Calibration of Machine Learning–Based Probabilistic Hail Predictions for Operational Forecasting

AMANDA BURKE

School of Meteorology, University of Oklahoma, Norman, Oklahoma

NATHAN SNOOK

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

DAVID JOHN GAGNE II

National Center for Atmospheric Research, Boulder, Colorado

SARAH MCCORKLE

Indiana University, Bloomington, Indiana

AMY MCGOVERN

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

(Manuscript received 16 May 2019, in final form 13 November 2019)

ABSTRACT

In this study, we use machine learning (ML) to improve hail prediction by postprocessing numerical weather prediction (NWP) data from the new High-Resolution Ensemble Forecast system, version 2 (HREFv2). Multiple operational models and ensembles currently predict hail, however ML models are more computationally efficient and do not require the physical assumptions associated with explicit predictions. Calibrating the ML-based predictions toward familiar forecaster output allows for a combination of higher skill associated with ML models and increased forecaster trust in the output. The observational dataset used to train and verify the random forest model is the Maximum Estimated Size of Hail (MESH), a Multi-Radar Multi-Sensor (MRMS) product. To build trust in the predictions, the ML-based hail predictions are calibrated using isotonic regression. The target datasets for isotonic regression include the local storm reports and Storm Prediction Center (SPC) practically perfect data. Verification of the ML predictions indicates that the probability magnitudes output from the calibrated models closely resemble the day-1 SPC outlook and practically perfect data. The ML model calibrated toward the local storm reports exhibited better or similar skill to the uncalibrated predictions, while decreasing model bias. Increases in reliability and skill after calibration may increase forecaster trust in the automated hail predictions.

1. Introduction

Hail is a high-impact severe weather hazard, annually causing in excess of $1 billion (U.S. dollars) of property damage and $1 billion of crop damage (Jewell and Brimelow 2009). Isolated hail events, especially those impacting large urban areas, are particularly damaging. For example, a single hailstorm during the afternoon rush hour in the Denver, Colorado, metropolitan area on 8 May 2017 resulted in $2.3 billion of insurance claims (Svaldi 2018). The economic impacts of severe hail underscore the need for accurate and timely predictions, which allow individuals and businesses to take action toward mitigating risk to their property and safety. Accurate predictions of hail remain a challenge given the rapid evolution of hail-producing convective storms, coupled with uncertainties and limitations of atmospheric observation data needed to properly resolve the small-scale convective environment.

To produce skillful hail forecasts through explicit hail prediction, numerical weather prediction (NWP) models must accurately predict the development of convective

Corresponding author: Amanda Burke, aburke1@ou.edu

DOI: 10.1175/WAF-D-19-0105.1

storms, as well as produce reasonably accurate representations of hail within the model's microphysical scheme (Labriola et al. 2017). At these small scales, model forecast errors can lead to large uncertainties in the timing and location of convective storms (e.g., Kain et al. 2010b; Durran and Weyn 2016). For hail prediction on longer time scales (up to 48 h), most methods rely on approximating environmental data at the convective scale (e.g., Johns and Doswell 1992). However, the spatial and temporal coverage of atmospheric soundings are generally insufficient to provide accurate initial conditions for explicit prediction of storms on the convective scale.

Where limitations in scale cause uncertainties in local storm characteristics, convection-allowing models (CAMs) have shown skill in predicting convective morphologies (e.g., Weisman et al. 2008). In recent years, CAMs have been employed in NOAA's Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE). For example, during the 2010 HWT SFE operational forecasters subjectively indicated that the CAM guidance improved convection forecasts, compared to traditional convective-parameterizing schemes (Clark et al. 2012). Also, Gallo et al. (2017) noted that CAMs played an important role in reliable short-term forecasts, especially hourly forecasts, during the 2015 HWT SFE. For day-ahead forecasts (12–36-h lead time), CAM ensemble forecasts have shown improved skill compared to individual deterministic CAM forecasts (Loken et al. 2017).

Explicit hail prediction using storm-scale ensembles has been previously studied (e.g., Adams-Selin and Ziegler 2016; Snook et al. 2016; Labriola et al. 2017, 2019). However, limitations exist when explicitly predicting hail, including the sensitivity of microphysical scheme choice, initial and boundary conditions varying across ensembles, and physical assumptions needed to predict hail with computational efficiency. Additionally, the multitude of operational models and ensembles predicting hail that are available to forecasters can lead to cognitive overload. Wilson et al. (2017) found that forecaster overload increased with an increase in the number of datasets monitored, especially when multiple warning decisions are needed.

Recently, studies have focused on using machine learning (ML) to synthesize large amounts of atmospheric data, reducing the amount of monitored data while producing skillful forecasts products without explicit prediction assumptions (e.g., Gagne 2016; Gagne et al. 2017, hereafter G17; McGovern et al. 2017; Lagerquist et al. 2017; Herman and Schumacher 2018b,a). Instead of explicit prediction, ML models map a set of inputs to a given output by optimizing the model's structure, such that the differences between the ML predictions and the output observations, or "ground truth", are minimized. Using these learned structures, ML models are able to make predictions on new sets of model data with relatively minimal computational expense. Low computational expense, compared to other postprocessing methods applied to gridded NWP output for hail prediction (e.g., Adams-Selin and Ziegler 2016), is a major advantage of ML forecasts. ML-based hail prediction studies over the contiguous United States (CONUS) have demonstrated that ML predictions exhibit greater forecasting skill over direct prediction of hail from NWP model output or the use of proxy variables (Gagne 2016; G17). However, subjective commentary from the 2018 HWT SFE indicated that forecasters do not trust ML guidance if the output is unfamiliar or dissimilar to human-produced forecasts.[1]

