2019 DTC Visitor Program Final Report:

Improving Heavy Rainfall and Severe Weather Guidance from Convection-Allowing Ensembles Using Machine Learning

Eric D. Loken[1,2] and Adam J. Clark[2,3]

[1]Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma
[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[3]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

(Email: eloken@ou.edu)

1. Background and introduction

Despite their attempts to account for model and initial condition errors, numerical weather prediction (NWP) ensembles provide imperfect convection-related forecasts. For example, due to small ensemble membership, many ensembles have sub-optimal reliability and are under-dispersive for precipitation forecasts (e.g., Schwartz et al. 2014). Ensembles may also suffer from precipitation spatial displacement errors if they have imperfect forecasts of convective initiation and evolution. Moreover, even convection-allowing ensembles (CAEs) lack horizontal grid-spacing fine enough to explicitly simulate severe weather hazards. Thus, forecasting severe weather from CAEs typically requires the use of proxies correlated with observed severe weather, such as large values of 2-5 km updraft helicity (UH; e.g., Kain et al. 2008; Sobash et al. 2011, 2016, 2019) and upward vertical velocity (e.g., Roberts et al. 2019).

Machine learning (ML) techniques offer a promising way to post-process imperfect CAE precipitation and severe weather forecasts, since these techniques can learn multi-variate, flow-dependent, non-linear relationships between ensemble forecast variables and observed precipitation or severe weather. The random forest (RF; Breiman 2001) algorithm may be particularly well-suited for post-processing because it typically produces reliable output probabilities even when input predictors are biased and/or nonlinearly related to the predictands (Breiman 2001).

Several studies have previously demonstrated the promise of RF-based post-processing. For example, Gagne et al. (2014) used a RF to forecast probabilistic precipitation from a 19-member convection-allowing ensemble produced by the Center for Analysis and Prediction of Storms (CAPS), but they only analyzed forecasts over the eastern 2/3 of the contiguous United States (CONUS) and had to artificially under-sample low-precipitation forecast points during training. Meanwhile, Herman and Schumacher (2018) used multiple RFs to post-process CONUS-wide day 2 and 3 heavy precipitation forecasts from the Global Ensemble Forecast System Reforecast (GEFS/R) system but did not test their methods on a CAE for day 1 (i.e., 12-36-h) lead times.

The work done in this visit expanded on Gagne et al. (2014) and Herman and Schumacher (2018) to post-process high-resolution ensemble precipitation and severe weather forecasts for the CONUS at day 1 (i.e., 1200 UTC – 1200 UTC; 12-36 h) time and space scales. The primary goal of this work was to develop and analyze a first-guess RF-based product to post-process and summarize ensemble output for operational forecasters. To that end, the RF-based post-processing was applied to the High-Resolution Ensemble Forecast System version 2 (HREFv2; Jirak et al. 2018; Roberts et al. 2019) and the Short-Range Ensemble Forecast System (SREF; Du et al. 2015) for precipitation prediction and the Storm Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016) for severe weather prediction. RF post-processing was also applied to the Warn-on-Forecast Ensemble System (WoFS; formerly the NEWS-e; e.g., Wheatley et al. 2015; Jones et al. 2016; Skinner et al. 2018; Flora et al. 2019) for predicting springtime severe weather at regional and sub-daily scales.

2. Objectives and outcomes

Major objectives of the visit included the following:

(a) Create and verify CONUS-wide post-processed, first-guess probabilistic guidance for precipitation and severe weather.

(b) Make post-processed forecasts available to operational forecasters via the Storm Prediction Center's (SPC's) public HREF viewer website (http://www.spc.noaa.gov/exper/href/).

(c) Investigate the impact of training dataset length on RF-based forecast skill to determine the feasibility of applying RF-based post-processing operationally.

(d) Transfer expertise to the Developmental Testbed Center (DTC) on the use of publicly-available, python-based ML codes.

(e) Disseminate results to the research community through peer-reviewed journal articles and conference proceedings.

With respect to these objectives, the following outcomes were achieved:

(a) A RF-based post-processing technique was developed and analyzed for next-day (1200 UTC – 1200 UTC; 12-36 h) probabilistic precipitation and severe weather forecasts from multiple ensembles. Post-processed forecasts were found to be reliable and skillful and tended to compare well against common operational baselines (see *Summary of results* below).

(b) 0000 UTC probabilistic precipitation forecasts are running daily on the public SPC HREF viewer website (http://www.spc.noaa.gov/exper/href/; accessible by selecting the 00:00 UTC run and clicking "RF probs" under the Precipitation tab).

(c) Encouragingly, substantial post-processing skill was noted with approximately 3 months of training data (see *Summary of results* below).

(d) The following steps were taken to transfer ML expertise to the DTC:

1. A beginner-level machine learning tutorial was developed to teach the basics of Scikit-Learn and Keras to scientists interested in applying ML to their own work. While the tutorial was presented to DTC scientists at the end of the visit (i.e., January 2020), the presentation and relevant codes from the tutorial remain accessible on Cheyenne in the following directory: "/glade/work/eloken/ML_tutorial_final/files". The tutorial covers Scikit-Learn random forests and neural networks as well as deep learning with Keras.

2. It is intended that the tutorial materials and other relevant codes will be uploaded to a Github repository (eloken-weather/NCAR_ML) to allow more permanent access by DTC scientists.

(e) The following peer-reviewed journal article was published in *Weather and Forecasting* during the visit:

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Post-processing next-day ensemble probabilistic precipitation forecasts using random forests. *Weather and Forecasting*, **34**, 2017-2044.

The following article has also been submitted to *Weather and Forecasting* and is currently undergoing peer-review:

Loken, E. D., A. J. Clark, and C. D. Karstens, in review: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. Submitted to *Weather and Forecasting.*

The submission of an additional *Weather and Forecasting* article is planned by mid-2020 on the application of RF post-processing to the WoFS.

The following conference proceedings were also given on work done in the visit:

Loken, E. D., A. J. Clark, A. McGovern, and K. Knopfmeier, 2019: Post-processing 12-36 hour multi-model ensemble PQPFs using a random forest. *1st NOAA Workshop on Leveraging AI in the Exploitation of Satellite Earth Observations & Numerical Weather Prediction*, College Park, Maryland, NOAA.

Loken, E. D., and A. J. Clark, 2020: Generating ensemble-derived next-day probabilistic severe weather forecasts with machine learning. *19th Conf. on Artificial Intelligence for Environmental Science*, Boston, MA.

