, Norman, OK

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

Kent H. Knopfmeier, Joshua Martin, and Brian C. Matilla
NOAA/National Severe Storms Laboratory, Norman, OK

1 INTRODUCTION

Lack of reliable verification is a serious impediment to ongoing efforts to improve severe wind forecasts. Various prior studies (e.g., Weiss et al. 2002, Trapp et al. 2006, Smith et al. 2013, Miller et al. 2016) have highlighted the potential for substantial mischaracterization of the spatial extent and intensity of severe wind events through reliance on in situ observations and local storm reports (LSRs). Observation networks and potential damage indicators vary widely by geographic region, and wind speed estimates (whether obtained from human observers at the time of the event or damage surveys conducted afterward) are notoriously error-prone. Doppler radar observations potentially afford better coverage but also suffer from errors due to beam elevation and wind-direction-relative azimuth; furthermore, there are substantial holes in the current WSR-88D network, particularly over the western United States.

Ideally, an objective analysis of the near-surface wind field in the region of interest, leveraging these (and other) information sources, would be sufficiently accurate to serve as "truth" for forecast verification. However, considering the rapid evolution and small-scale fluctuations that often play prominent roles in severe wind events, such an analysis would require rapid updates on a high-resolution grid. One promising candidate is the Warn-on-Forecast System (WoFS; Stensrud et al. 2009) maintained by the NOAA National Severe Storms Laboratory.

WoFS currently consists of 36 WRF-ARW members cycled every 15 minutes over a targeted (sub-CONUS) domain at 3-km grid spacing, potentially fulfilling both the spatial and temporal requirements of the task at hand. It has shown great promise in providing both raw and calibrated guidance for a variety of weather hazards (e.g., Yussouf and Knopfmeier 2019, Flora et al. 2019, Galarneau et al. 2022) and has played an integral role in the annual NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments over the past several years (Clark et al. 2020, 2022a, 2022b).

In particular, the investigation of WoFS as a severe wind verification tool was motivated by its assessment of the derecho that swept

* Corresponding author address: Dr. Nathan Dahl, Cooperative Institute for Severe and High-Impact Weather Research and Operations, 120 David L. Boren Blvd., Suite 2100, Norman, OK 73072. email: *nathan.dahl@noaa.gov*

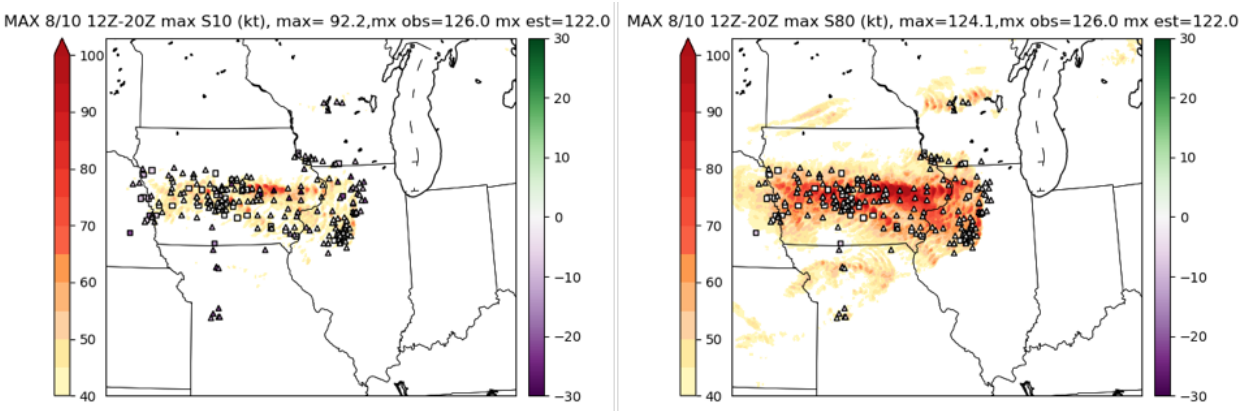

Figure 1 Comparison of WoFS wind speed and severe wind LSRs for the 10 August 2020 upper Midwest derecho. The color fill based on the left color bar depicts ensemble maxima of 15-minute forecast maximum 10 m wind speed (left) and 15-minute forecast instantaneous 80 m wind speed (right) from successive WoFS cycles from 12 to 20 UTC, while the markers show locations of "estimated gust" LSRs (triangles) and "measured gust" LSRs (squares). The marker fill color (based on the right color bar) indicates the difference between the LSR gust magnitude and the most similar WoFS wind speed within a 20 km radius.

through Iowa on the morning of 10 August 2020. Retrospectively, WoFS was cycled on a 600 x 400 grid for the period from 12 to 20 UTC, and the ensemble maxima of the 15-minute forecast maximum 10 m wind speed and the 15-minute forecast instantaneous 80 m wind speed for each cycle were stitched together and compared with locations and magnitudes of severe wind LSRs from the *Storm Data* archive as shown in Fig. 1. While the WoFS 10 m winds substantially underrepresented both the spatial extent and the magnitude of the damaging winds, the WoFS 80 m winds (used as an analog for the actual surface gusts) was highly accurate in both respects. These questions are considered herein: (1) Are those results representative of the general quality of the WoFS depiction of near-surface winds for potentially-severe situations?; (2) Can further improvements be realized through simple adjustments and/or machine learning?

