Verification of Hail Forecasts Produced by Machine-Learning Algorithms

SARAH MCCORKLE*

Indiana University, Bloomington, IN

NATHAN SNOOK

Center for Analysis and Prediction of Storms, Norman, Oklahoma

AMY MCGOVERN

School of Computer Science, University of Oklahoma, Norman, Oklahoma

ABSTRACT

Hail can result in billions of dollars with of damage every year. The ability to forecast for significant hail events even just a day in advance can greatly mitigate severe hail risk. Machine-learning (ML) algorithms have already shown skill in producing skillful hail forecasts, as they can identify the areas that hail will be a threat. Using output from the High Resolution Ensemble Forecast version 2 (HREFv2) model, new forecasts were produced during the Hazardous Weather Testbed (HWT) Spring 2018 experiment for days April 30th to June 1st. Verification is necessary to identify weaknesses in these algorithms in order to make improvements. The ultimate goal of verification of these forecasts is to show that these ML algorithms can skillfully forecast for hail to increase trust to eventually implement them into operational forecasting. By verifying these forecasts using reliability diagrams, it was discovered that there was a bias of over-forecasting. Isotonic regression was used to correct for the HREFs tendency to over-forecasting. The raw HREFv2 data was calibrated to both the SPC practically perfect forecasts and to the observations. When calibrated to the observations, the corrected HREFv2 produced more reliable forecasts.

1. Introduction

Hail is extremely damaging to property, crops, and livestock, resulting in billions of dollars of damage every year (National Severe Storms Laboratory). With hail being as dangerous and costly as it is, the ability to better predict when and where hail will occur one or more days in advance becomes extremely valuable. Furthermore, compared to other thunderstorm hazards (such as heavy rain, strong winds, and tornadoes), the prediction of hail itself has been lacking (Snook et al. 2016). Models that have the capabilities to explicitly predict for hail will be beneficial in helping mitigate its impacts, especially in the case of severe hail.

Previous studies have shown that machine-learning algorithms (ML) have shown skill in producing more reliable hail forecasts by identifying areas where hail will be a threat (McGovern et al. 2017; Gagne et al. 2017). ML algorithms operate by searching for and identifying patterns within a large database. For the case of weather prediction, the algorithm searches through a database of past weather events. Specifically, the algorithm will match observations or NWP forecasts to events of meteorological significance that occurred in the past. By matching these separate pieces of data, the model is being trained to recognize patterns. Once new data is presented (i.e. new observations or NWP model forecasts), the model can be used to create accurate predictions of future events (in this case, forecasts of hail) based on this new data.

With the abundance of weather data available, artificial intelligence methods such as machine-learning are efficient and timely. These algorithms search and find relevant information that the model can use to produce hail forecasts, which in turn forecasters can utilize. However, forecasters need to be able to trust these ML models in order to implement them when forecasting, especially in the case of severe weather (McGovern et al. 2017).

The goal of this study is to demonstrate that ML algorithms can be used to skillfully forecast hail and produce real-time forecasts that are useful in a variety of operational applications. By verifying ML models run in real-time, specifically the HREFv2 model, it will become evident as to how to further improve such ML models in the future in support of the goal of implementing these methods into everyday operational forecasting procedures.

*Corresponding author address: Sarah McCorkle, Indiana University, 501 S. Woodlawn Ave, Bloomington, IN 47401
E-mail: smmccork@iu.edu

2. Data and Methods

The forecasts that will be verified in this study were produced during the 2018 Hazardous Weather Testbed (HWT) spring forecast experiment, an annual program which is conducted jointly by the Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL). The purpose of the HWT is to test emerging technologies and ideas that hope to improve prediction of hazardous weather (Experimental Forecast Program of the NOAA/Hazardous Weather Testbed 2018). In 2018, the HWT spring forecast experiment ran from April 30 to June 1; forecasts from these days will be verified in this study. Included in the spring experiment were ML hail forecasts produced from the HREFv2 model, which is an operational model used by NOAA.