Clark, A. J., E. D. Loken, P. S. Skinner, and K. H. Knopfmeier, 2020: Machine-learning-derived severe weather probabilities from a Warn-on-Forecast System. *10th Conf. on Transition of Research to Operations*, Boston, MA.

3. Summary of results

A RF-based method was designed to post-process next-day (12 – 36 h) precipitation and severe weather forecasts. The basic approach for both precipitation and severe weather applications is as follows: first, take a temporal aggregation (e.g., 24-h mean or maximum) of the predictors on the native grid; then remap all variables (predictors and observations) to a coarser grid (20-km for precipitation, 80-km for severe weather) for verification; next, take ensemble statistics (e.g., an ensemble mean) on the coarse grid for use as predictors; and, if necessary, randomly sample points in the domain to reduce the dimensionality of the dataset. The final set of predictors for each point includes forecast variables from the point of prediction as well as the 24 closest grid points. This general procedure is summarized, as it applies to the precipitation forecasts, in Fig. 1.

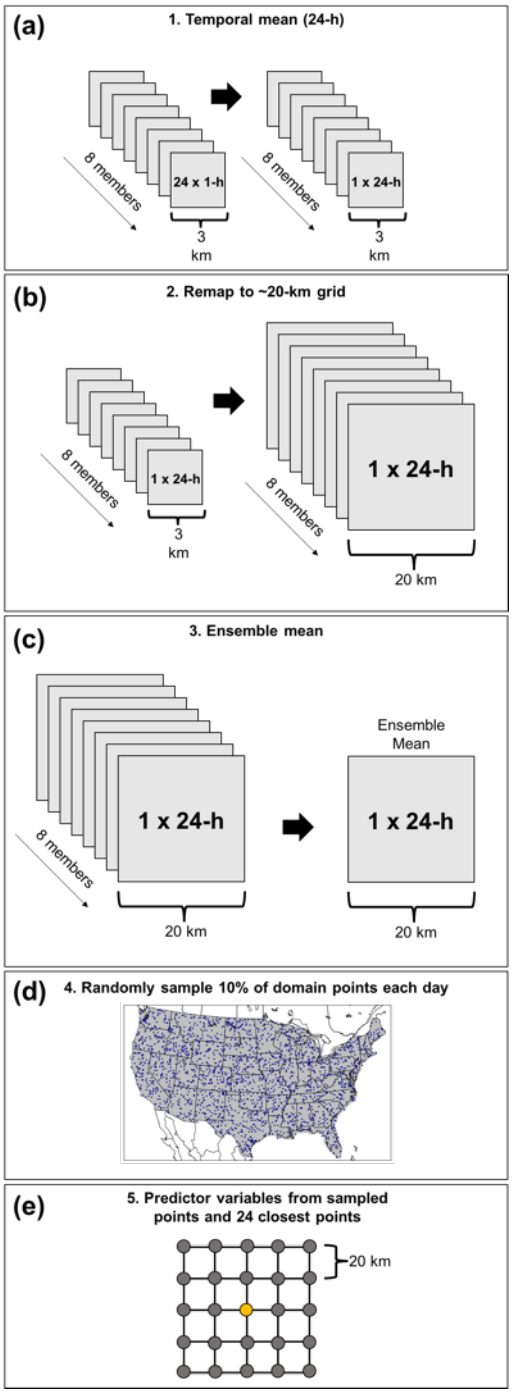

Figure 1 *Schematic illustrating the RF preprocessing for the HREFv2 for precipitation prediction. (a) The temporal mean is taken over 24-h at each native grid point for each ensemble member. (b) The temporally-averaged data is remapped to an approximately 20-km grid. (c) An ensemble mean is taken at each 20-km grid point. (d) 10% of the domain is randomly sampled for training. (e) Training data consists of the predictor variables at each sampled point (yellow) and the 24 closest 20-km points.*

a. Precipitation post-processing

The above RF-based method was applied to 0.1-, 0.5-, 1.0-, and 3.0-inch precipitation exceedance forecasts from the HREFv2—a 7-member, 3-km horizontal grid-spacing CAE—and the SREF–a 26-member, 16-km horizontal grid-spacing convection parameterizing ensemble. RF-based forecasts from each ensemble were compared against the simple fraction of ensemble members exceeding the threshold (hereafter referred to as raw baseline probabilities) and raw baseline probabilities smoothed spatially using an isotropic 2-dimensional Gaussian kernel density function (hereafter referred to as smoothed baseline probabilities). Forecast and observation data over 496 days from April 2017 – November 2018 were considered. 16-fold cross-validation with 31 days per fold was used to verify all forecasts. Predictors included 18 (16) forecast variables from the HREFv2 (SREF; Table 1), while the NCEP Stage IV Precipitation dataset was used for the observations.

Predictor Variable	**Atmospheric Level**
Temperature	500-, 700-, 850-hPa, and 2-m AGL
Dewpoint Temperature	500-, 700-, 850-hPa, and 2-m AGL
Max. Hourly Simulated Reflectivity*	1 km AGL
CAPE	Surface-based
CIN	Surface-based
PWAT	Entire Column
Max. Hourly Simulated UH*	2-5 km AGL
Max. Hourly U, V Wind	10 m AGL
Max. Hourly Upward Vertical Velocity (UVV), Downward Vertical Velocity (DVV)	100-1000 hPa (400-1000 hPa for NAM members of HREFv2)
Forecast 24-h Precipitation	Surface
Lat., Lon.	N/A

Table 1 *Predictor variables from the HREFv2 and SREF for precipitation prediction. Asterisks denote variables only used for the HREFv2.*

Skill benefits [i.e., greater area under the relative operating characteristics curve (AUC), higher Brier Skill Scores (BSSs), lower Brier Score (BS) reliability values, and higher BS resolution values] tended to be greater for the smaller precipitation thresholds (Fig. 2a-p), likely since the smaller precipitation thresholds provided more positive exceedance cases from which the RF could learn. Additionally, the SREF benefited more from the RF post-processing, since the SREF had more biases and less initial skill compared to the HREFv2.

A 1-sided paired permutation test (e.g., Good 2006) was used to test whether the AUC and BSS values from one set of forecasts was significantly greater than those from another set of forecasts. The procedure is explained in depth in Loken et al. (2019), and results are summarized in Fig. 3. Overall, compared to corresponding raw and smoothed ensemble probabilities, the RF-based forecasts tended to have significantly greater AUCs (Fig. 3a,c,e,g) and BSSs (Fig. 3b,d,f,h). The RF forecasts also had better reliability (Fig. 2i-l) and resolution (Fig. 2m-p) and fewer spatial biases (not shown), although statistical significance was not assessed for those metrics.