Previous studies on applying automation in the forecasting process emphasize human interaction (e.g., Snellman 1977; Bosart 1989; Moller et al. 1994), and the importance of reliable guidance over simple competence (Hoffman et al. 2013). Similar to the 2018 HWT SFE, forecasters during the 2014 HWT SFE did not trust guidance without knowing the reliability and skill of new products (Karstens et al. 2015). However, Karstens et al. (2018) found that when proper training and forecast verification results are provided, the addition of automation can increase forecaster productivity. One way to increase the skill and reliability of probabilistic forecasts is through calibration (e.g., Raftery et al. 2005; Hagedorn et al. 2008; Hamill et al. 2008). In addition to increases in forecast performance, calibrating ML output to resemble existing operational forecasts, specifically those produced by the Storm Prediction Center (SPC), could result in greater trust in automated guidance for operations.

In this study, adapted from Burke (2019), we present newly-developed hail forecast guidance products using ML algorithms and output from the operational HREFv2 model. We demonstrate that these day-ahead forecast products can be successfully calibrated to increase reliability and skill, as well as resemble SPC hail products, all to increase forecaster trust in automated hail guidance.

2. Data and methods

a. Data

The ML-based hail prediction models investigated in this study use HREFv2 (Jirak et al. 2018) model output as input data. Starting in April 2017, the SPC

[1] Forecasters from 2018 HWT SFE on14–18 May.

TABLE 1. Configuration for the eight-member High-Resolution Ensemble Forecast system, version 2 (HREFv2), model output, including planetary boundary layer (PBL) schemes and an additional conditional random forest (RF) threshold for classifying member storm objects as producing hail. The ensemble members include the high-resolution window (HiResW) Advanced Research version of the Weather Research and Forecasting (WRF) Model (ARW), HiResW Nonhydrostatic Multiscale Model on the B-grid (NMMB), HiResW National Severe Storm Laboratory (NSSL)-like version of the WRF-ARW, and nested North American Mesoscale Model (NAM-NEST).

Members	Initializations	PBL	Microphysics	Vertical levels	Grid spacing (km)	Threshold
HiResW ARW	0000 UTC	YSU	WSM6	50	3.2	0.17
HiResW ARW	1200 UTC	YSU	WSM6	50	3.2	0.17
HiResW NMMB	0000 UTC	MYJ	Ferrier–Aligo	50	3.2	0.14
HiResW NMMB	1200 UTC	MYJ	Ferrier–Aligo	50	3.2	0.23
HiResW NSSL	0000 UTC	MYJ	WSM6	40	3.2	0.12
HiResW NSSL	1200 UTC	MYJ	WSM6	40	3.2	0.15
NAM Nest	0000 UTC	MYJ	Ferrier–Aligo	60	3	0.22
NAM Nest	1200 UTC	MYJ	Ferrier–Aligo	60	3	0.14

began running the High-Resolution Ensemble Forecast system, version 2 (HREFv2). The HREFv2 is based off the Storm Scale Ensemble of Opportunity (SSEO), a ''poor man's ensemble'' (Ebert 2001) that combines existing operational CAMs produced by NOAA (Jirak et al. 2012) to produce a computationally efficient ensemble. Between the 2012–15 HWT SFEs, the SSEO scored positively in both objective and subjective metrics and provided a baseline for evaluation of CAM model guidance during the 2016 HWT SFE (Clark et al. 2016). The success of the SSEO in probabilistic severe weather prediction has brought attention to the HREFv2 dataset for use in skillful weather prediction. Compared to the SSEO, the HREFv2 produces slightly smaller grid spacing and includes an Advanced Research version of the Weather Research and Forecasting (WRF) Model (WRF-ARW; Skamarock and Klemp 2008) for a total of eight members.

Developed by National Centers for Environmental Prediction (NCEP)/Environmental Modeling Center (EMC), the HREFv2 is run daily by NCEP Central Operations (NCO).[2] The HREFv2 is an eight-member ensemble, with time-lagged members initialized at 0000, 0600, 1200, and 1800 UTC. Only the members initialized at 0000 and 1200 UTC were available for use in this study. The HREFv2 is a diverse ensemble consisting of multiple microphysical schemes, including the WRF single-moment six-class (Hong et al. 2010) and Ferrier–Aligo (Aligo et al. 2014) scheme, as well as multiple initial conditions. The planetary boundary layer (PBL) schemes are the Yonsei University (YSU) (Hong et al. 2006) and the local Mellor–Yamada (MYJ) (Janjić 1990, 1994). All members use a horizontal model grid spacing of approximately 3 km. The number of vertical levels at which data are produced differs between ensemble members; the four high-resolution window (HiResW) members, including the time-lagged members, produce 50 vertical levels, the two ''National Severe Storm Lab (NSSL) like'' ARW models output 40 vertical levels, and the two nested North American Mesoscale Model (NAM-NEST) members include 60 vertical levels. Forecast products are generated from 1200 UTC to 1200 UTC the next day, the same period as a SPC day-1 convective outlook. A detailed description of the eight members of the HREFv2 ensemble is provided in Table 1.