The ML forecasts produced from HREFv2 model data predict severe hail over a domain covering the continental United States. The forecasts are produced using a horizontal 3-km grid spacing, on a grid which contains 1059 x 1779 points. The ML forecasts predict hail probabilities for both 25 mm (severe) and 50 mm (significant severe) hail, and contains forecasts for each of the eight HREFv2 ensemble members. The ML hail forecasts from HREFv2 predict the spatial distribution of the probability of hail. Fig. 1 shows an example ML hail forecast for May 29th, 2018.

In order to evaluate the performance of ML forecasts of severe hail using HREFv2 model data, these forecasts must be verified against what was observed. In this study, SPC hail reports will provide the observed data. The verification process will consider the surrounding 25 miles around each hail report, in order to produce forecasts which are directly comparable to SPC operational outlooks, which predict the probability of hazards (including hail) occurring within 25 miles of a point (Fig. 2).

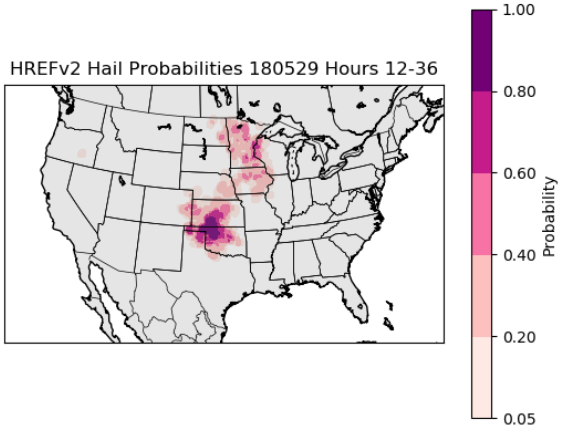

FIG. 1. HREFv2 forecast showing the probability of hail greater than 25 mm for May 29th, 2018. On this day, hail probabilities were particularly high on the Kansas/Oklahoma border.

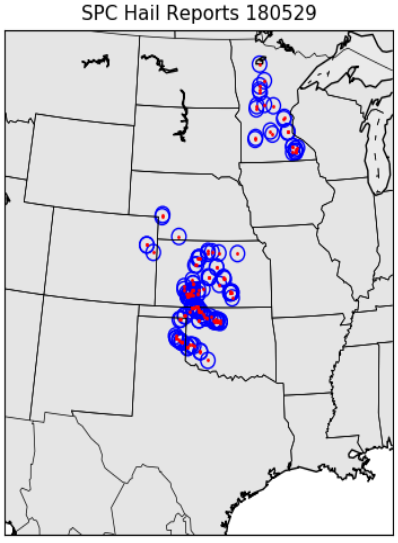

FIG. 2. Hail reports with 25-mile radius circles around each report plotted for May 29, 2018.

Every grid point from the HREFv2 model that is within 25 miles from the SPC hail report will be accounted for, as this creates the dataset that will be used to verify the model. If a model grid point falls within 25-miles of a storm report, that grid point is considered to have had (observed) hail occur there, while grid points more than 25 miles from any hail report are considered not to have had (observed) hail occur (Fig. 3).

3. Evaluation

The purpose of this paper is to determine if the ML hail forecasts produced using HREFv2 data produced quality hail forecasts, in terms of how well forecasts correspond to observations, or vice versa (Murphy 1993). Forecasts will be evaluated using reliability diagrams, which will provide insight on the skill of the HREFv2 forecasts by revealing biases. Reliability diagrams plot the observed frequency of hail reports against the forecast probability given by the HREFv2 model. The observed frequency refers to the fraction of points with observed hail for a given forecast probability bin. A forecast is perfectly reliable if the forecast probability and observed frequency are the same. Any deviation from the perfectly reliable line provides the conditional bias of the forecast (World Weather Research Programme 2017).