2 DATA AND METHODS

The data set examined here consists of WoFS output for 132 event days from 2019 to 2021, most of them produced in conjunction with the annual HWT Spring Forecasting Experiment. As in the example shown previously, the 15-minute forecast maximum 10-m wind speed and the 15-minute forecast instantaneous 80-m wind speed for each member and each cycle were stored. (Unfortunately, the 15-minute forecast maximum 80 m wind speed was not available for the dates in question.) The magnitude, time, and location (mapped to the nearest grid point) of daily maximum ASOS wind speed measurements greater than 20 kt and severe wind LSRs listed in the *Storm Data* archive (with the "measured gust" label) within the WoFS domain were used for verification. The distribution of those observations is shown in Table 1.

For each gust observation, the largest WoFS ensemble 50th percentile, 75th percentile,

90th percentile, and maximum gust over a +/- 1 hr window around the observation time (to allow for errors in reporting) were obtained at the observation location and within 10 km and 20 km neighborhoods. In addition, "subgrid coverage" features were extracted within a 40 km neighborhood (corresponding to the grid spacing for the current SPC mesoanalysis). Specifically, the ensemble-averaged fractions of the neighborhood covered by 10 m or 80 m wind speeds over 20, 35, 50, or 65 kt were calculated, along with the fractions covered by radar reflectivity over 20, 30, or 40 dBZ and the fractions covered by 2-5 km updraft helicity over 10, 25, or 50 $m^{-2}s^2$.

Table 1 Distribution of wind speed observations used for verification

observed speed (kt)	sample size (n)
20-35	4704
35-50	806
50-65	506
65+	54
TOTAL	6070

The data were split 2/3 for training and calibration and 1/3 for testing. For the machine-learning experiments, a gradient

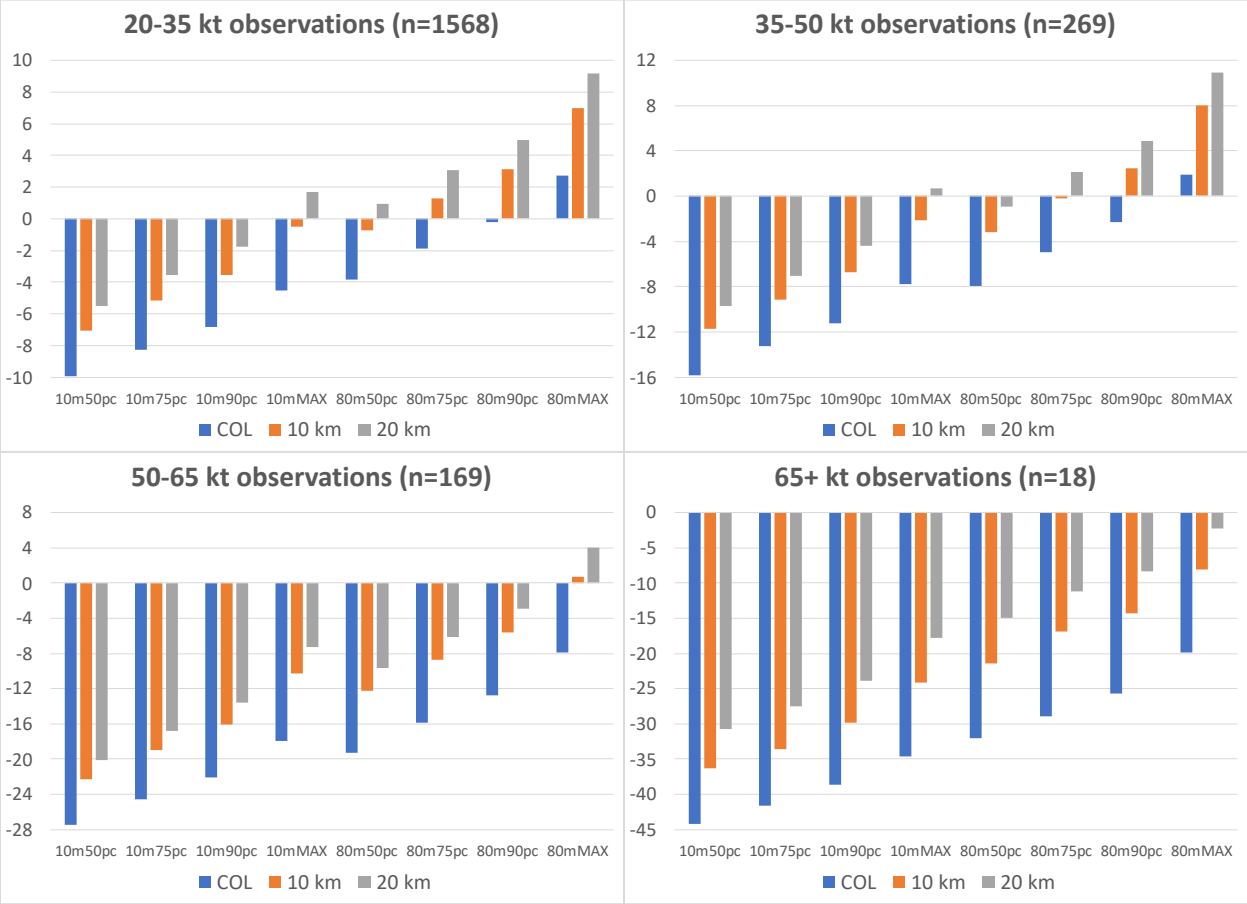

Figure 2 Bias within different observed wind speed bins for ensemble 50th percentile ("50pc"), 75th percentile ("75pc"), 90th percentile ("90pc"), and maximum ("MAX") of WoFS 10 m ("10m") and 80 m ("80m") wind speed, verified point-to-point ("COL", blue) or using the 10 km (orange) or 20 km (grey) neighborhood maximum.