The target data to train the ML models are the Maximum Expected Size of Hail (MESH) (Witt et al. 1998) derived from NOAA/NSSL Multi-Radar Multi-Sensor (MRMS) radar data (Zhang et al. 2011; Smith et al. 2016). Because MESH outputs exhibit greater skill for values exceeding 19 mm (Wilson et al. 2009), only values greater than 19 mm are considered. Although MESH has known biases, such as overprediction of higher values (e.g., Wilson et al. 2009; Cintineo et al. 2012; Ortega 2018; Murillo and Homeyer 2019), MESH was chosen over the local storm reports (LSRs) as the observational dataset for training the ML models because of the known population and size biases with LSRs (e.g., Schaefer et al. 2004; Cintineo et al. 2012). Also, Melick et al. (2014) found that MESH hail swaths were more skillful than LSRs in observing hail objects and act as a useful independent dataset in low population areas.

The ML models are trained on data between 1 April and 31 July 2017, and tested on data from 1 May through 31 August 2018. Different years are used for training and testing to create independent datasets and reduce the chance of overfitting. The duration of the training period (April–July) is selected based on greater hail potential and number of observations over the CONUS in the spring and early summer. The testing period includes the 2018 HWT SFE, from 30 April

[2] https://www.spc.noaa.gov/exper/href/.

TABLE 2. The 29 HREFV2 storm and environment variables extracted during object tracking. Multiple levels indicate the variable was investigated at each separate level. CAPE is convective available potential energy, CIN is convective inhibition, MAXUVV is the maximum hourly upward vertical velocity, and MAXDVV is the maximum hourly downward vertical velocity.

Storm		Environmental	
Variable	Level(s)	Variable	Level(s)
MAXUVV	—	Precipitable water	—
Storm relative helicity	1 and 3 km		
Hourly max reflectivity	1 km	Temp	500, 700, 850, and 1000 hPa
MAXDVV	—	Dewpoint temp	
Hourly max UH	2–5 km	Geopotential height	500, 700, and 850 hPa
		U wind	
		V wind	
		Hourly max *U* wind	—
		Hourly max *V* wind	—
		Surface lifted index	—
		CAPE	—
		CIN	—

through 1 June 2018, during which forecasts were provided to HWT SFE participants for evaluation and feedback. April 2018 is not included for testing because of incomplete HREFv2 data.

b. *Data preprocessing*

Both the input and observational datasets are preprocessed with object tracking to address the relative rarity of hail (and severe weather in general) at any given location within the CONUS. The object-tracking method and ML models evaluated for hail prediction are based upon those used in G17. This object-tracking algorithm identifies potential storm objects where a chosen variable field exceeds a user-specified threshold. For the HREFv2 dataset, storm objects are identified using the maximum hourly upward vertical velocity (MAXUVV) with a threshold of $8\,\mathrm{m\,s^{-1}}$, rather than column total graupel greater than $3\,\mathrm{kg\,m^{-2}}$ from G17. The selection of updraft speed rather than updraft helicity or column total graupel, and the use of a relatively low threshold value, were designed to capture all possible hail storms rather than only high-end supercells. Although supercells are typically responsible for the most severe hail events, marginal hail is also important to the public, the insurance industry, and agriculture. For observations, potential MESH storm objects are generated for values exceeding 19 mm, differing from the 12-mm threshold used in G17, for reasons outlined above.

After identification, potential storm objects are matched in time and space to create storm tracks. For the HREFv2 storm tracks, additional hourly maximum variables are extracted at each grid point throughout the tracks. While instantaneous model fields may miss variations in storm intensity at time scales less than 1 h, hourly maximum fields can record maximum intensities without needing to output model data at every time step. Kain et al. (2010a) found that hourly maximum values provide skill for severe weather forecasting, particularly in determining hail threats in nonsupercellular storms. In addition, hourly maxima have been found skillful as guidance when forecasting severe weather, with minimal calibration needed (Sobash et al. 2011). Although CAMs, which output hourly maximum variables, cannot resolve individual hazards (such as severe wind, severe hail, or tornadoes), they can resolve severe hazard proxies such as updraft helicity or updraft speed.

The HREFv2 maximum hourly variables, statistically evaluated over each storm track after extraction, include storm and environmental variables (Table 2). Storm variables are directly related to convection, such as hourly maximum reflectivity, storm relative helicity (SRH), hourly maximum updraft helicity (UH), and so on. Environmental variables consist of near-storm fields, such as temperature, dewpoint temperature, geopotential height, and so on. The environmental variables are extracted from the previous forecast hour to mitigate contamination of storms on environmental conditions, with storm variables extracted at the current forecast hour. Statistical evaluations of the storm tracks for both variable types include the mean, maximum, minimum, standard deviation, skewness, and 10th, 50th, and 90th percentiles. The HREFv2 hourly maximum variable statistics serve as the input data to the ML models. The number of HREFv2 storm tracks identified for training the ML models ranges from 10 000 to 25 000 per member, and from 23 000 to 63 000 per member in the input test set. The 2018 test set is larger as it contains more days with model runs, while the 2017 data was limited because it was the first year of running the HREFv2 operationally. Increasing the number of storm tracks in the 2017 dataset would be ideal, however, any changes would

also increase the track number in 2018, maintaining the dataset imbalance issues.