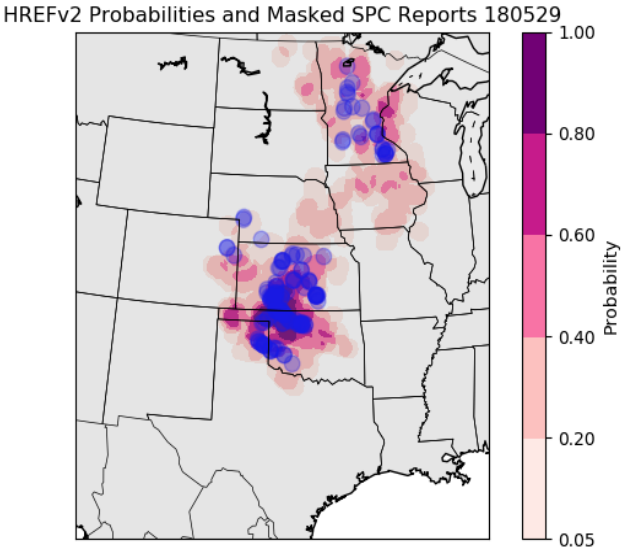

FIG. 3. HREFv2 hail probabilities plotted with masked 25-mile SPC hail reports.

4. Results

a. Full Season Reliability Diagram

The full season reliability diagram indicates that the HREFv2 ML hail forecasts exhibited an over-prediction bias for hail, evident in the reliability diagrams (Fig. 4). Interestingly, the HREFv2 ML hail forecasts show better skill at predicting hail (less bias) in more severe events. When the risk for hail on a certain day is not as high, the HREFv2 ML hail forecasts generally exhibit greater over-prediction (not shown).

b. Calibration with SPC Practically Perfect Forecasts

In order to improve the reliability of the HREFv2 forecasts, the forecasts must be properly calibrated. To correct for the over-prediction bias noted above, the HREFv2 reliability will be calibrated to the SPCs practically perfect forecasts. A forecast is considered practically perfect when it resembles a forecast that would have been created if the forecaster had perfect knowledge of the events beforehand (Hitchens et al. 2013). By calibrating the HREFv2 ML hail forecasts to the SPC practically perfect forecasts, more operationally-useful predictions of hail can be obtained.

SPC produces practically perfect hail forecasts for 4-hour intervals that include 17-23Z, 19-23Z, and 21-01Z, as well as for 20-hour intervals from 16-12Z. For each time increment, the practically perfect forecasts predict for all hail and significant hail threats. In this study, each 4-hour interval will be evaluated for all hail and compared to the respective HREFv2 model run.

The SPCs practically perfect forecasts tend to under-forecast hail, which is to be expected, as SPC probabilities will never exceed 60%. Using isotonic regression, with the SPC practically perfect data as the truth or target point for the HREFv2 calibration, the HREFv2 ML forecasts are calibrated to produce a similar distribution of probabilities. The green line in Fig. 5 becomes the new corrected reliability line for the HREFv2 model. For each time interval evaluated, the calibrated HREFv2 ML forecast reliability closely follows the SPC practically perfect reliability line, indicating that the isotonic regression was successful in producing ML forecasts with similar reliability properties to the SPC practically perfect data.

c. Calibration with Observations

It is also desirable to produce calibrated versions of the HREFv2 ML hail forecasts which are as reliable as possible. To calibrate the HREFv2 ML hail forecasts for maximum reliability (minimal bias), the hail observations are used as the target in the isotonic regression. In the resulting calibrated forecasts (Fig. 6), there is very little bias, and the forecasts exhibit near perfect reliability, as opposed to the under-prediction bias of the SPC practically perfect forecasts and the over-prediction bias of the raw HREFv2 ML forecast output.

d. Cross-Validation

In order to ensure that the post-calibrated HREFv2 forecasts are skillful when presented with new data, the calibrated forecasts will be cross-validated. The raw HREFv2