boosting regressor (GBR) was used with five-fold cross validation for tuning. All splits were done chronologically and stratified by wind speed to preserve the sample imbalance shown in Table 1. (Additional machine-learning tests were attempted with oversampling of higher wind speed bins to produce an even distribution; however, the outcome did not substantially differ from the results shown here.)

3 RESULTS

3.1 Raw output and simple adjustments

First, note that all of these results pertain to the portion of the data set aside for testing. The bias and mean absolute error (MAE) for the raw WoFS output, verified point-to-point or using the 10 km or 20 km neighborhood maximum, are shown in Figs. 2 and 3. The WoFS 10-m wind showed a consistent strong negative bias and substantial MAE at all thresholds, worsening as the observed wind speed increased. The 80-m wind performed better overall, although the results were sensitive to the selection of a "representative" percentile and neighborhood size; for example, the ensemble maximum within a 10-km neighborhood had a strong positive bias and large MAE for winds below 50 kt but was the best choice for winds over 50 kt.

It is perhaps more pertinent to examine WoFS' ability to discriminate between winds below 35 kt and potentially damaging winds above 35 kt or severe winds above 50 kt. The

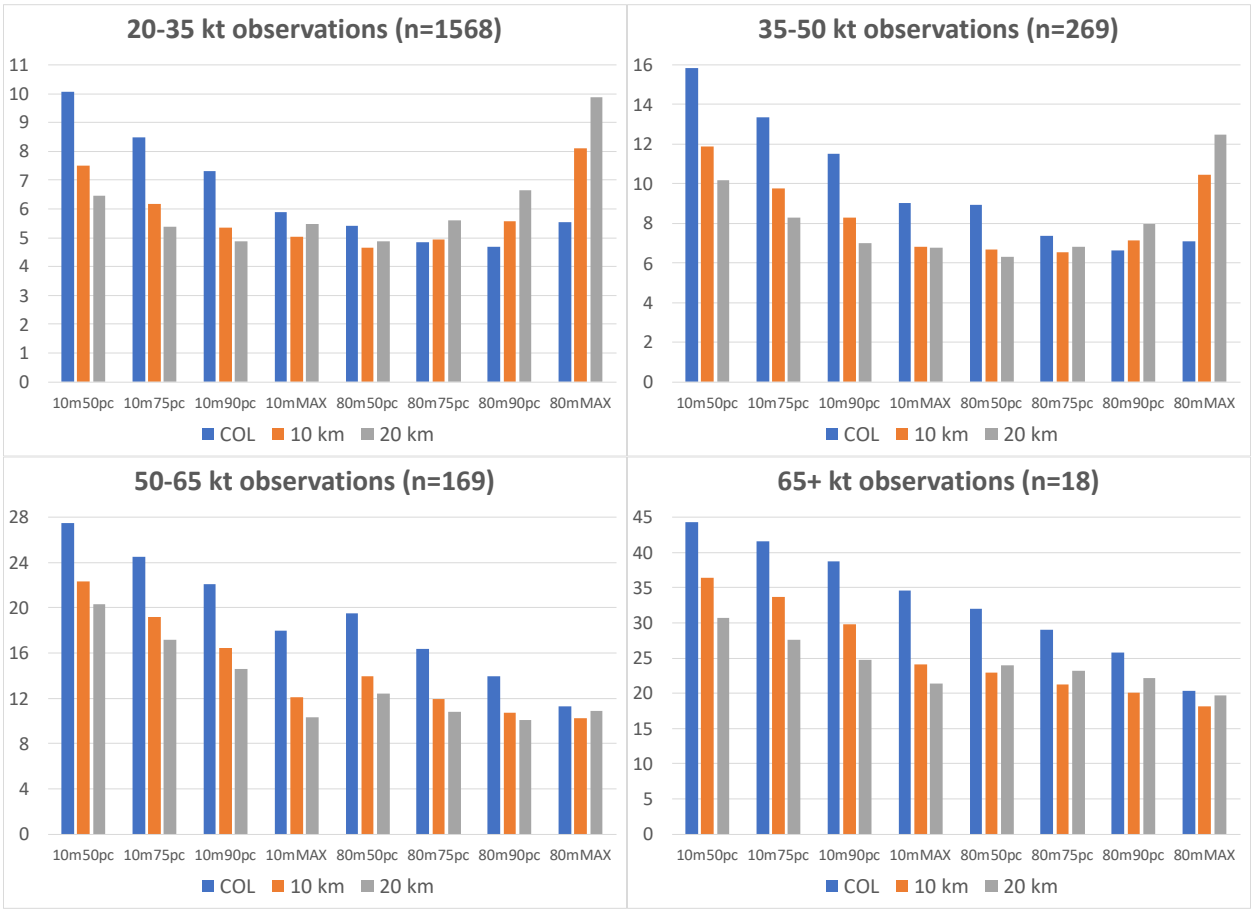

Figure 3 Same as Fig. 2, but for mean absolute error.