The final preprocessing step matches the HREFv2 storm tracks to MESH tracks. Track pairings occur where the calculated distance between member storm tracks and observed hail tracks is less than 80 km. If any of the fields used to calculate distance (differences in the observed and modeled track starting times, locations, durations, and sizes) exceed the thresholds defined in G17, the tracks are not matched. Match classifications, a binary dataset, as well as the shape and scale parameters of the paired MESH tracks provide the observational datasets for training the ML models.

c. *Machine learning methods*

The algorithm for operationally calibrating the ML hail predictions includes three models (Fig. 1): a random forest (RF) (Breiman 2001) classification model, RF regression model, and an isotonic regression (IR) model (Niculescu-Mizil and Caruana 2005). Both RF model configurations include 500 trees, a square root number of features chosen per tree, and a requirement of 1 sample per leaf. The number of random features chosen for each tree is relatively low (about 14) to reduce the chance of overfitting to the limited training dataset. The regression model optimizes the mean square error, while the classification RF optimizes the Gini index. These hyperparameters are similar to those in G17 and are relatively standard. The choice of RFs for producing operational severe hail forecasts is partly due to the speed for both training and forecasting. Computational efficiency and cost are large considerations for operations in addition to skillful forecasts. For forecasting rare events, RFs have shown greater skill over other models that assume linearity, such as logistic regression or elastic nets (G17; Herman and Schumacher 2018a). The randomness associated with RFs decreases model bias and variance, creating strong classifiers and regressors (Breiman 2001). Finally, RFs are easily interpretable (Herman and Schumacher 2018a), which is also a large consideration when employing a postprocessing method for operations.

Before calibration, described in section 2d, a RF classification model is trained to predict the probability of HREFv2 member storm tracks being matched with hail observations, where the binary truth dataset is described in section 2b. For each ensemble member, fivefold cross validation (Breiman and Spector 1992) produces five probability predictions, one per validation test fold, from the training dataset. Calculation of contingency table metrics over all five probability estimations, at 1% intervals, determines the threshold with the highest equitable threat score (ETS). This threshold "filters out" ensemble member storm objects because only storm tracks with probabilities exceeding the estimated threshold are classified as hail producing. Table 1 includes the thresholds for each HREFv2 member.

The member storm tracks classified as hail producing (with probability values exceeding the threshold values mentioned above, about 600–2000 storm tracks per ensemble member in the training dataset and 4000–7500 tracks in the testing set) are evaluated using a RF regression model. The regression model predicts the scale and shape parameters of a MESH gamma distribution for each input storm track. As in G17, the scale and shape parameters of the hail size gamma distribution are log-normalized and predicted together, to maintain the negative correlation between the shape and scale parameters. Comparing the predicted shape and scale parameters from each ensemble member to the values found through object-tracking produces average root mean squared error (RMSE) values of 0.64 for the shape parameter, and 4.51 for the scale parameter. The average ensemble mean absolute error (MAE) values are 0.51 and 3.28 for the shape and scale parameters, respectively. A higher scale parameter error could indicate that the RF regression model predicts a larger range of hail size values than observed.

Using the predicted shape and scale parameters, hail sizes from the gamma distribution are extracted such that the highest storm object values, MAXUVV in this case, in a track are associated with the largest MESH values. Unlike G17, the distribution of storm objects created for each member are based off data from the entire training period. To output hail sizes for a given storm object, G17 matched percentiles from a daily hail size distribution with a daily storm object distribution. However, if the range of storm objects on a given day is relatively low, higher percentile hail values are matched to storm object values that would not result in large hail on a different day. Preliminary testing identified a subjective high MESH bias on marginal days when using the daily values.

After predicting hail sizes from each ensemble member, the ensemble maximum size and neighborhood maximum ensemble probability (NMEP) of hail within 42 km of a point are calculated from 1200 UTC to 1200 UTC the following day. The NMEP predictions, based off the definition of ensemble probabilities in Schwartz and Sobash (2017), are evaluated at the severe (>25 mm) and significant severe (>50 mm) hail thresholds on the 3-km HREFv2 grid. The grid is further smoothed with a 2D Gaussian filter (σ = 42 km). In addition to managing the uncertainty of severe weather, the

