DETERMINING WHEN AND HOW A RANDOM FOREST ADDS VALUE TO DAY 1 SPC HAIL FORECASTS

Shannon O. McCloskey[1,5], Eric D. Loken[2], David E. Jahn[2,3], Christopher D. Karstens[3,4], Bryan T. Smith[3]

[1]National Weather Center Research Experience for Undergraduates Program
Norman, Oklahoma
[2]CIMMS/University of Oklahoma
Norman, Oklahoma
[3] NOAA/NWS/Storm Prediction Center
Norman, Oklahoma
[4]NOAA/OAR/National Severe Storms Laboratory
Norman, Oklahoma
[5]University of Florida
Gainesville, Florida

ABSTRACT

This study investigates when an RF algorithm's day 1 severe hail probabilities differ from corresponding Storm Prediction Center (SPC) human-generated probabilities by at least one SPC outlook category. The goal of this study is to determine when an RF is most and least likely to add value to day 1 SPC human hail forecasts. RF forecasts are trained on forecast variables from the High-Resolution Ensemble Forecast System, version 2.1 (HREFv2.1) and observed SPC hail reports, using 627 days of data from May 2018 through April 2020. RF forecasts are compared against a continuous version of human-generated day 1 SPC hail forecasts, produced daily at 06z.

Analysis shows that the RF is especially skillful in reducing false alarm by forecasting one outlook category lower than that of the SPC. Additionally, when the RF forecasts at least one outlook category higher (lower) than the SPC, ensemble mean storm attribute variables including maximum 2-5 km updraft helicity, maximum upward vertical velocity, and maximum downward vertical velocity tend to have higher (lower) absolute values. Meanwhile, the distribution of these variables does not change much when the SPC forecasts at least one outlook category higher or lower than the RF. These findings suggest that RFs add value to the SPC by calibrating their probabilities based on the strength of simulated storms, while SPC forecasters add value to the RF by analyzing other (meteorological and non-meteorological) variables.

1. INTRODUCTION[1]

Random forests (RFs) produce skillful, reliable probabilistic guidance for next-day precipitation (e.g., Gagne et al. 2014; Herman and Schumacher 2018; Loken et al. 2019) and severe weather (e.g., Loken et al. 2020; Hill et al. 2020), and they are starting to be incorporated in real-time forecasting operations (e.g., Schumacher et al., in press). Indeed, Loken et al. (2020) showed that over many cases, day 1 RF severe hail probabilities are now at least as good as corresponding Storm Prediction Center (SPC) human forecasts. However, machine learning (ML) methods have difficulty extrapolating beyond the relationships contained in their training dataset, which occasionally leads to poor performance. This can pose a problem for forecasters, since useful operational prediction tools must be trustworthy and reliable (Karstens et al. 2015). By

[1] *Corresponding author address:* Shannon O. McCloskey, University of Florida, Ste. 2300, 120 David L. Boren Blvd., Norman, OK 73072. E-mail: shannonmccloskey11@gmail.com

studying the error characteristics of ML methods in severe weather, forecasters and researchers may be able to identify when ML forecasts are trustworthy and when they may struggle (McGovern et al. 2017).

One way to determine when RF forecasts are trustworthy is to compare them to analogous human forecasts. Loken et al. (2020) found that RF and corresponding SPC severe weather forecast probabilities often covered similar areas but had different magnitudes. However, a systematic analysis of when, where, and how often RF forecasts substantially differ from SPC human forecasts has never been done. This study aims to fill that knowledge gap by exploring differences between RF and SPC day 1 hail forecasts, with the quality of RF deviations from SPC guidance assessed using observed SPC hail reports. Through this analysis, this study hopes to determine when an RF algorithm is most and least likely to add value to day 1 SPC human hail forecasts.