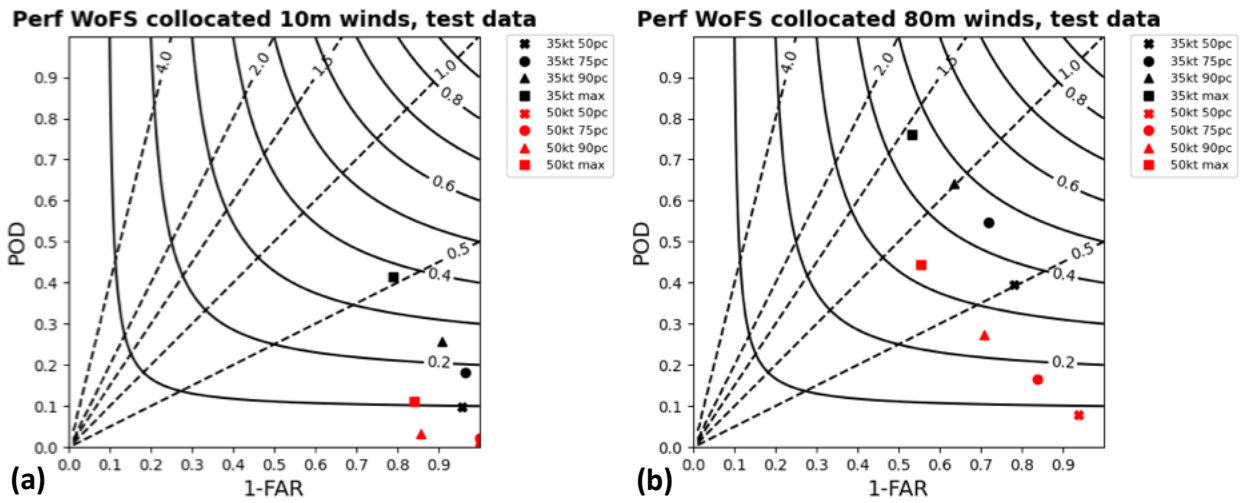

Figure 4 Performance diagrams for the use of the ensemble 50th percentile ("50pc", crosses), 75th percentile ("75pc", circles), 90th percentile ("90pc", triangles), and maximum ("max", squares) of WoFS 10 m winds (left) and 80 m winds (right) to detect winds over 35 kt (a) or 50 kt (b).

performance at detecting winds above 35 or 50 kt point-to-point is displayed in Fig. 4. As might be expected from the bulk error statistics shown earlier, the 10-m winds struggled with detection, particularly at the 50-kt threshold. The 80-m winds were substantially more skillful (in terms of Critical Success Index [CSI]), with the 90th percentile performing best for the 35 kt threshold and the ensemble max performing best for the 50-kt threshold. However, the best probability of detection (POD) for the 50-kt threshold was still under 0.5, which is not ideal for a verification source. Therefore, the discussion now turns to efforts to improve upon these results by through adjustments based on additional ensemble and neighborhood-based information.

First, seeking to account for the bias shown in Fig. 2, the assumption of equivalence between WoFS winds and observations (e.g., that a WoFS wind speed of 50 kt should correspond to an observed wind speed of 50 kt) was set aside and the most skillful combination of ensemble percentile and wind speed threshold was sought for the observed 35-kt and 50-kt thresholds. Note that, for these adjustment experiments, the settings were determined using the training data and then applied to the testing data.

For the WoFS 10-m wind, setting a 22-kt minimum for the 50th percentile worked best for discriminating observed winds above and below the 35-kt threshold (POD 0.64, CSI 0.49), while setting a 22-kt minimum for the 50th percentile worked best for the 50-kt threshold (POD 0.59, CSI 0.41); these values are much better than the raw results shown in Fig. 4a. For the WoFS 80-m wind, setting a 31-kt minimum for the 75th percentile worked best for the 35-kt threshold (POD 0.68, CSI 0.46), while setting a 46-kt minimum for the ensemble maximum worked best for the 50-kt threshold (POD 0.61, CSI 0.37); these values do not improve much on the raw results shown in Fig. 4b and actually produce lower CSI than the adjustments based on the 10-m winds.

A similar process was used to examine the use of WoFS wind speed, reflectivity, or updraft helicity "subgrid coverage" in discriminating at the observed 35-kt and 50-kt thresholds. The percentile, minimum value, and minimum subgrid coverage fraction that produced the greatest skill for the training data was applied to the testing data, and in no case did it improve upon the results already obtained. For example, using the optimal settings for the WoFS 10-m winds to discriminate at the observed 50-kt threshold (50th percentile, 35 kt subgrid coverage of 0.02 or greater) produced a POD of only 0.50 and a CSI of only 0.35.