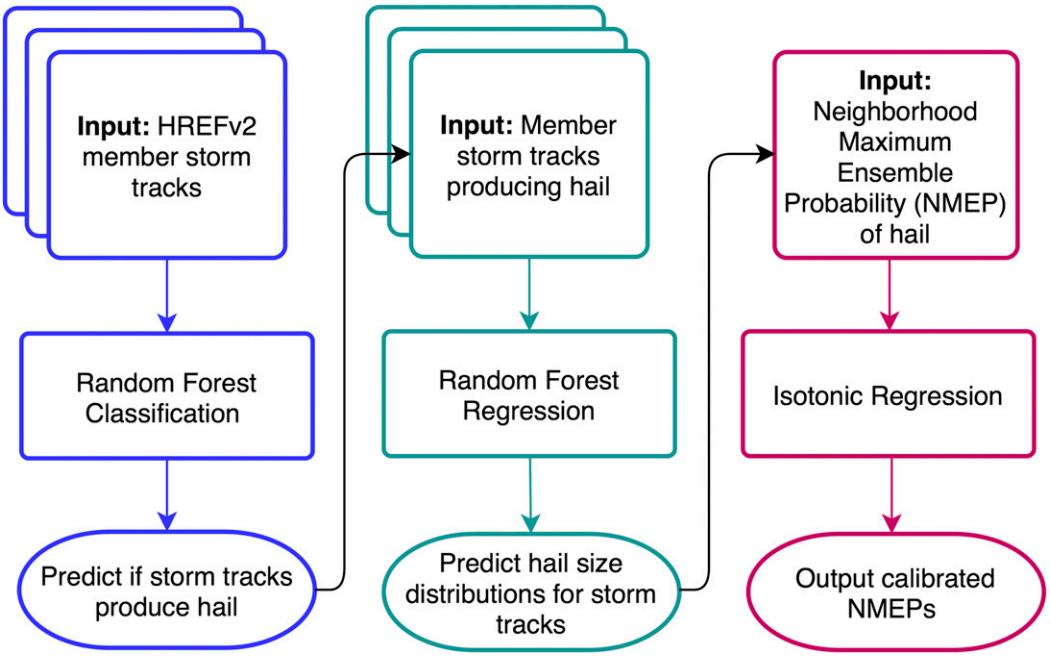

FIG. 1. Process for producing and calibrating machine learning hail predictions. The random forest classification and regression models are trained using input from each HREFv2 member separately, while the aggregated ensemble predictions are input to the isotonic regression model for training.

smoothed grid allows for direct comparison of predictions to SPC products for verification. The uncalibrated NMEP predictions were generated daily at 0700 UTC during the test period, valid at 1200 UTC. For a complete description of the data preprocessing and machine learning methods, we refer the reader to G17.

d. Calibration methods

In the uncalibrated NMEP hail predictions, probabilities range from 0% to 100% while SPC hail outlooks never predict probabilities of hail exceeding 60%. Ideally, to build forecaster trust in the ML-based model, the hail forecasts should be reliable and comparable in probability magnitude with operational products such as the SPC hail outlook. To address these issues, an IR model, chosen for its computational efficiency and lack of linearity assumptions, calibrates the RF output for operations. Previous studies have shown that RF probabilistic forecasts are more reliable after calibration using IR (Niculescu-Mizil and Caruana 2005; Lagerquist et al. 2017). With IR, which uses a nondecreasing function to map an input dataset toward a target dataset, only nonzero values are altered.

The operational target datasets used to calibrate the RF NMEP hail forecasts include the local storm reports (LSRs) and SPC practically perfect (PP) (Davis and Carr 2000; Hitchens et al. 2013) probability of severe hail occurrence. The LSR target dataset is a binary field, such that the observed hail grid points are within 40 km of at least one severe or significant severe hail report. The SPC PP probability field serves, during the HWT SFE, as an estimate of the optimal outlook a forecaster would issue if all LSR locations were known beforehand. After identifying the LSRs as a binary field, application of a two-dimensional Gaussian smoothing filter ($\sigma = 2$ for operations) creates PP smoothed probabilistic forecasts to account for uncertainties in LSR placement.

The calibration model trains and tests over the RF predictions (121 days between 1 May and 31 August 2018), as the model for creating calibrated probabilities can only be applied to existing probability predictions. A total of 84 training (70% of data) and 37 testing (30% of data) days were randomly selected to limit biases resulting from synoptic or seasonal patterns. We acknowledge that including a validation dataset would be ideal, as well as separate years of training and testing data to decrease the chances of overfitting. We plan to split the data accordingly in future iterations when a larger dataset is available. The calibrated probabilities range in time from the day-ahead forecasts to three 4-h periods used by 2018 HWT SFE forecasters (1700–2100, 1900–2300, and 2100–0100 UTC). The target PP data does not include data from the full 24-h period, instead considering only a 20-h period from 16 to 36 h of forecast

time. However, the IR model maps the two datasets together for calibration purposes.

3. Results

The calibrated day-ahead NMEP predictions are qualitatively verified against the SPC daily hail outlook and the 20-h PP output, valid the same forecast day. We analyze two case studies to examine the robustness of the ML model to accurately predict hail occurrence over varied weather regimes. Both the severe and significant severe (sig-severe) thresholds are mapped for consistency with the G17 study. Additional quantitative evaluations over the IR test period (37 random days between 1 May and 31 August 2018) investigate the ML-based hail predictions, both uncalibrated and calibrated, against the PP dataset, the 1200 UTC SPC hail outlook, and a hail proxy variable (updraft helicity). The 2–5-km HREFv2 UH data at thresholds related to severe ($>75\,m^2\,s^{-1}$) and sig-severe ($>150\,m^2\,s^{-1}$) hail (G17) provide another non-ML baseline.

a. *Marginal hail case study*

On 8 May 2018, a trough and surface cold front moved into Oregon, where a deep mixed layer provided enough support for a few nonsevere hail-producing storms despite relatively weak convective available potential energy (CAPE) values. Storms over the mid to lower Missouri Valley had ample CAPE, but a dry boundary layer restricted growth ahead of a surface trough. The SPC hail outlook valid 1200 UTC displays two regions with 5% probability of severe hail (Fig. 2a), located over Oregon and portions extending from South Dakota to Missouri. The PP probability of severe hail on 8 May 2018 includes values between 5% and 15% highlighting South Dakota, Minnesota, and northwestern Iowa (Fig. 2b). There were no areas of sig-severe hail probability (Fig. 2c) and no severe hail reports received over the western United States.

The uncalibrated ML NMEP prediction, valid 1200 UTC 8 May 2018, displaces the severe hail probabilities up to 35% over portions of Iowa, South Dakota, Nebraska, and Missouri to the southeast of the observed severe hail reports (black dots) (Fig. 3a). Portions of Oklahoma, Kansas, and Oregon also exhibit severe probabilities up to 22%, however the PP output does not contain probabilities in these regions (Fig. 2b). Although severe hail was not reported in Oregon, the SPC hail outlook (Fig. 2a) displays a similar area of probabilities as the uncalibrated prediction. At the sig-severe threshold (Fig. 3b), the uncalibrated model outputs probabilities up to 4% in eastern Iowa and southern Minnesota, while the PP output does not produce sig-severe probabilities (Fig. 2c). Overall, the uncalibrated prediction exhibits higher magnitude probabilities than the SPC outlook and PP output, however the severe hail prediction produces spatially similar areas of nonzero hail threat as the SPC outlook.