2. DATA & METHODS

Variable	Description (Units)
2-5 UH	Maximum daily simulated 2-5 km updraft helicity (m^2/s^2)
MAXUVV	Maximum daily simulated upward vertical velocity m/s)
MAXDVV	Maximum daily simulated downward vertical velocity (m/s)

Table 1: Abbreviations and explanations of the variables analyzed in this study.

Using the method discussed in Loken et al. (2020) and Loken (2021), a severe-hail-predicting RF is trained and verified on 627 days of HREFv2.1 (Roberts et al. 2020; Loken 2021) and observed SPC hail report data. While the RF considers a variety of forecast field inputs from the HREFv2.1 (Table 4.3 in Loken 2021), this study primarily analyzes the variables described in Table 1.

To align with SPC's objective to forecast severe weather within 40-km of a point, we predict the probability of at least one observed hail report falling within an 80 km grid box each day. Specifically, observed SPC hail reports are remapped to an approximately 80 km grid and are recorded in a binary manner, where a value of one (zero) indicates one or more (zero) observed daily SPC hail reports in the given box. RF predictions are output on an approximately 80-km grind that covers the contiguous United States (CONUS; Fig. 1).

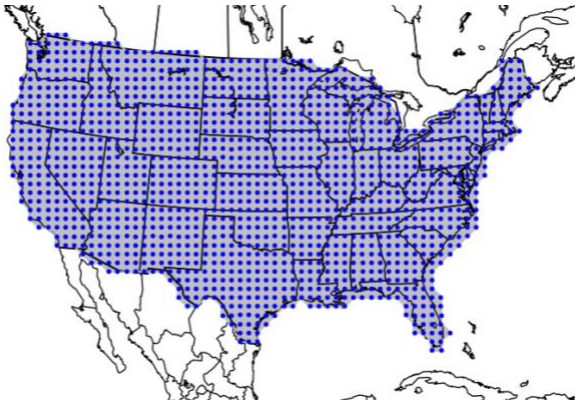


Figure 1: RF analysis domain (gray shading) and forecast points (blue dots).

Murphy (1993) defines three measurements of forecast goodness: consistency, quality, and value. Here, we are interested in assessing value, which can be difficult to quantify. For this paper, we assume that the RF potentially adds value to SPC forecasts when the RF's forecast probabilities differ from those of the SPC by more than one outlook category (Table 2).

An important consideration is that it is difficult to compare discrete SPC probability forecasts with continuous RF forecast probabilities. Therefore, we analyze experimental continuous SPC probabilities, created in the same way as described by Loken et al. (2020). The impact of significant severe weather on a forecast probability level is also a limitation, as it can impact the SPC category level of a given data point even though significance is not considered in this study.

Day 1 SPC Outlook Probability	Hail
5-15%	MRGL
15-30%	SLGT
30-45%	ENH
>45%	MDT

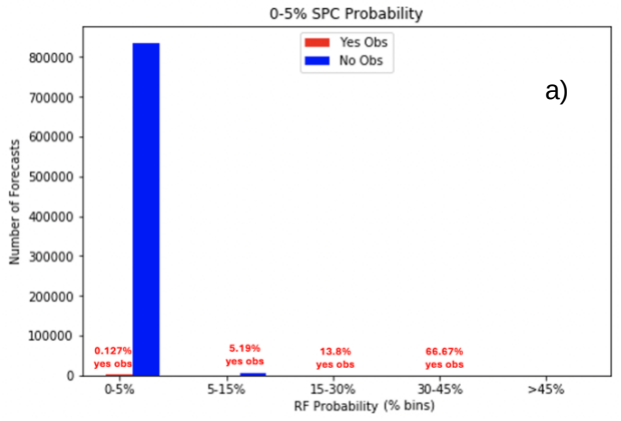
Table 2: Day 1 SPC convective outlook probabilities by category for hail. MRGL = marginal, SLGT = slight, ENH = enhanced, MDT = moderate, and HIGH = high.