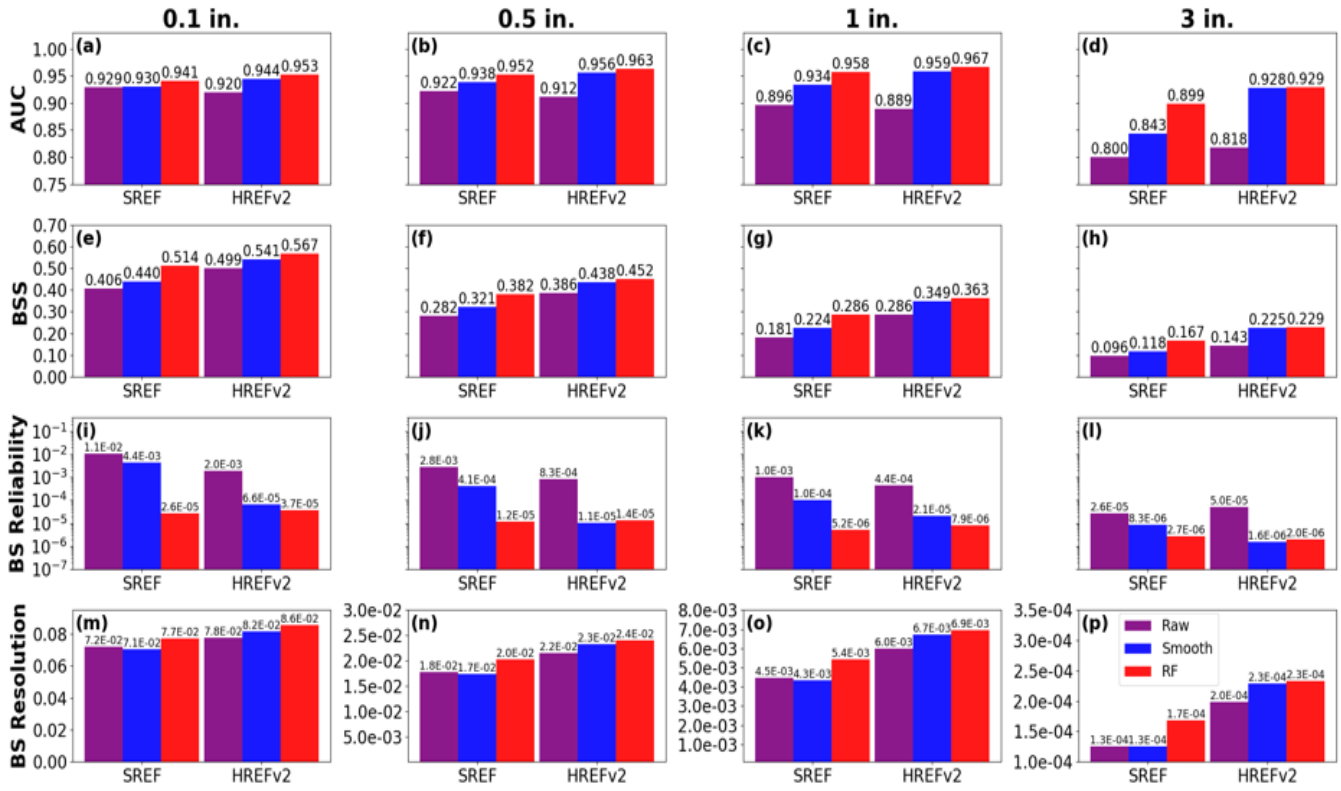

Figure 2 *AUC for SREF and HREFv2 raw (purple), smooth (blue), and RF forecasts (red) for the 0.1-in. threshold. (b)-(d) As in (a) but for the 0.5-, 1-, and 3-in. thresholds, respectively. (e)-(h) As in (a)-(d) but for BSS. (i)-(l) As in (a)-(d) but for the reliability component of the BS. (m)-(p) As in (a)-(d) but for the resolution component of the BS. Note the different y-axes for (m)-(p), and note that lower values of BS reliability are better.*