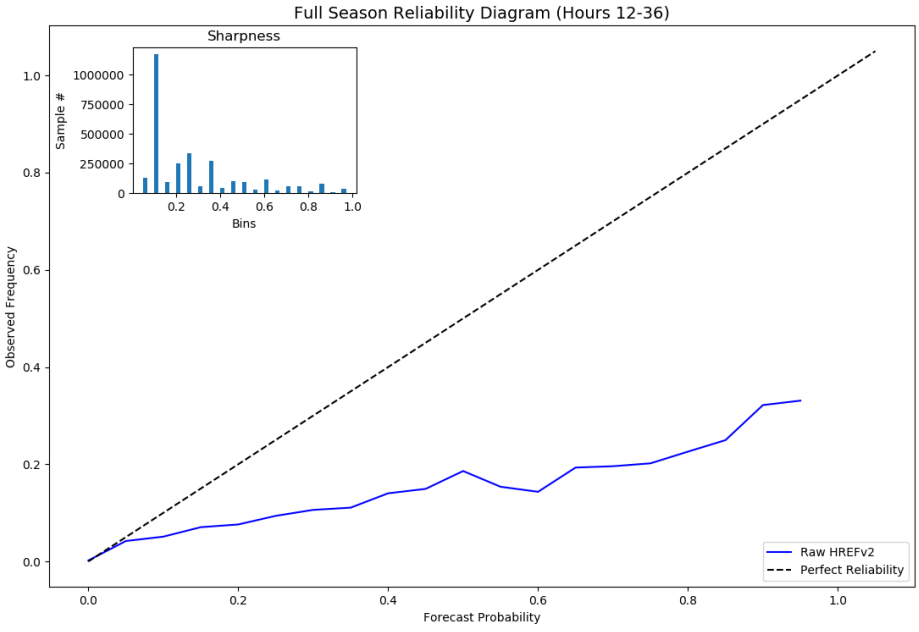

FIG. 4. HREFv2 forecast reliability diagram and sharpness for the full season (April 30th to June 1st).

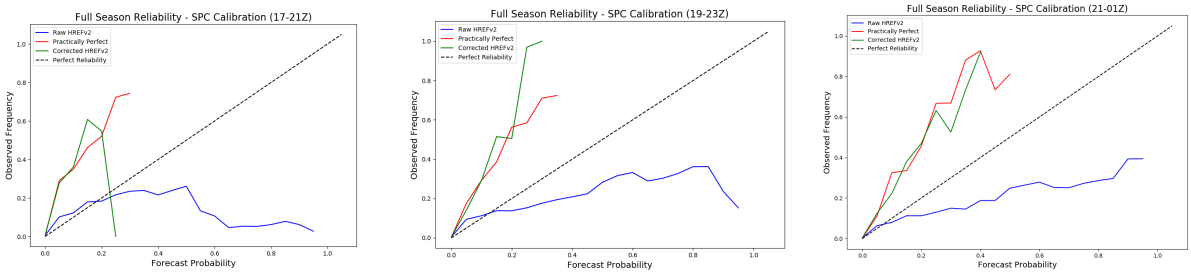

FIG. 5. Reliability diagrams for the full season, for times 17-21Z, 19-23Z, and 21-01Z, where the blue line shows the uncorrected HREFv2 reliability, red line shows SPC practically perfect reliability, and green line is corrected HREFv2 forecast reliability targeted to SPC forecasts.