It is also important to note that SPC forecasters must consider more than just meteorological conditions when issuing their probabilities. They must be mindful of what risk category their probabilities may trigger and how their forecasts will be interpreted by the public. So, they must consider impact and meteorological probability when making their forecasts, while the RF only considers quantitative likelihood and meteorological significance.

Figures 2a-2e are bar charts binned by SPC probability of hail, based on the day 1 SPC outlook probabilities in Table 2. The data within this subset is then conditioned based on the RF probabilities of hail, and then grouped by this percentage on the x-axes. The y-axes of all the figures shows the number of forecasts in each bin. Finally, the data was further adapted to split the forecasts in each RF bin by “yes obs” and “no obs”, The percentage above each red bar indicates the percentage of yes observations out of the total number of forecasts in each bin. For example, the third bar from the left in Figure 2c contains all of the points in the original dataset that had a 15-30% (SLGT) SPC hail probability, and a 15-30% RF hail probability. 22.7% of the forecasts in this bin had hail reports associated with them as well. Similarly, the first bar on the left indicates forecasts with an SPC probability of 15-30% and an RF probability of 0-5%, and only 4%of the forecasts in this bin had associated hail observations reported. Each of these figures follows the same idea.

Violin plots are created for variables deemed most important to the RF by the Python tree interpreter model (Saabas 2016; Loken 2021) and show the distribution of these variables given a particular SPC and RF forecast outlook category. Figures 3a and 3b show two sample violin plots of simulated daily maximum 2-5 km updraft helicity (2-5 UH) compared with forecast probabilities. Figure 3b follows the same basic method for conditioning the data set that was used for the bar plots shown in Figure 2, however the y-axes differ in the violin plots. Instead of having the number of forecasts in each bin plotted on the y-axis, the violin plots have the range of 2-5 UH values in each bin. Figure 3a follows the same logic, however the RF and SPC conditioning is flipped. So, the second violin from the left in figure 3a includes all the forecast points that had a 5-15% RF probability and a 5-15% SPC probability of hail and shows the range of 2-5 UH values associated with those forecasts. Figures 4 and 5 show violin plots that follow the same conditioning as Figure 3, but for two other variables of interest. These variables are shown below in Table 2. Extremities differ among these three variables. Higher values of 2-5 UH and MAXUVV and lower (more negative) values of MAXDVV indicate stronger, more intense storms.