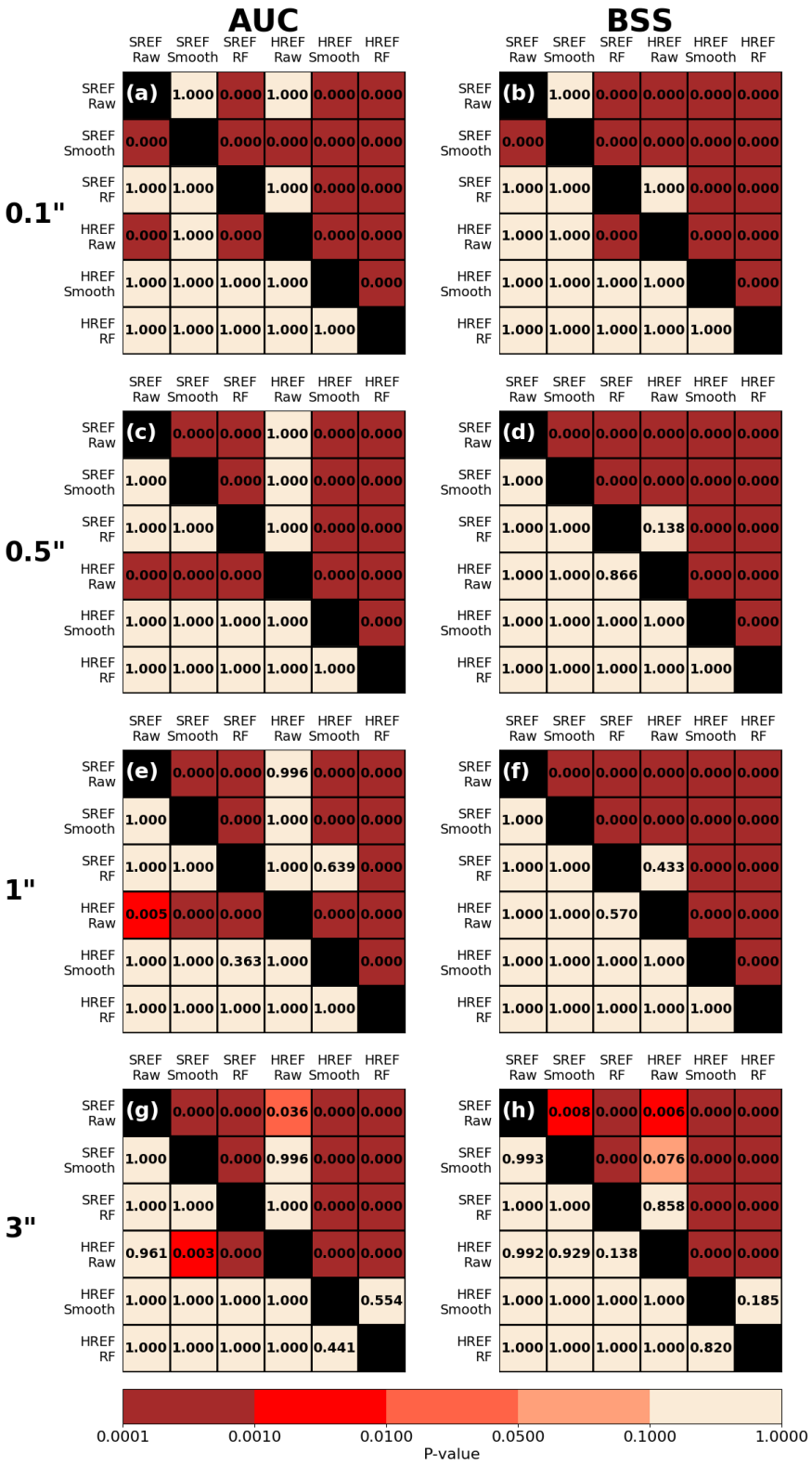

Figure 3 *(a) P-values from the 1-sided AUC permutation tests for the 0.1-inch threshold. (b) As in (a) but for BSS. (c)-(d), (e)-(f), (g)-(h) As in (a)-(b) but for the 0.5-, 1-, and 3-inch thresholds, respectively. Each square shows the p-value from testing whether the forecast in the top row has a significantly greater metric than the forecast in the left column.*

To test the impact of dataset length on RF skill, RFs were re-trained and re-evaluated using a dataset containing the first 62, 124, 248, and 372 days (i.e., the first 1/8, 1/4, 1/2, and 3/4) of the full dataset, respectively. Big gains in AUC and BSS were noted by increasing the dataset from 62 to 124 days, especially for the higher exceedance thresholds (Fig. 4). Large gains in skill were also seen by increasing the dataset to 248 days, but further increases did not noticeably impact skill. Because the dataset size in Fig. 4 includes a 31-day testing set, the results suggest adequate AUCs and BSSs with only about 3 months (93 days) of training data and substantial skill with only about 7 months (217 days) of training data.

Ultimately, the full dataset (496 days) was used to train 4 RFs (i.e., RFs predicting 0.1-, 0.5-, 1.0-, and 3.0-inch exceedances, respectively) for real-time prediction. Real-time RF exceedance forecasts are currently produced daily from the 0000 UTC HREFv2 forecasts and are publicly available via the HREF viewer (http://www.spc.noaa.gov/exper/href/).

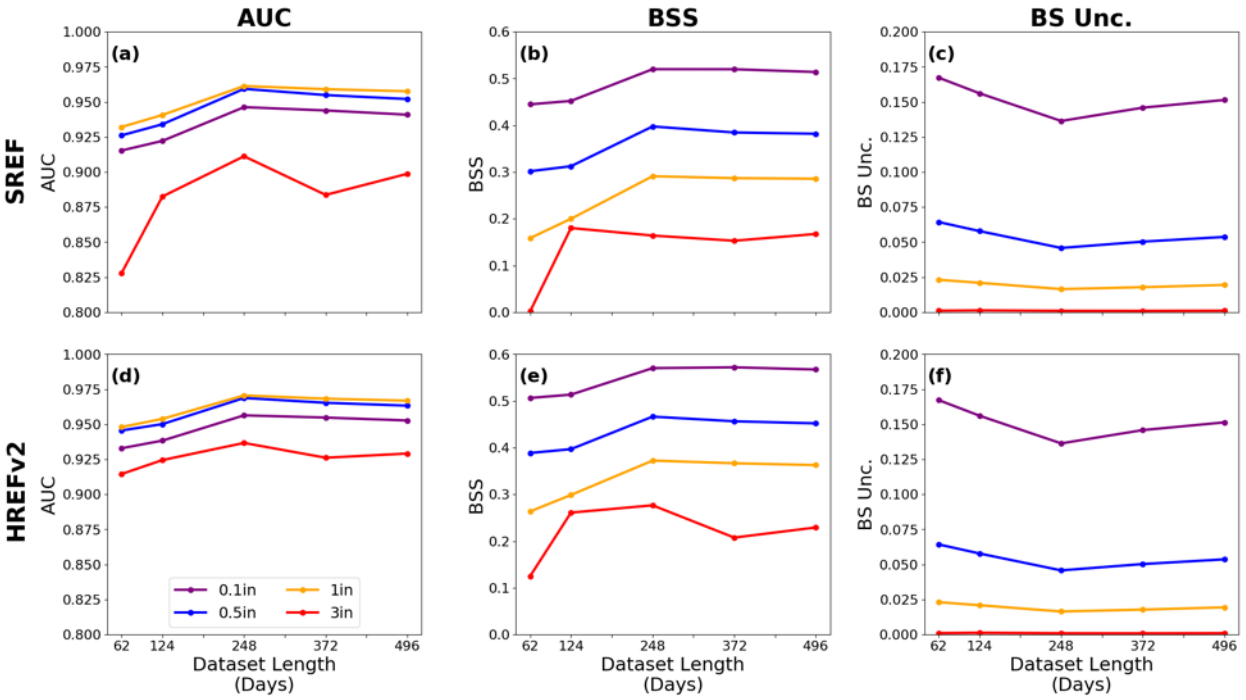

Figure 4 *(a) AUC as a function of dataset length for the SREF. (b)-(c) As in (a) but for the BSS and uncertainty component of the BS, respectively. (d)-(f) As in (a)-(c) but for the HREFv2.*

b. Severe weather post-processing

A similar RF-based post-processing technique was applied to two CAEs for severe weather prediction. These CAEs include the 7-member SSEO and the 18-member WoFS. SSEO members each have 4-km horizontal grid-spacing but have different initial and lateral boundary conditions, initialization times, and microphysics and turbulence parameterizations. Meanwhile, the WoFS, which is run experimentally during annual Hazardous Weather Testbed Spring Forecasting Experiments (HWTSFEs; e.g., Clark et al. 2012; Gallo et al. 2017), uses 3-km

horizontal grid-spacing and covers a moveable 900 x 900 km domain. It is run every 30 minutes out to 6-h and assimilates radar, satellite, and surface observation data every 15 minutes.

i) Post-processing SSEO data

Eight RFs were trained to provide probabilistic severe weather guidance based on SSEO forecast output and observed Storm Prediction Center (SPC) storm reports. RFs respectively predict: tornadoes, significant tornadoes (i.e., those with an Enhanced Fujita rating ≥ 2), severe wind [i.e., wind speed ≥ 50 kts (58 mph)], significant severe wind [i.e., wind speed ≥ 65 kts (75 mph)], severe hail (i.e., hailstone diameter ≥ 1 inch), significant severe hail (i.e., hailstone diameter ≥ 2 inches), all-hazards severe weather, and all-hazards significant severe weather. Predictors include storm attribute and environmental fields as well as latitude and longitude and spatially smoothed UH probabilities (Table 2).