Using neighborhood maximum wind speeds instead of verifying point-to-point yielded mixed results. Performance diagrams for the 10-km and 20-km neighborhoods are shown in Fig. 5. For the 35-kt threshold, the 10-m ensemble maximum and the 80-m 50th percentile were on par with one another and slightly more skillful than the best point-to-point result in Fig. 4 (i.e. using the 80-m 90th percentile). For the 50-kt threshold, the 10-m winds continued to struggle; the 80-m 90th percentile appeared best overall, improving somewhat on the best result in Fig. 4 (i.e., using the 80-m ensemble maximum.)

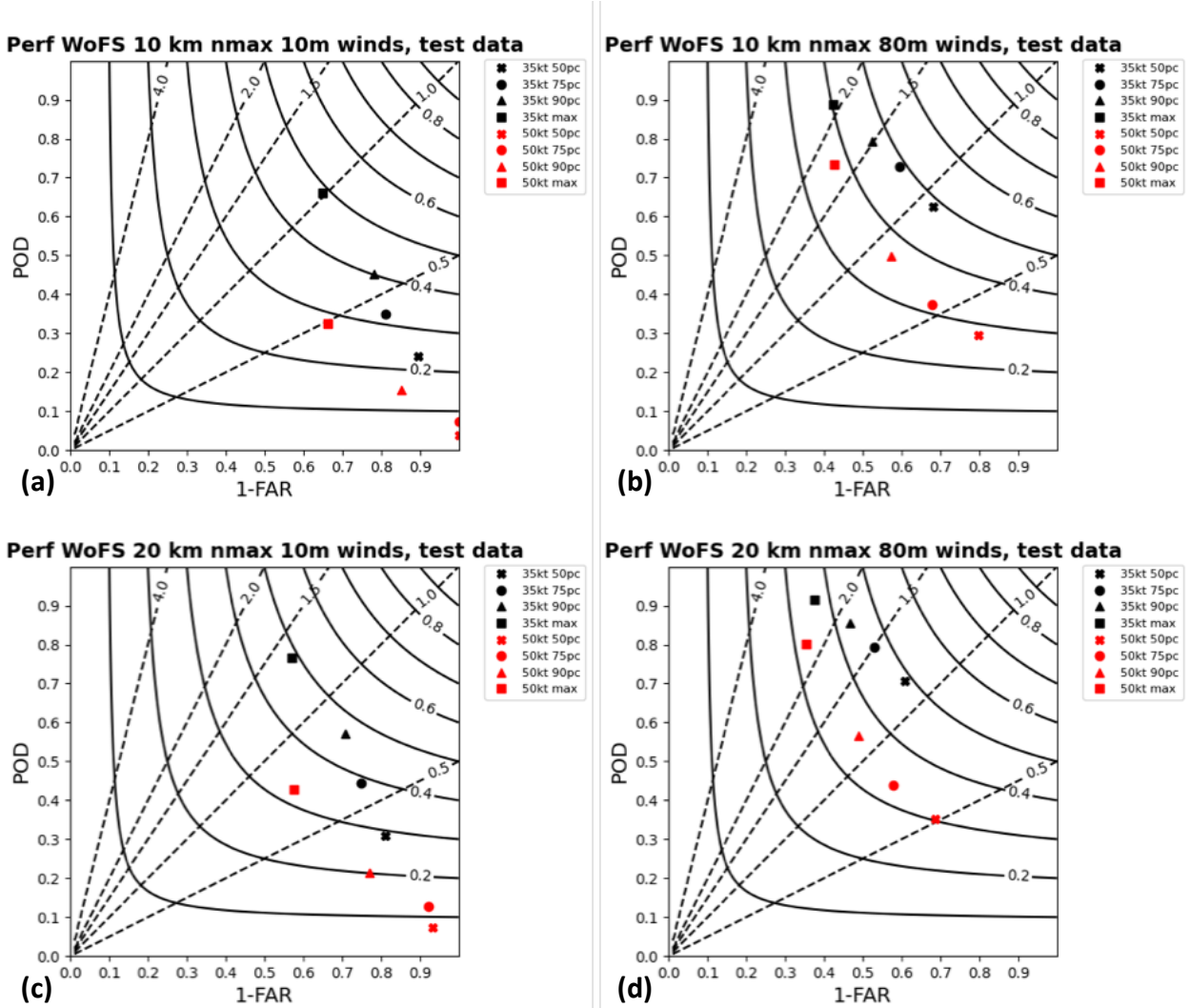

Figure 5 Same as Fig. 4, except for WoFs 10-km neighborhood maximum wind speeds at 10 m (a) and 80 m (b) and 20-km neighborhood maximum wind speeds at 10 m (c) and 80 m (d).

Ultimately, none of these simple adjustments produced a dramatic improvement in optimized skill. There are indications that further improvement based on local information may be possible, however. As shown in Fig. 6, if the "best" wind speed (i.e., closest to the observed value) within the neighborhood was used instead of the maximum wind speed, a verification-quality analysis would seem more attainable, with optimal CSI approaching 0.7 for the 35-kt threshold and 0.6 for the 50-kt threshold. Obviously, the feasibility of selecting of the "best" neighborhood wind speed *a priori* is dubious; nevertheless, the fact that WoFS usually produced a reasonable approximation of the truth in close proximity to the observation location motivated efforts to further benefit from proximity information through machine learning.