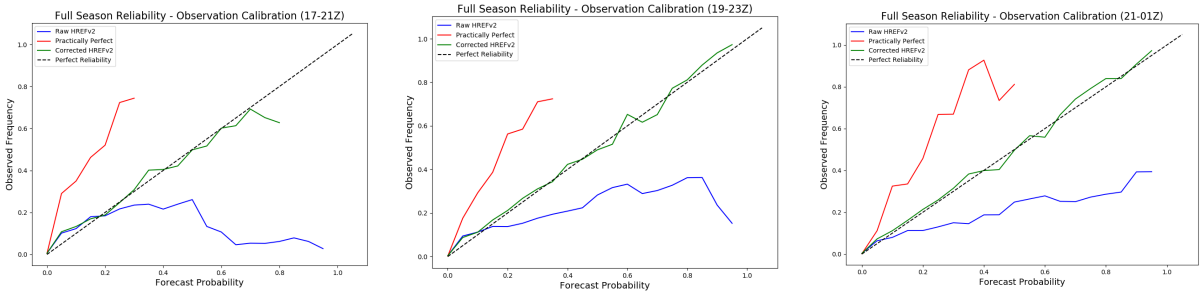

FIG. 6. Reliability diagrams for the full season, for times 17-21Z, 19-23Z, and 21-01Z, where the blue line shows the uncorrected HREFv2 reliability, red line shows SPC practically perfect reliability, and green line is corrected HREFv2 forecast reliability targeted to observational data.

ML forecasts and observations will be used as the training data for four weeks worth of data, and one week of raw HREFv2 data will serve as the testing data. The testing data are forecasts that the algorithm has not yet seen, and this will give insight as to whether or not the ML algorithm can produce reliable forecasts upon receiving new

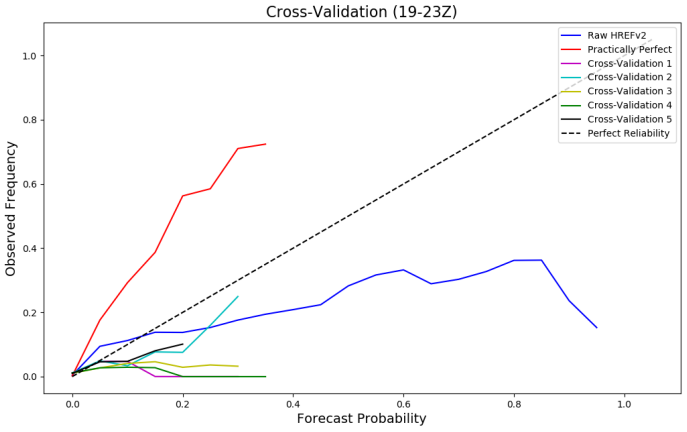

FIG. 7. Cross-Validation results when trained on raw HREFv2 forecast data for 19-23Z.

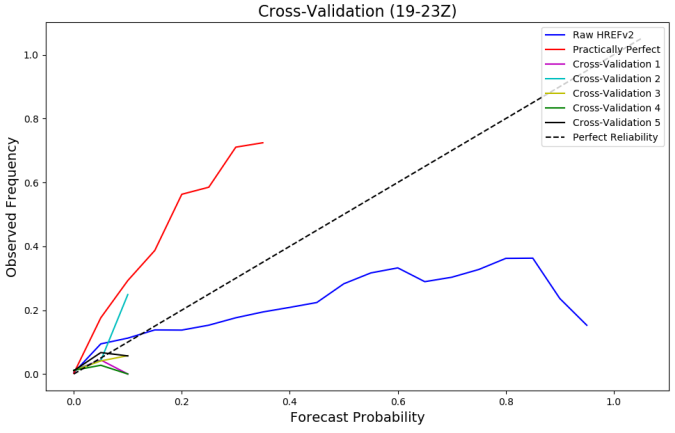

FIG. 8. Cross-Validation results when trained on SPC practically perfect forecasts for 19-23Z.

TABLE 1. The following table shows the breakdown of days that were cross-evaluated for both training and testing.

	Training Days	Testing Days
Cross-Validation 1	May 6 - June 1	April 30 - May 5
Cross-Validation 2	April 30 - May 5, May 13 - June 1	May 6 - May 12
Cross-Validation 3	April 30 - May 12, May 20 - June 1	May 13 - May 19
Cross-Validation 4	April 30 - May 19, May 28 - June 1	May 20 - May 27
Cross-Validation 5	April 30 - May 27	May 28 - June 1

data. The split of the train and test data can be seen in the Table 1.