3. RESULTS



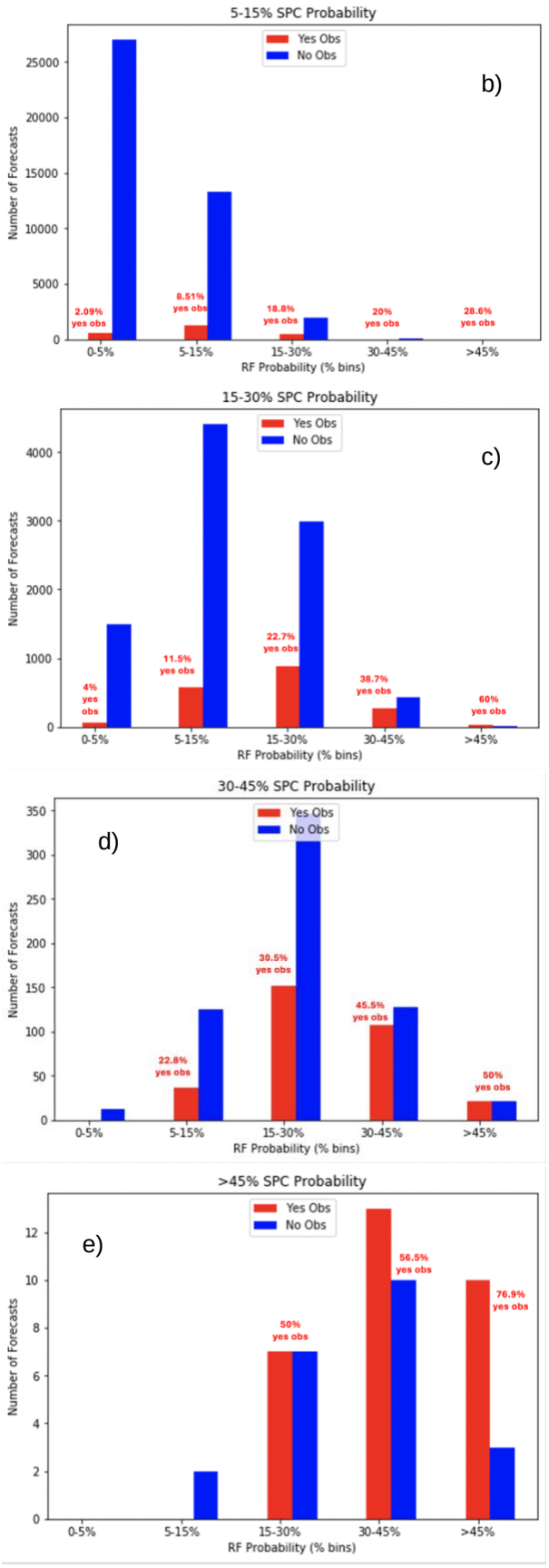
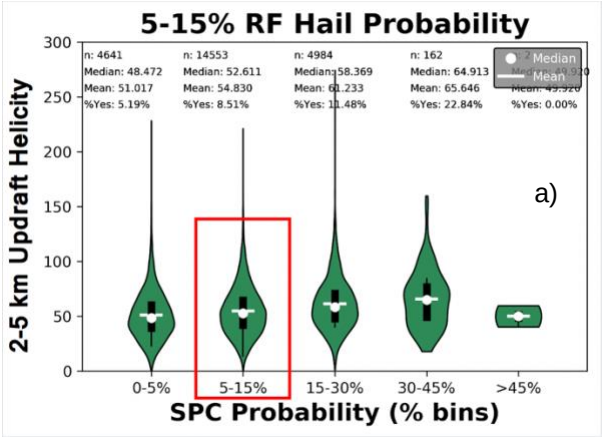


Figure 2: Bar charts showing different SPC probability bins further conditioned by the RF. **Fig. 2a** shows less than 5% SPC probability, **fig. 2b** shows 5-15% SPC probability, **fig. 2c** shows 15-30% SPC probability, **fig. 2d** shows 30-45% SPC probability, and **fig. 2e** shows greater than 45% SPC probability. Each number above the bars represents the percentage of yes observations associated with the number of forecast points in that particular bin.

The bar charts shown in Figures 2a-2e show that the RF tends to forecast the majority of its points plus or minus one category from the SPC, which implies that they are generally forecasting similar probabilities. In Figures 2b-2e, the RF consistently has the highest number of forecasts in the bin that is one category lower than what the SPC predicts. This trend in the data tells us that the RF may be skilled at reducing false alarm in human forecasts. While the RF continues to follow this trend in Figure 2e, it is incorrect in doing so. This is because when the SPC forecasts an MDT hail probability, there are high percentages of hail observations associated with the forecast points in the SLGT and ENH RF bins. However, this observation may be occurring because of the relatively low sample size the bins in this figure have, thus conclusions cannot be drawn. The RF may also be skilled at enhancing POD, as represented by the percentage of hail observations in the MDT% RF probability bins of each figure, but especially in the middle categories (Figures 2b, 2c, and 2d).