Storm Attribute Fields	**Environment-related Fields**	**Other**
Max. Hourly Simulated Reflectivity	2-m Temperature	Latitude
Accumulated 1-h Precipitation	2-m Dewpoint Temperature	Longitude
Max. Hourly Updraft Speed	2-m Relative Humidity	Smoothed UH probabilities
Max. Hourly UH	MUCAPE	
	CIN	
	0-6 km Shear	
	CAPE × Shear	

Table 2 *SSEO-based predictors for obtaining severe weather probabilities.*

Probabilistic individual-hazard predictions were then used to construct categorical day 1 outlooks using the same criteria as the SPC. Continuous RF probabilities were compared against: continuous probabilities based on spatially smoothing the fraction of members exceeding a 2-5 km updraft helicity (UH) threshold (calibrated for each hazard), discrete and continuous (Karstens et al. 2019) SPC probabilities from the 0600 UTC day 1 convective outlook, and discrete RF probabilities truncated at the same probability levels used by the discrete SPC forecasts. The dataset contained 629 days, from April 2015 – July 2017. 17-fold cross-validation (with 37 days per fold) was used to verify the forecasts. 95% BSS confidence intervals (CIs) were determined by resampling each forecast's individual-day BS score with replacement (i.e., bootstrapping; e.g., Wilks 2011). 10,000 bootstrapping iterations were performed for each set of forecasts. The 95% CIs were recorded by noting the 2.5- and 97.5 percentile BSS values from all iterations for each type of forecast.

Overall, the continuous RF probabilities produced the best BSS for each hazard, drastically outperforming the UH forecasts for almost all hazards and the SPC forecasts for severe wind and hail (Fig. 5a). While the truncated RF did not convincingly outperform the discrete SPC forecasts for significant severe hazards, the continuous RF did for significant severe wind and hail (Fig. 5a) due to its superior resolution (Fig. 5b) and reliability (Fig. 5c), resulting from its ability to forecast continuous probabilities below 10%.

The categorical outlooks produced by the RF (e.g., Fig. 6a) generally compared favorably to those produced by the SPC at 0600 UTC (e.g., Fig. 6b). Thus, it is envisioned that these CAE-derived categorical outlooks could be used by forecasters as a skillful automated "first-guess" upon which forecasters could improve.

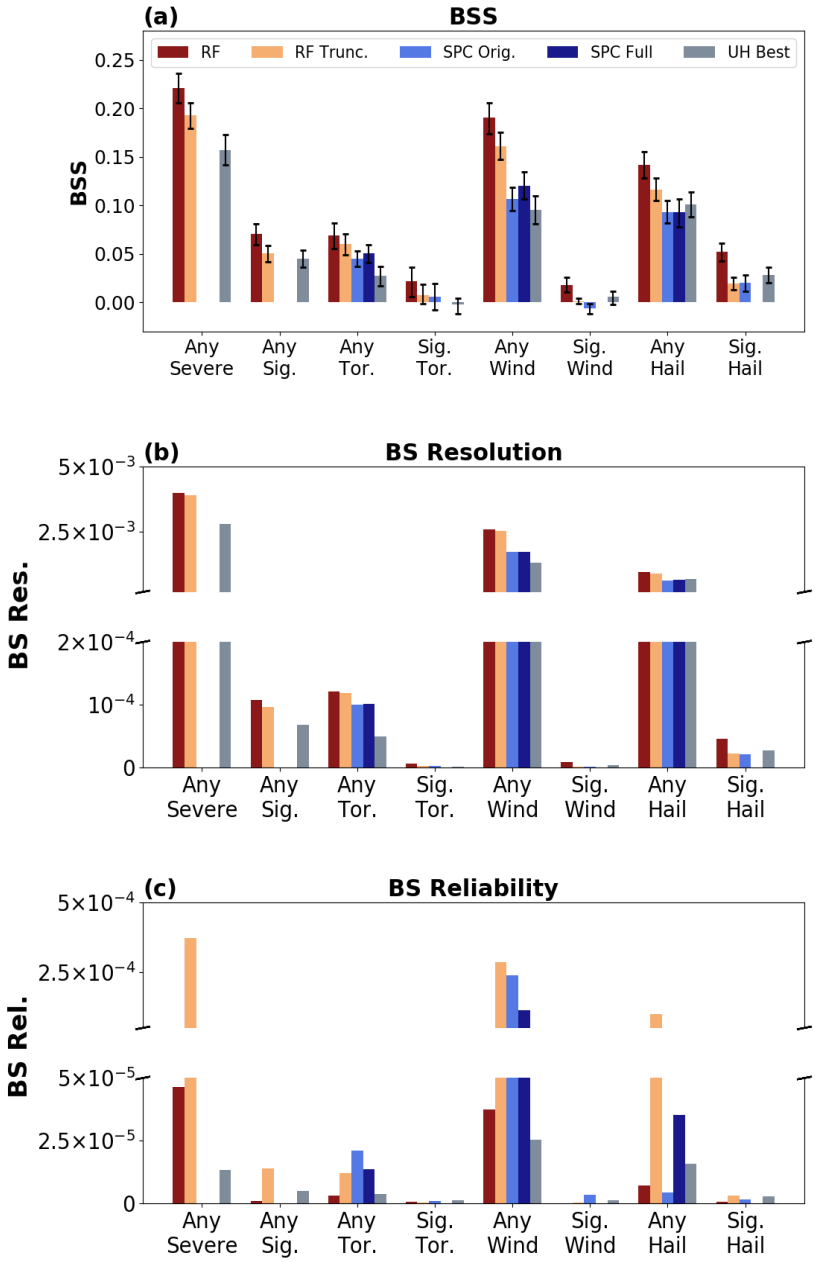

Figure 5 *(a) BSS for the continuous RF-based probabilities (dark red), truncated RF-based probabilities (yellow), original SPC probabilities (light blue), full/continuous SPC probabilities (dark blue), and calibrated UH-based probabilities (gray). (b)-(c) As in (a) but for the resolution and reliability components of the BS, respectively. Black bars denote 95% confidence intervals in (a). The abbreviations "sig." and "tor." refer to "significant" and "tornado," respectively.*