Compared to the uncalibrated ML output, the probability magnitudes of the LSR calibrated severe hail prediction (Fig. 3c) better resemble the SPC hail outlook and the severe hail PP output. The LSR calibrated output exhibits the same spatial coverage of nonzero probabilities as the uncalibrated prediction, although values less than 1% are not displayed. Differing from the uncalibrated output, the LSR calibrated probabilities do not exceed 14%, similar to the severe PP output and SPC outlook magnitudes (Figs. 2a,b). At the sig-severe threshold, the LSR calibrated prediction (Fig. 3d) outputs spatially similar probabilities as the uncalibrated model, however again overforecasting compared to the PP sig-severe probabilities (Fig. 2c). In general, the LSR calibrated model corrects the high magnitude bias present in the uncalibrated ML severe hail forecasts, although a high spatial bias persists at both thresholds.

The PP calibrated predictions, where the target dataset for calibration is the PP dataset, output lower probabilities of severe (Fig. 3e) and sig-severe (Fig. 3f) hail compared to the LSR calibrated forecast (Figs. 3c,d). The severe hail PP dataset provides higher magnitude probabilities (up to 14%) than the PP calibrated prediction (up to 4%). At the sig-severe threshold, the PP calibrated model correctly predicts no areas exceeding a 1% chance of sig-severe hail (Fig. 3f). For a marginal case, the PP calibrated model decreases the severe hail threat in areas observing hail reports but more closely resembles observations at the sig-severe threshold.

In general, the uncalibrated hail predictions overestimate the probability of severe hail, both spatially and in magnitude. An added calibration step decreases the high magnitude bias of the uncalibrated output, where the LSR calibrated model outputs severe hail probability values more comparable to the SPC outlook and PP data. Conversely, the predictions calibrated to the PP output underestimate the severe hail threat over South Dakota and Iowa. However, the PP calibrated model predicts the low sig-severe hail threat, compared to the slightly overestimated threat from the LSR calibrated model. At both size thresholds, at least one calibration model decreases the overprediction bias associated with the uncalibrated output.

b. *Hail outbreak case study*

Examination of a high-end severe hail event, occurring on 29 July 2018, investigates the robustness of the