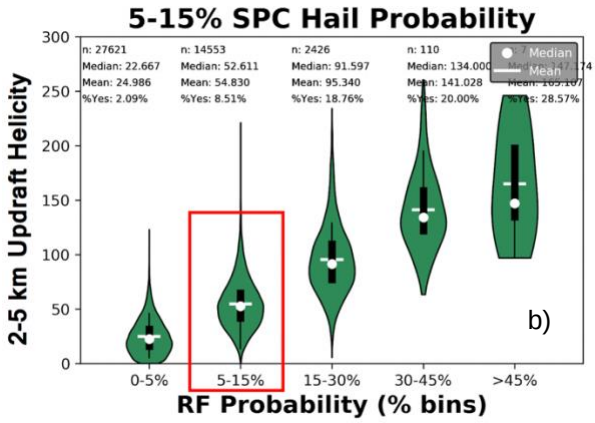


Figure 3: Violin plots for 2-5 UH values conditioned by a 5-15% RF (top) and 5-15% SPC (bottom) hail probability.

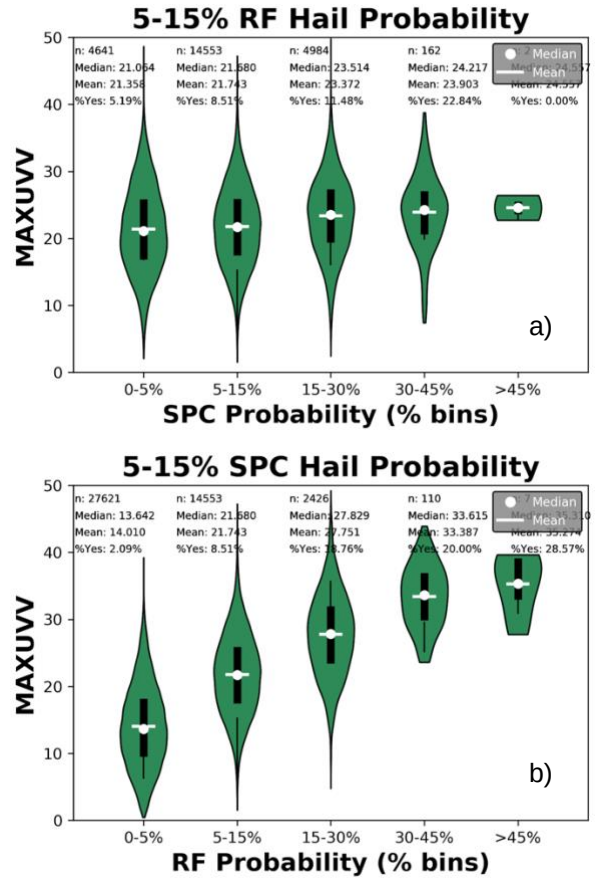


Figure 4: Violin plots for MAXUVV values conditioned by a 5-15% RF (top) and 5-15% SPC (bottom) hail probability.

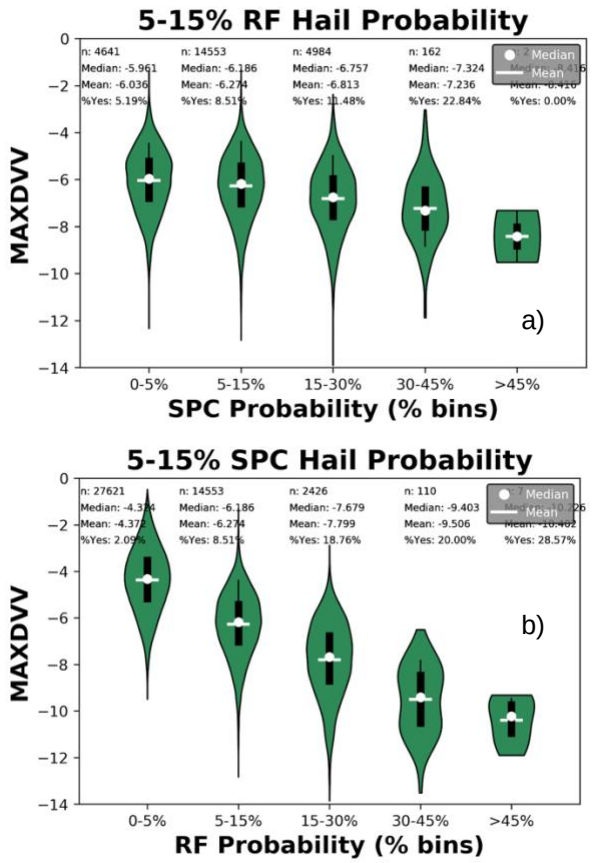


Figure 5: Violin plots for MAXDVV values conditioned by a 5-15% RF (top) and 5-15% SPC (bottom) hail probability.