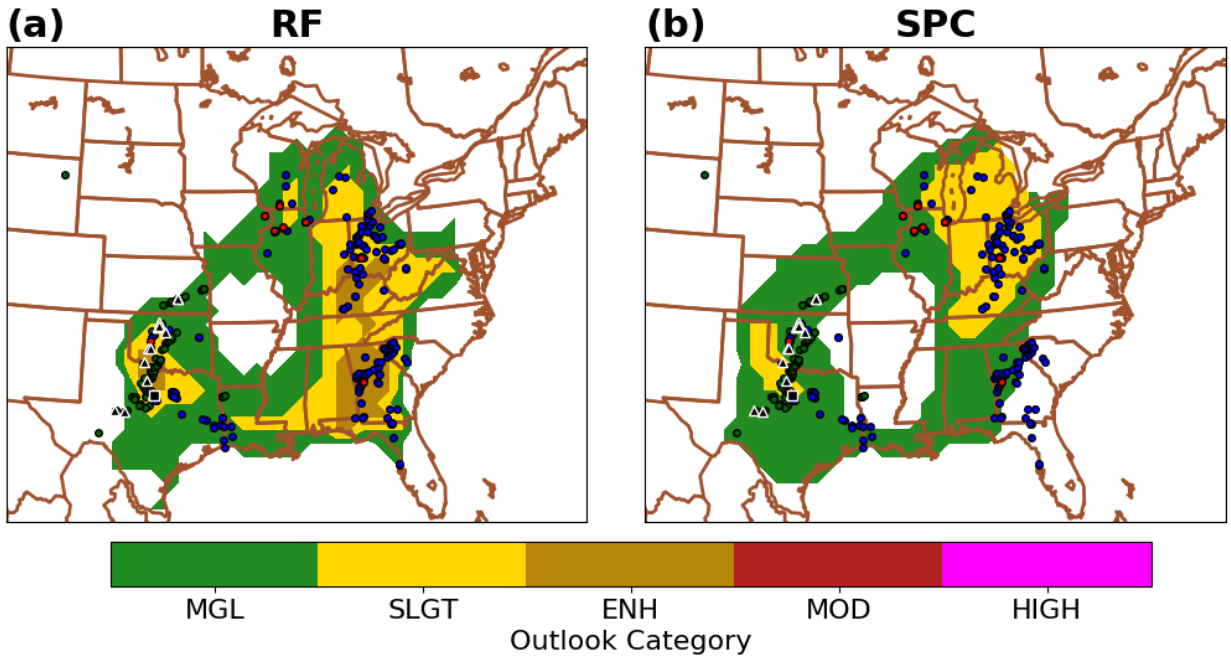

Figure 6 *Day 1 categorical outlook from the (a) RF approach and (b) SPC 0600 UTC forecast, valid for the 24-h period ending at 1200 UTC on 27 May 2015. Observed SPC storm reports are overlaid. Reports of tornadoes, severe wind, and severe hail are denoted by red, blue, and green circles, respectively, while significant tornado, wind, and hail reports are respectively represented by white-outlined red circles, black squares, and black triangles.*

RF individual-hazard probabilities also demonstrated considerable skill (e.g., Fig. 7a-f). For example, for the 24-h period ending on 27 May 2015, the RF correctly shifted the 2% tornado probabilities farther west in the Upper Midwest to better capture the tornado reports there compared to the SPC (Fig. 7a-b). Relative to the SPC forecast, the RF also had higher severe wind probabilities in the Ohio Valley and Southeast, where numerous severe wind reports occurred (Fig. 7c-d), and better highlighted the threat for significant severe wind and hail in the Southern Plains (Fig. 7c-f). However, the RF also had a relatively large area of false alarm in the eastern U.S. for severe hail (Fig. 7e).

RF forecasts also improved on UH-based forecasts for the prediction of all-hazards severe and significant severe weather (e.g., Fig. 8a-d). For example, the RF has higher severe probabilities in the East and Southeast U.S., where numerous severe wind reports occurred, as well as higher probabilities near the tornado reports in northwestern Illinois (Fig. 8a-b). The RF significant severe probabilities are much more similar to the UH-based probabilities (Fig. 8c-d), although the RF indicates non-zero significant severe probabilities in the Upper Midwest and slightly changes the orientation of the higher probabilities in the Southern Plains to more closely align with the observed significant severe reports.