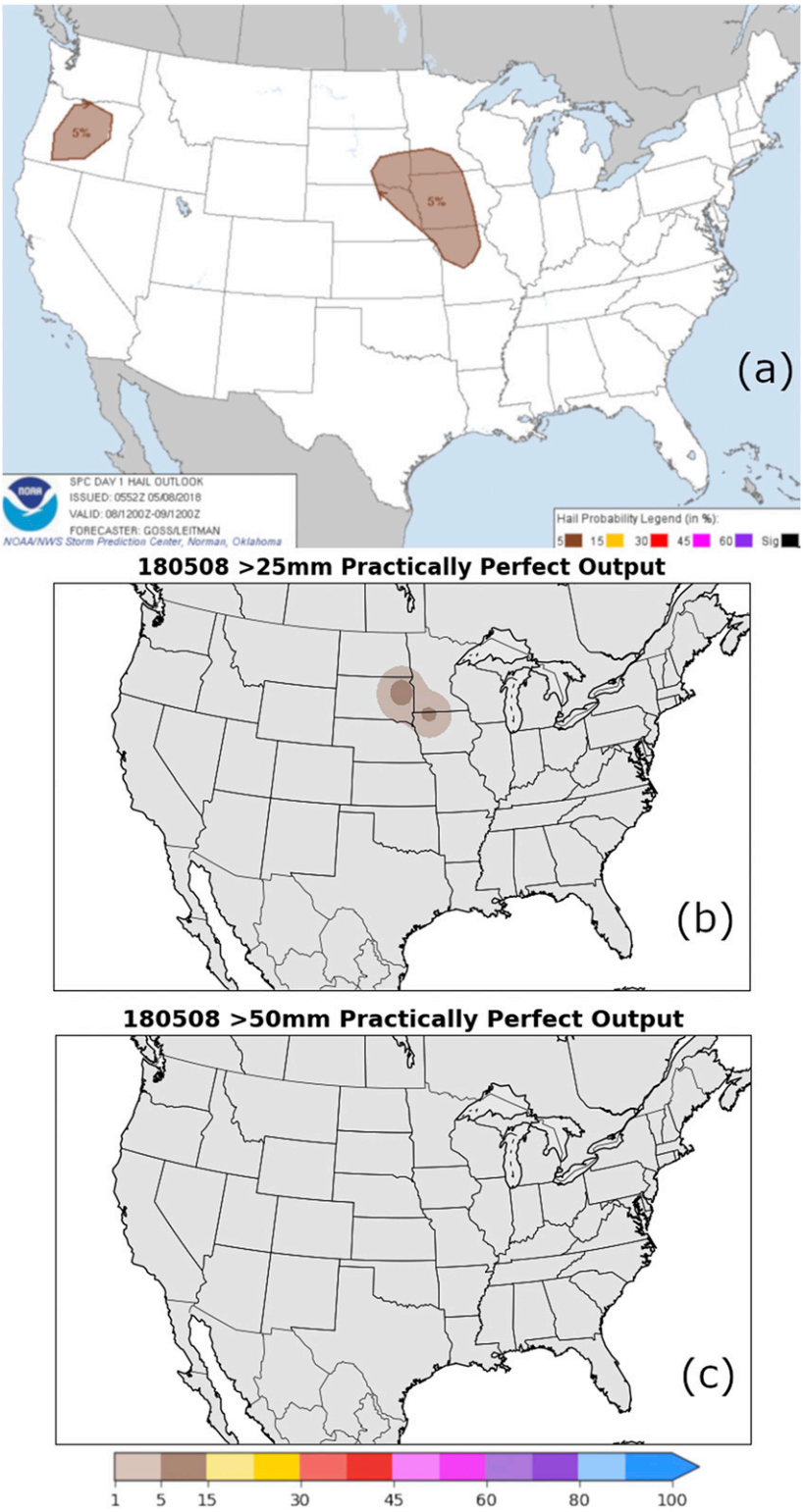

FIG. 2. SPC forecasts from 8 May 2018 including (a) the day-1 hail outlook valid 1200 UTC and practically perfect output at the (b) severe and (c) significant severe threshold.

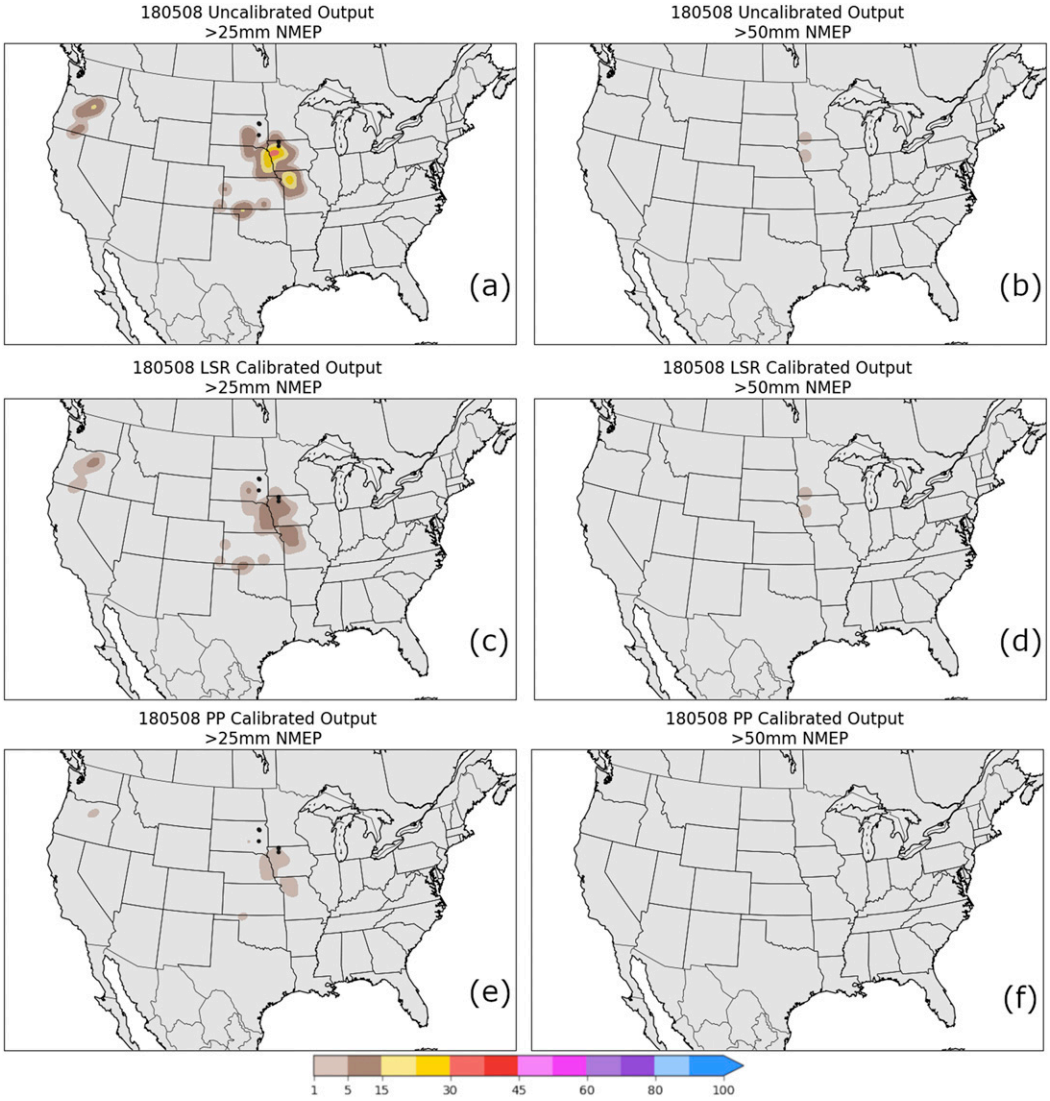

FIG. 3. ML neighborhood maximum ensemble probabilities of hail for 8 May 2018. Both calibrated and uncalibrated predictions are produced at the (a),(c),(e) severe and (b),(d),(f) significant severe hail thresholds. Predictions are calibrated to the LSRs and PP probabilities. The black dots are severe and significant severe hail reports.