3.2 Machine learning results

The process of tuning hyperparameters produced five GBR models, one for each validation fold. All five models were applied to the test data, producing five output wind speeds which were then averaged to obtain the model wind speed corresponding to

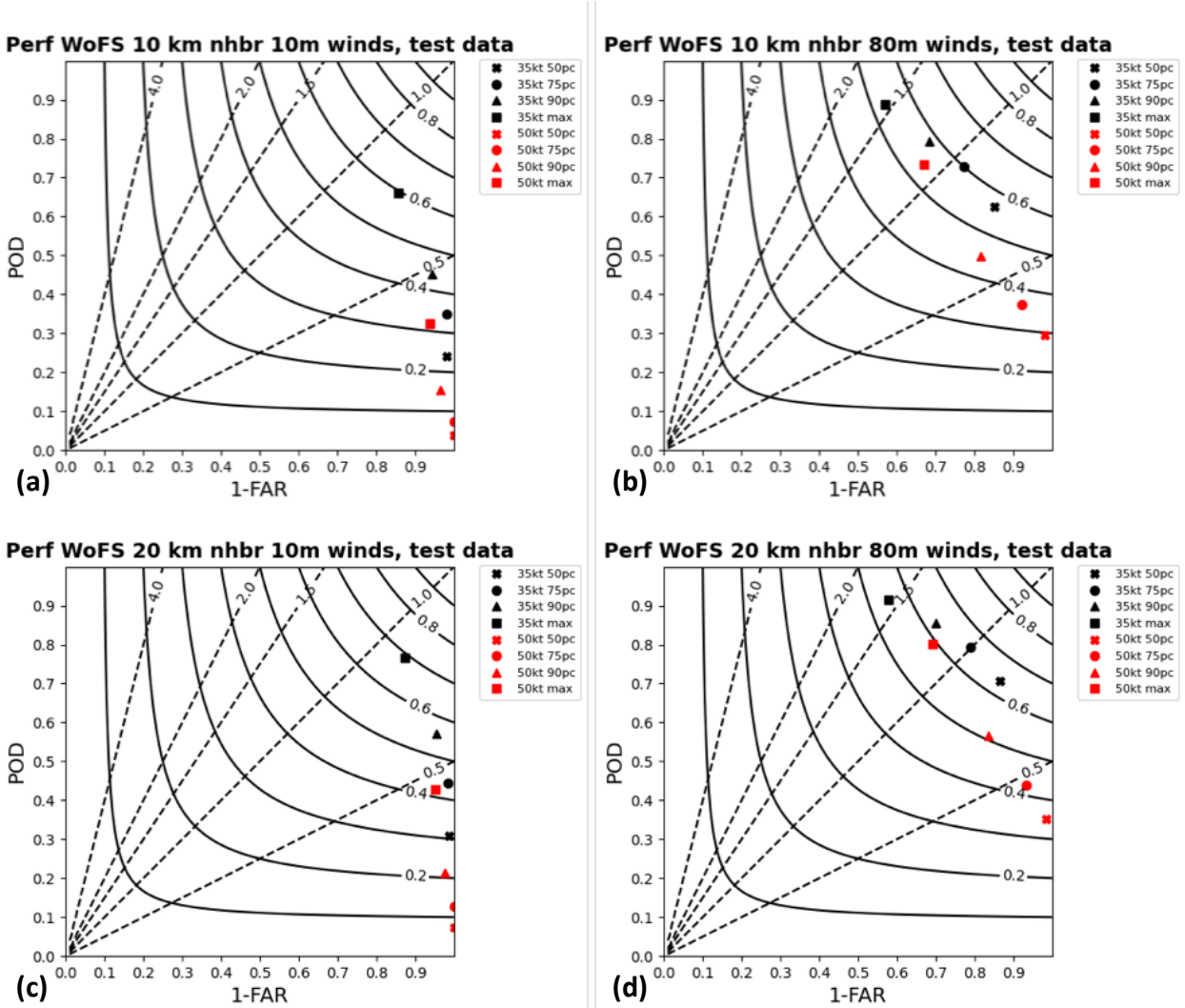

Figure 6 Same as Fig. 4, except for WoFs 10-km neighborhood "best" wind speeds at 10 m (a) and 80 m (b) and 20-km neighborhood "best" wind speeds at 10 m (c) and 80 m (d).

each observation. (Other methods, such as training a model on the full training dataset using the tuned hyperparameters from the cross-validation, produced similar outcomes to what is shown here.)

The result is displayed in Fig. 7a. First, as a side note, the scatter plot highlights a scarcity of observations between 40 and 50 kt. How much of this is due to the selection of data (i.e., strictly ASOS observations and LSRs, with no sub-severe observations from Mesonet stations or other sources that enter the *Storm Data* pool once the wind speed reaches 50 kt) and how much is due to rounding and/or misclassification of "estimated gust" reports as "measured gust" reports in the *Storm Data* archive is unclear. It is apparent that, while the model discriminates well at the 35-kt threshold, it has a pronounced low bias for observed winds of 50 kt or more. However, the cross-validation process produced intermediate results (viz. the output of the models for the validation folds) with their own biases. By calculating and inverting lines of best fit for the validation output for each model, a set of linear corrections was obtained which was then applied to the testing output.