Figures 3, 4, and 5 all show sample violin plots for 5-15% RF hail probabilities conditioned by the SPC probabilities (Figs. 3a, 4a, and 5a), as well as the 5-15% SPC hail probabilities conditioned by the RF forecasts (Figs. 3b, 4b, and 5b). In Figure 3, consider the violins outlined in red to be the “benchmarks”, as they represent the range of 2-5 UH values measured when (in 3a) the RF forecasted a 5-15% probability of hail and the SPC forecast the same, and (in 3b) where SPC gave a 5-15% probability and the RF had the same. Figure 3b shows that when the RF deviates from the SPC by a category, say moving to a 15-30% hail probability, there is a clear increase in the 2-5 UH values associated with those forecast points. This consistently increasing trend in RF probabilities is noticeable not only with 2-5 UH but can be seen when looking at MAXUVV in Figure 4b as well. MAXDVV follows the same trend as well, but the trend appears negative because more

negative MAXDVV values indicate stronger simulated storms. This trend among these ensemble mean storm attribute variables suggests that the RF strongly emphasizes higher (and lower) absolute values of variables such as 2-5 UH, MAXUVV, and MAXDVV because they are indicative of stronger (weaker) simulated storms.

Conversely, Figure 3a shows that when the SPC forecasts at least one outlook category higher or lower than the RF, the distribution of simulated storm attribute variables does not change much. This is consistent in Figures 4a and 5a, because as the SPC deviates from the RF, MAXUVV and MAXDVV values fluctuate slightly but do not significantly change. This supports the notion that the RFs add value to the SPC by calibrating their probabilities based on the strength of simulated storms, while SPC forecasters add value to the RF by analyzing both meteorological and non-meteorological factors. Both of these findings are consistent throughout all SPC and RF bins of other outlook categories as well.

4. CASE STUDY

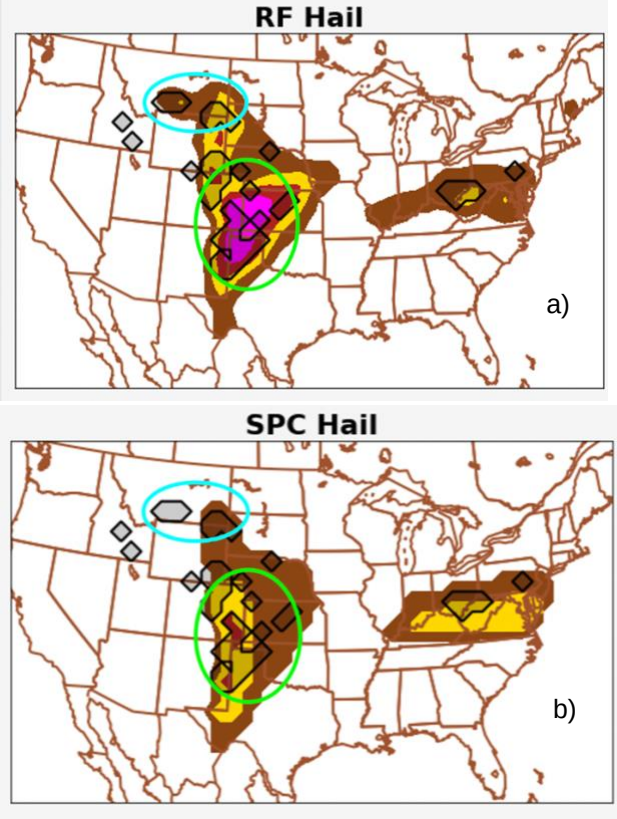
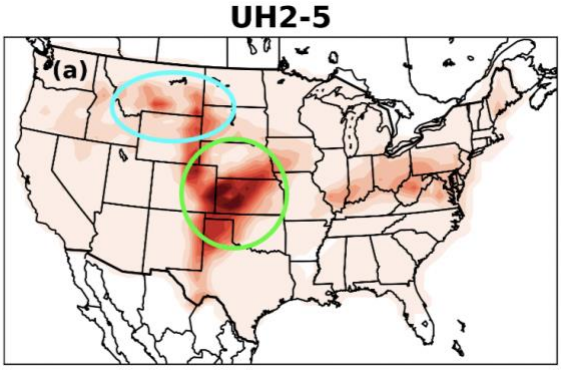