Cross-validation was performed for time interval 19-23Z. The observational (Fig. 7) and SPC practically perfect forecasts (Fig. 8) were used as training data. Based off these figures, the cross-validated outputs are not producing skillful forecasts. Fig. 7 demonstrates that when trained on the raw HREFv2 output and the observations, the cross-validated results are more reliable when compared to the SPC practically perfect cross-validation results. Cross-validation 5 is the most optimized train/test combination, as it is the most reliable compared to the other four validations.

5. Discussion

The aim of this research is to improve the HREFv2 ML hail prediction model, which was used to produce real-time predictions of severe hail during the 2018 HWT spring forecast experiment. Because the HREFv2 ML forecasts exhibit a consistent bias towards over-prediction of hail, it is evident that some sort of calibration must take

place to maximize the operational usefulness of the forecasts. Isotonic regression is used to calibrate the HREFv2 data against both SPC practically perfect forecasts and observed hail reports. When tested on the SPC practically perfect forecasts, the corrected HREFv2 ML hail forecasts had a similar under-prediction bias as the SPC practically perfect reliability trend. When calibrated against the observed hail data, the corrected HREFv2 output produced more reliable forecasts which exhibited very little bias.

The overarching goal is to improve these algorithms to be implemented into operational severe weather forecasting. Once calibrated, the HREFv2 ML hail forecast model can be made to produce highly reliable forecasts, or, alternatively, forecasts which closely mimic the probability distribution of the SPC practically perfect forecasts. Depending on the end user, either form of calibration data could be utilized when training the HREFv2 forecasts. When targeted to the SPC practically perfect data, the calibrated forecast under-predicts hail, but produces output consistent with the practically perfect forecasts that are utilized in operational forecasting today. If a more reliable forecast is desired, then using observed hail reports as calibration data for the HREFv2 ML forecasts is ideal.

Cross-validation is also useful for ensuring that the ML algorithm will produce skillful forecasts upon seeing new data. In order to produce better cross-validation results, a larger dataset would be ideal. Cross-validating on only five weeks of data was not sufficient. The results were limited by the size of the dataset, which was constrained by the duration of the HWT spring experiment. However, moving forward, a larger dataset of days used to train and test the ML algorithms will be used.

In conclusion, a post-calibrated version of the HREFv2 ML model is necessary if it is to be used for operational forecasting purposes, since the raw HREFv2 forecasts over-predict for severe hail.

Acknowledgments. This work would not have been possible without the funding from the National Science Foundation, Grant AGS 1560419. The corresponding author would like to thank Dr. Daphne LaDue and the National Weather Center Research Experience for Undergraduates for the opportunity to conduct research. Additionally, she would like to thank Melanie Schroers, Amanda Burke, and Jacqueline Waters for all their help and guidance during this summer.

References

Experimental Forecast Program of the NOAA/Hazardous Weather Testbed, 2018: Spring forecasting experiment: Program overview and operations plan. URL https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE2018_operations_plan.pdf.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and Forecasting*, **32 (5)**, 1819–1840, doi:10.1175/WAF-D-17-0010.1.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Weather and Forecasting*, **28 (2)**, 525–534, doi:10.1175/WAF-D-12-00113.1.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, **98 (10)**, 2073–2090, doi:10.1175/BAMS-D-16-0123.1.

Murphy, A. H., 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8 (2)**, 281–293, doi:10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2.

National Severe Storms Laboratory, ????: Nssl research: Hail. URL https://www.nssl.noaa.gov/research/hail/.

Snook, N., Y. Jung, J. Brotzge, B. Putnam, and M. Xue, 2016: Prediction and ensemble forecast verification of hail in the supercell storms of 20 may 2013. *Weather and Forecasting*, **31 (3)**, 811–825, doi:10.1175/WAF-D-15-0152.1.

World Weather Research Programme, W. C. R. P., 2017: Forecast verification methods across time and space scales. URL http://www.cawcr.gov.au/projects/verification/#Introduction.