The outcome is shown in Fig. 7b. The linear corrections dramatically improved the POD for severe wind observations while only modestly increasing the number of false alarms. Figure 8 compares the GBR performance with the earlier point-to-point verification of WoFS 80 m wind speeds. The GBR performance at the 35-kt threshold (which is not substantially affected by the linear corrections) improves noticeably, albeit modestly, upon prior efforts. At the 50-kt threshold, the raw GBR model performance is poor, driven (as expected) by a low POD. The linear corrections greatly improve the POD and CSI, giving the most skillful result of any method. However, the fundamental shortcomings noted in the previous section remain; in other words, applying machine learning, while potentially beneficial, did not produce a "silver bullet" in this case.

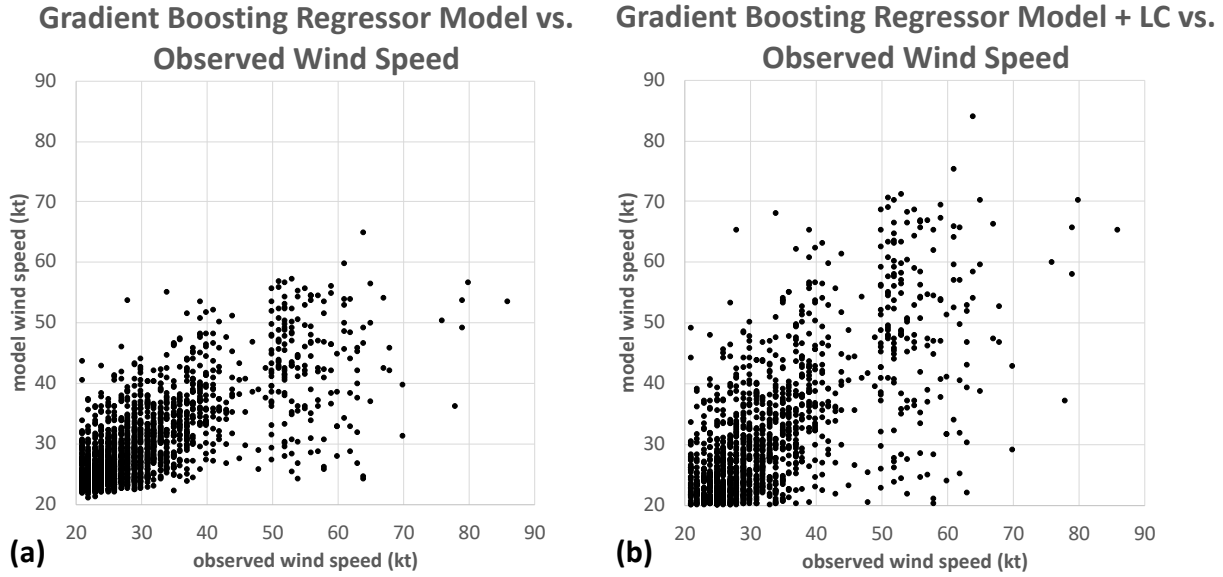

Figure 7 Scatter plots of model wind speed vs. observed wind speed for raw GBR output (a) and GBR output with linear corrections applied (b).

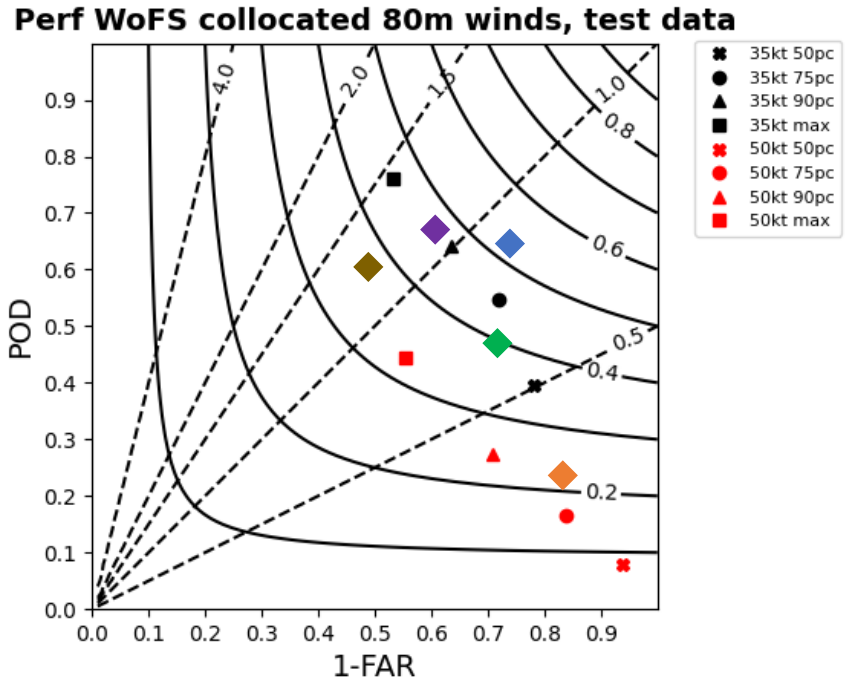

Figure 8 As in Fig. 4b, updated to include point-to-point detection of observed winds at or above 35 kt using a 31 kt minimum for the WoFS 75th percentile 80 m wind speed (purple), point-to-point detection of observed winds at or above 50 kt winds using a 46 kt minimum for the WoFS ensemble maximum 80 m wind speed (brown), detection of 35 kt winds using the raw GBR model output (blue), detection of 50 kt winds using the raw GBR model output (orange), and detection of 50 kt winds using the GBR model output with linear corrections applied (green).