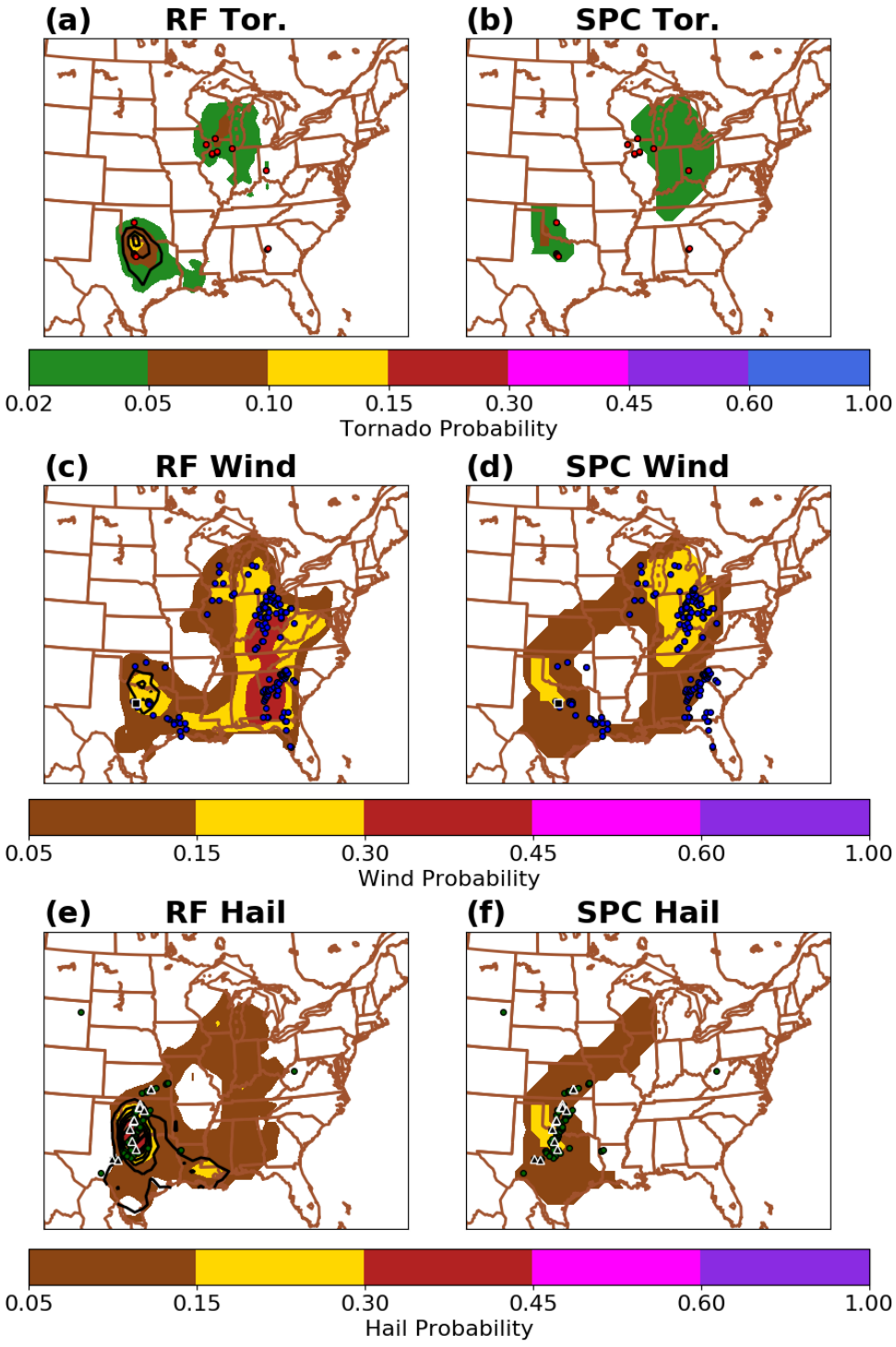

Figure 7 *(a) RF-based tornado probabilities (shaded) and significant tornado probabilities (contoured every 2% with ≥ 10% probabilities hatched), valid for the 24-h period ending at 1200 UTC on 27 May 2015. (b) As in (a) but for SPC forecasts issued at 0600 UTC. (c)-(d) As in (a)-(b) but for severe wind forecasts. (e)-(f) As in (a)-(b) but for severe hail forecasts. Observed SPC storm reports are overlaid, using the same representations as in Fig. 6.*

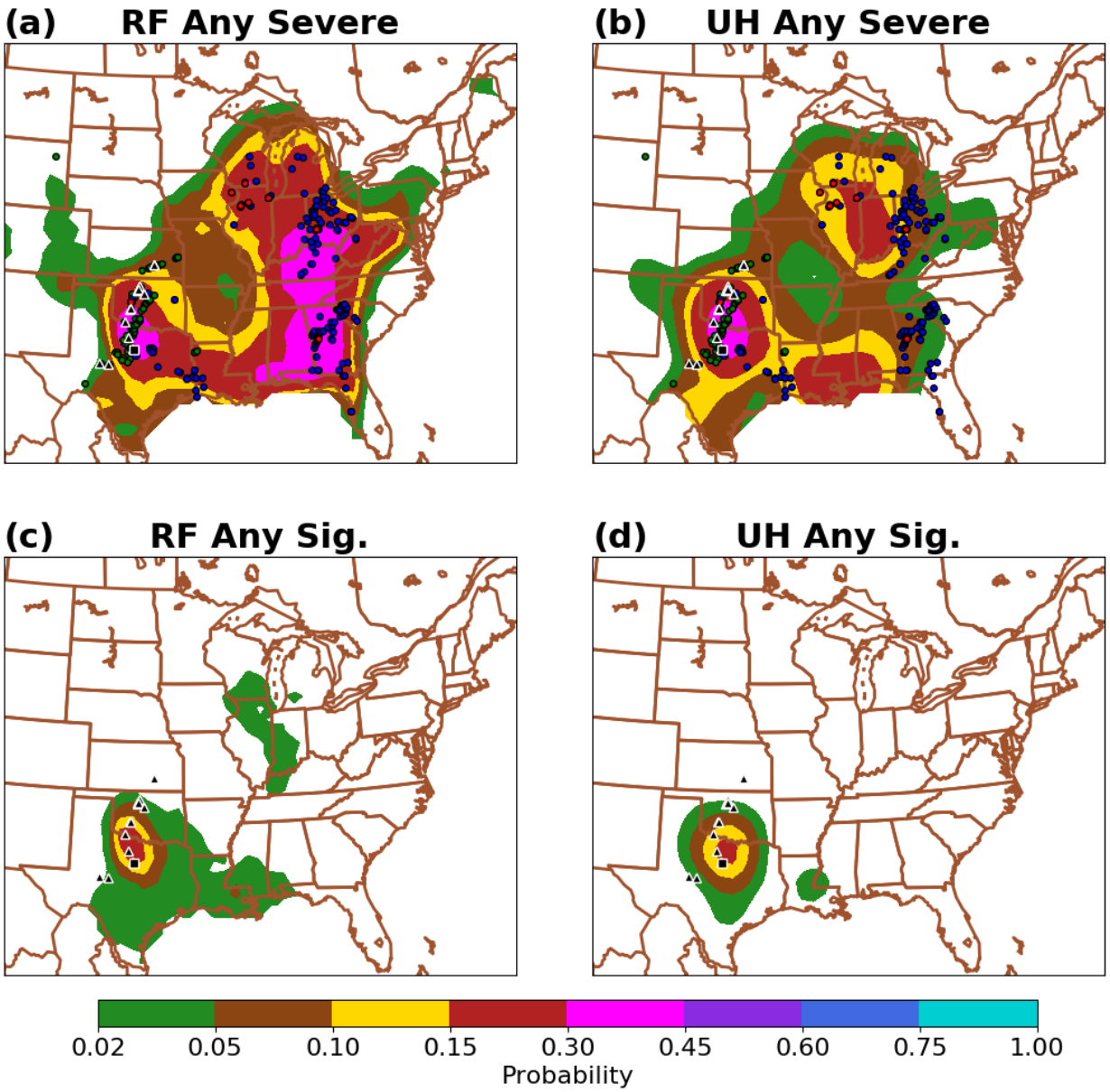

Figure 8 *(a) RF- and (b) UH-based probabilities of all-hazards severe weather, valid for the 24-h period ending at 1200 UTC on 27 May 2015. (c)-(d) As in (a)-(b) but all-hazards significant severe weather probabilities are plotted. Relevant observed SPC storm reports are overlaid, using the same representations as in Fig. 6.*

ii) Post-processing WoFS data

WoFS-derived RF severe weather forecasts were produced for hourly initializations from 19-03 UTC for 24 cases from May of 2018. Predictors included point-based WoFS environment and hourly maximum storm forecast variables (Table 3). SPC local storm reports (all-hazards) were used as the observational dataset.