Figure 6: Hail probability forecasts (colored shading) on 26 May 2019 by the RF (top) and SPC (bottom). Locations with at least one observed SPC hail report are contoured and shaded black. Two areas of interest have been circled and are discussed in the text.

RF and SPC forecast probabilities of hail from 26 May 2019 are shown in Figure 6. This case study provides a clear example of how the RF places such a strong emphasis on high values of storm attribute variables such as 2-5 UH and MAXUVV. It also shows another example of how the RF can add value to the SPC forecasts. Two areas of interest have been circled on the figures to the left. The blue circle covers an area in southeastern Montana, and the green circle covers a much broader range of the Central Plains, including much of western Kansas and the area where Oklahoma, Texas, Kansas, Colorado, Nebraska, and New Mexico meet up.

The green circle, covering a broad region of the Central Plains, the RF forecasted significantly higher probabilities of hail than the SPC, even deviating by two categories in certain areas. Looking at the pink area on these figures and the same area highlighted in Figures 7 and 8, the RF clearly placed a large importance on the very high 2-5 UH and MAXUVV values. Looking at the blue circle, the RF forecasted one category higher than the SPC did and extended its forecast to cover a greater area than the SPC. Figures 7 and 8 shows that in this same area, there are patches of relatively higher 2-5 UH and MAXUVV values.



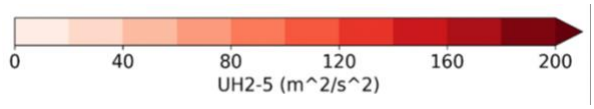


Figure 7: Map plotting the 2-5 UH values from May 26th, 2019. Darker areas represent higher 2-5 UH values. The areas of interest are identified by the blue and green circles.

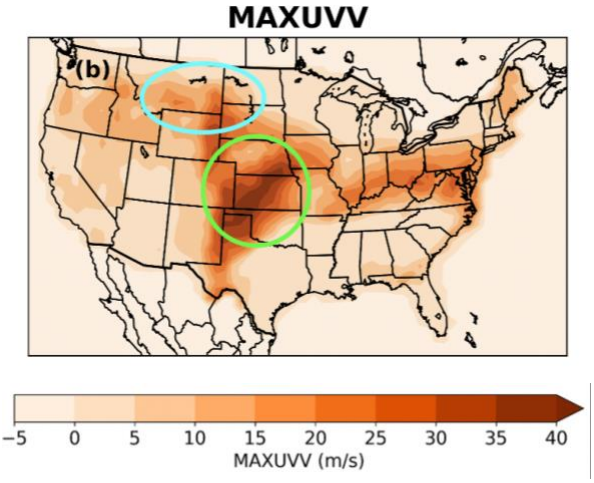


Figure 8: Map plotting the MAXUVV values from May 26th, 2019. Darker areas represent higher MAXUVV values. The areas of interest are once again identified by the blue and green circles.