4 Questions and Future Work

Several caveats must be included when evaluating these results. First, the sample size is clearly limited, particularly for observations at or above 50 kt. This mainly stems from the decision to only include measured gusts in order to focus on obtaining as accurate a near-surface wind speed analysis as possible. With regard to the LSRs excluded because they lacked a gust measurement, the authors are aware of ongoing efforts to develop methods for discriminating severe gusts from sub-severe gusts in those cases (e.g., the NOAA-ML project at Iowa State University) which could help alleviate the sample size issue in the future.

Second, the lack of 15-minute forecast maximum 80-m winds in the WoFS output used here may have artificially reduced the quality of the results. The forecast maximum was available for the 2022 Spring Forecast Experiment, and detailed analysis has not yet been done (with the *Storm Data* archive only recently updated to include reports through May 2022). However, preliminary evaluation of bulk error statistics and raw data performance using the forecast maximum for 2022 show a marked improvement over what is shown here for the instantaneous 80-m winds from previous years, although it is impossible (due, again, to small sample size) to know whether this finding will prove significant.

Finally, reviewing results day-by-day (not shown) tends to show that the WoFS wind analyses tend to do relatively well for larger-scale events (e.g. severe MCSs and derechos) and struggle for smaller-scale events (e.g., isolated or weakly-forced convection) for which the narrow wind maxima cannot be adequately resolved on the 3-km grid used by WoFS. While efforts to run WoFS at higher resolution are ongoing (e.g., Wang et al. 2022), it is possible that a probabilistic treatment of such events would yield better results in the meantime.

ACKNOWLEDGEMENTS

This extended abstract was prepared by Nathan Dahl with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA16OAR4320115, U.S. Department of Commerce. Kent Knopfmeier, Joshua Martin, and Brian Matilla were funded by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA16OAR4320115. The authors benefitted from collaboration with David Harrison, Burkely Gallo, and David Jahn and assistance from Storm Prediction Center and National Severe Storms Laboratory Information Technology staff, the Oklahoma Supercomputing Center for Education and Research, and user support from NCEP MADIS. 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.

REFERENCES

Clark, A., and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bull. Amer. Meteor. Soc.*, **101**, E2022-E2024.

Clark, A., and Coauthors, 2022a: The 2nd Real-Time Virtual Spring Forecasting Experiment to Advance Severe Weather Prediction. *Bull. Amer. Meteor. Soc*., **103**, E1114-E1116.

Clark, A., and Coauthors, 2022b: The 3rd Real-Time, Virtual Spring Forecasting Experiment to Advance Severe Weather Prediction Capabilities. *Bull. Amer. Meteor. Soc.*, DOI: 10.1175/BAMS-D-22-0213.1.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast System. *Wea. Forecasting*, **34**, 1721-1739.

Galarneau, T. J., Jr., L. J. Wicker, K. H. Knopfmeier, W. J. S. Miller, P. S. Skinner, and K. A. Wilson, 2022: Short-Term Prediction of a Nocturnal Significant Tornado Outbreak Using a Convection-Allowing Ensemble. *Wea. Forecasting*, **37**, 1027-1047.

Miller, P. W., A. W. Black, C. A. Williams, and J. A. Knox, 2016: Maximum wind gusts associated with human-reported nonconvective wind events and a comparison to current warning issuance criteria. *Wea. Forecasting*, **31**, 451-465.

Smith, B. T., T. E. Castanellos, A. C. Winters, C. M. Mead, A. R. Dean, and R. L. Thompson,

2013: Measured severe convective wind climatology and associated convective modes of thunderstorms in the contiguous United States, 2003-09. *Wea. Forecasting*, **28**, 229-236.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast System: A vision for 2020. *Bull. Amer. Meteor. Soc*., **90**, 1487-1500.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408-415.

Wang, Y., N. Yussouf, C. A. Kerr, D. R. Stratman, B. C. Matilla, 2022: An experimental 1-km Warn-on-Forecast System for Hazardous Weather Events. *Mon. Wea. Rev.*, **150**, 3081-3102.

Weiss, S. J., J. A. Hart, and P. R. Janish, 2002: An examination of severe thunderstorm wind report climatology: 1970-1999. Preprints, *21st Conf. on Severe Local Storms,* San Antonio, TX, Amer. Meteor. Soc., 446-449.

Yussouf, N., and K. H. Knopfmeier, 2019: Application of Warn-on-Forecast system for flash-flood producing heavy convective rainfall events. *Quart. J. Roy. Meteor. Soc.*, **145**, 2385-2403.