Environment (Ensemble Mean)	Hourly Max. Storm Fields (Max., 90th Percentile, Smoothed Mean)	Miscellaneous
CAPE (0-3 km AGL)	Updraft Helicity (2-5 km AGL)	Initialization Time
Temperature (2-m)	Updraft Helicity (0-2 km AGL)	
Dewpoint (2-m)	Vertical Vorticity (0-2 km AGL)	
Specific Humidity (2-m)	Max. Updraft Speed	
Lifted Condensation Level (mixed layer)	Hail Size (HAILCAST)	
Virtual Potential Temperature (2-m)	Wind Speed (80-m)	
CAPE (mixed layer)		
CIN (mixed layer)		
SRH (0-1 km AGL)		
SRH (0-3 km AGL)		
Sig. Tornado Parameter (mixed layer)		

Table 3 *WoFS-based predictors for obtaining severe weather probabilities.*

RF-based severe weather forecasts were compared against baseline severe weather probabilities, derived from WoFS UH. Baseline probabilities were created by spatially smoothing the fraction of ensemble members with UH exceeding 120 m^2s^{-2} with an isotropic, 2-dimensional Gaussian kernel density function (standard deviation of 48 km). The UH threshold and density function standard deviation were chosen to optimize the BS.

6-fold cross validation was used to create/verify the RF forecasts. Preliminary results suggest that the RF forecasts had better AUCs, BSs, and BS reliability values compared to the baseline forecasts (not shown). RF probabilities also had finer-scale details and tighter gradients relative to the baseline; however, RF probabilities were also occasionally high in areas without reports (not shown).

To assess variable importance, new RFs were trained with subsets of predictors. The following three experiments were run:

1. Storm fields vs. Environmental fields (2 RFs)
2. Individual storm fields (6 RFs)
3. Type of storm field (i.e., smoothed ensemble mean, 90th percentile, ensemble max; 3 RFs)

Overall, it was found that storm fields added more skill than environmental fields, with UH and maximum updraft speed as the best storm field predictors (Table 4). However, at times, the environment fields provided large positive contributions to RF skill (not shown). The smoothed ensemble mean was found to be the best *type* of storm field predictor (Table 4).

	RF Predictors	AUC	BS	BS_{rely} x 10^3
Control	Full Set of Predictors	0.8897	0.0451	0.139
Experiment 1: Storm vs. Environmental Fields	Hourly Max. Storm Fields	0.8813	0.0454	0.176
	Environmental Fields	0.7434	0.0578	2.061
Experiment 2: Individual Storm Fields	2-5 km UH	0.8803	0.0462	0.156
	Max. Updraft Speed	0.8843	0.0463	0.253
	0-2km UH	0.8735	0.0474	0.207
	0-2 km Vertical Vorticity	0.8662	0.0475	0.118
	Hail Size (HAILCAST)	0.8529	0.0475	0.437
	80 m Wind Speed	0.8228	0.0527	1.562
Expeirment 3: Type of Storm Field	Smoothed Mean	0.8864	0.0455	0.234
	90^{th} Percentile	0.8731	0.0472	0.081
	Max	0.8657	0.0482	0.084

Table 4 *Results from the WoFS variable importance experiments. BS_{rely} denotes the reliability component of the Brier Score.*

4. Future work

Future work will focus on interpreting RF output probabilities to identify when and why they differ from UH- and human-based probabilities. Interpretability information may not only aid human forecasters in making better predictions but could also alert model developers of ensemble biases.

Future work may also apply similar RF techniques for post-processing forecasts on finer spatiotemporal scales (e.g., sub-daily time and/or sub-regional space scales). Finally, efforts are underway to apply RF-based post-processing to the operational HREFv2 and test real-time RF severe weather probabilities in upcoming HWTSFEs.

5. References

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, doi:https://doi.org/10.1023/A:1010933404324.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed experimental forecast spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/doi:10.1175/BAMS-D-11-00040.1.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Regional ensemble forecast systems at NCEP. Preprints, *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.5. [Available online at https://ams.confex.com/ams/27WAF23NWP/webprogram/Manuscript/Paper273421/NWP2015_NCEP_RegionalEnsembles_paper.pdf.]

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.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Good, P. I., 2006: *Resampling Methods*. Birkhauser Boston, 228 pp.

Herman, G. R., and R. S. Schumacher, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Jirak, I. L., A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *25th Conference on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

Jirak, I. L., C. J. Melick, and S. J. Weiss, 2016: Comparison of the SPC storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102, https://ams.confex.com/ams/28SLS/webprogram/Session41668.html.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC Storm-Scale Ensemble of Opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. 26th Conference on Severe Local Storms, Nashville, TN, Amer. Meteor. Soc., P9.137, https://www.spc.noaa.gov/publications/jirak/sseo_hwt.pdf.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system. Part II: Combined radar and satellite data experiments. *Wea. Forecasting*, **31**, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, doi:https://doi.org/10.1175/WAF2007106.1.

Karstens, C. D., R. Clark III, I. L. Jirak, P. T. Marsh, R. Schneider, and S. J. Weiss, 2019: Enhancements to Storm Prediction Center Convective Outlooks. *9th Conference on*

Transition of Research to Operations, Phoenix, AZ, Amer. Meteor. Soc., J7.3, https://ams.confex.com/ams/2019Annual/webprogram/Paper355037.html.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Post-processing next-day ensemble probabilistic precipitation forecasts using random forests. *Weather and Forecasting*, **34**, 2017-2044.

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., Z. Liu, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, doi:https://doi.org/10.1175/WAF-D-13-00145.1.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype warn-on-forecast system. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

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, doi:https://doi.org/10.1175/WAF-D-15-0138.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, and M. L. Weisman, 2019: Next-day prediction of tornadoes using convection-allowing models with 1-km horizontal grid spacing. *Wea. Forecasting*, **34**, 1117–1135, https://doi.org/10.1175/WAF-D-19-0044.1.

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, doi: https://doi.org/10.1175/WAF-D-10-05046.1.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences: Third Edition*. Elsevier Inc., 676 pp.