5. DISCUSSION/ SUMMARY/ CONCLUSIONS

This study found that in most cases, the RF and the SPC are more alike than different when it comes to forecasting hail probabilities. The RF generally differed from the SPC by no more than one outlook category. Most frequently, the RF probabilities fell one outlook category below that forecasted by the SPC. Over many cases, the smaller RF probabilities were associated with lower observed report frequencies, suggesting the RF often successfully reduced false alarm compared to the SPC. When the RF forecast at least one outlook category higher (lower) than the SPC, storm attribute variables such as those discussed in this paper tend to have higher (lower) absolute values. This indicates stronger or weaker simulated storms. Conversely, the distribution of the simulated storm attribute variables did not change much when the SPC forecasts at least one outlook category higher or lower than the RF. These findings suggest that RFs add value to the SPC by calibrating their probabilities based on the strength of simulated storms, while SPC forecasters add value to the RF by analyzing other meteorological and non-meteorological variables.

Ultimately, it is hoped that the results of this study will help severe weather forecasters better utilize RF guidance in operations.

6. ACKNOWLEDGEMENTS

This work was prepared by the authors with funding provided by National Science Foundation Grant No. AGS-2050267, and NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation, NOAA, or the U.S. Department of Commerce.

7. REFERENCES

Gagne, D. J., McGovern, A., Xue, M., 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.

Herman, G. R., Schumacher, R. S., 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.

Hill, A. J., Herman, G. R., Schumacher, R. S., 2020: Forecasting Severe Weather with Random Forests, *Mon. Wea. Rev.*, **148**, 2135-2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Karstens, C. D., and Coauthors, 2015: Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed, *Wea. Forecasting*, **30**, 1551-1570, https://doi.org/10.1175/WAF-D-14-00163.1.

Karstens, C. D., and Coauthors, 2018: Development of a Human-machine Mix for Forecasting Severe Convective Events, *Wea. Forecasting*, **33**, 715-737, https://doi.org/10.1175/WAF-D-17-0188.1.

Loken, E. D. (2021). *The Creation and Analysis of Next-Day Random Forest-Based High-Impact Weather Forecasts* (Publication No. [Doctoral Dissertation, University of Oklahoma]. SHAREOK.

Loken, E. D., Clark, A. J., Karstens, C. D., 2020: Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests, *Wea. Forecasting*, **35**, 1605-1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Loken, E. D., Clark, A. J., McGovern, A., Flora, M., Knopfmeier, K., 2019: Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests, *Wea. Forecasting*, **34**, 2017-2044, https://doi.org/10.1175/WAF-D-19-0109.1.

McGovern, A., Elmore, K. L., Gagne II, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., Williams, J. K., 2017: Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather, *Bull. Amer. Meteor. Soc.*, **98**, 2073-2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

McGovern, A., Lagerquist, R., Gagne II, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., Smith, T., 2019: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, *Bull. Amer. Meteor. Soc.,* **100**, 2175-2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Murphy, A. H., 1993: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, *Wea. Forecasting*, **8**, 281-293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Pliske, R., Crandall, B., Klein, G., 2004: Competence in Weather Forecasting, *Psychological Investigations of Competence in Decision Making*, 40-70.

Roberts, B., Gallo, B. T., Jirak, I. L., Dowell, D. C., Wang, X., Wang, Y, 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.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality, *Wea. Forecasting*, **24**, 601-608, https://doi.org/10.1175/2008WAF2222159.1

NOAA/SPC, 2020: SPC Products. Accessed 30 June 2021, https://www.spc.noaa.gov/misc/about.html.

Saabas, A. “Random Forest Interpretation with Scikit-learn.”*Diving into Data*, 12 August 2015, https://blog.datadive.net/random-forest-interpretation-with-scikit-learn/.

Schumacher, R. S., Hill, A. J., Klein, M., Nelson, A., Erickson, M. J., Trojniak, S. M., Herman, G. R., 2021, in press: From Random Forests to Flood Forecasts: A Research to Operations Success Story, *Bull. Amer. Meteor. Soc.,* https://doi.org/10.1175/BAMS-D-20-0186.1.