ML-based hail predictions in differing environments. On this day, multiple hail-producing storms resulted in 106 severe and 30 sig-severe hail reports concentrated in Colorado and extending into Wyoming, Kansas, South Dakota, and Nebraska. A strengthening upper-level trough and midlevel jet over the Midwest, strong diurnal heating, and a moist boundary layer set the stage for severe storms over the central high plains. The day-1 SPC hail outlook valid 1200 UTC (Fig. 4a) predicts a 15% chance of severe hail over eastern Colorado and 5% from Montana to Arkansas. Sig-severe hail was not anticipated until the outlook valid 1300 UTC. The PP output indicates severe hail probabilities up to 38% over northeastern Colorado (Fig. 4b), and up to 23% for sig-severe hail over eastern Colorado (Fig. 4c).

The uncalibrated hail prediction (Fig. 5a), valid 1200 UTC, features similar areas of 5% probability as the SPC hail outlook and PP output. However, neither the uncalibrated prediction nor the SPC outlook indicate the reported severe hail threat in North Dakota or Minnesota. In eastern Colorado, the uncalibrated severe hail prediction exceeds 60%, substantially overestimating the hail threat compared to observed PP severe hail output and SPC outlook (Figs. 4a,b). The uncalibrated sig-severe hail prediction (Fig. 5b) displays comparable probabilities in eastern Colorado compared to the sig-severe PP output (Fig. 4c) where

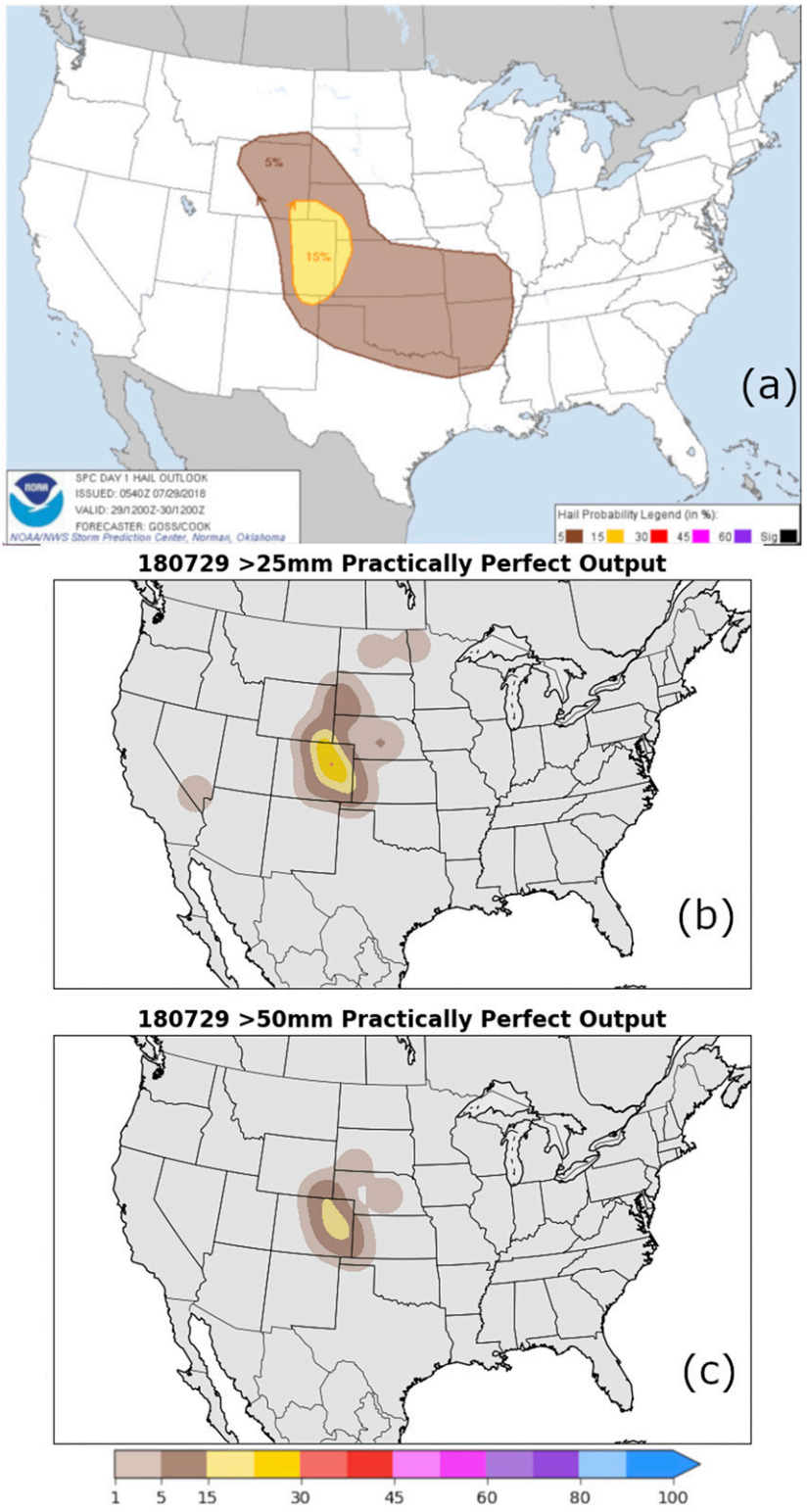

FIG. 4. As in Fig. 2, but for SPC forecasts including the (a) day-1 hail outlook valid at 1200 UTC 29 Jul 2018 and practically perfect output at the (b) severe and (c) significant severe threshold.

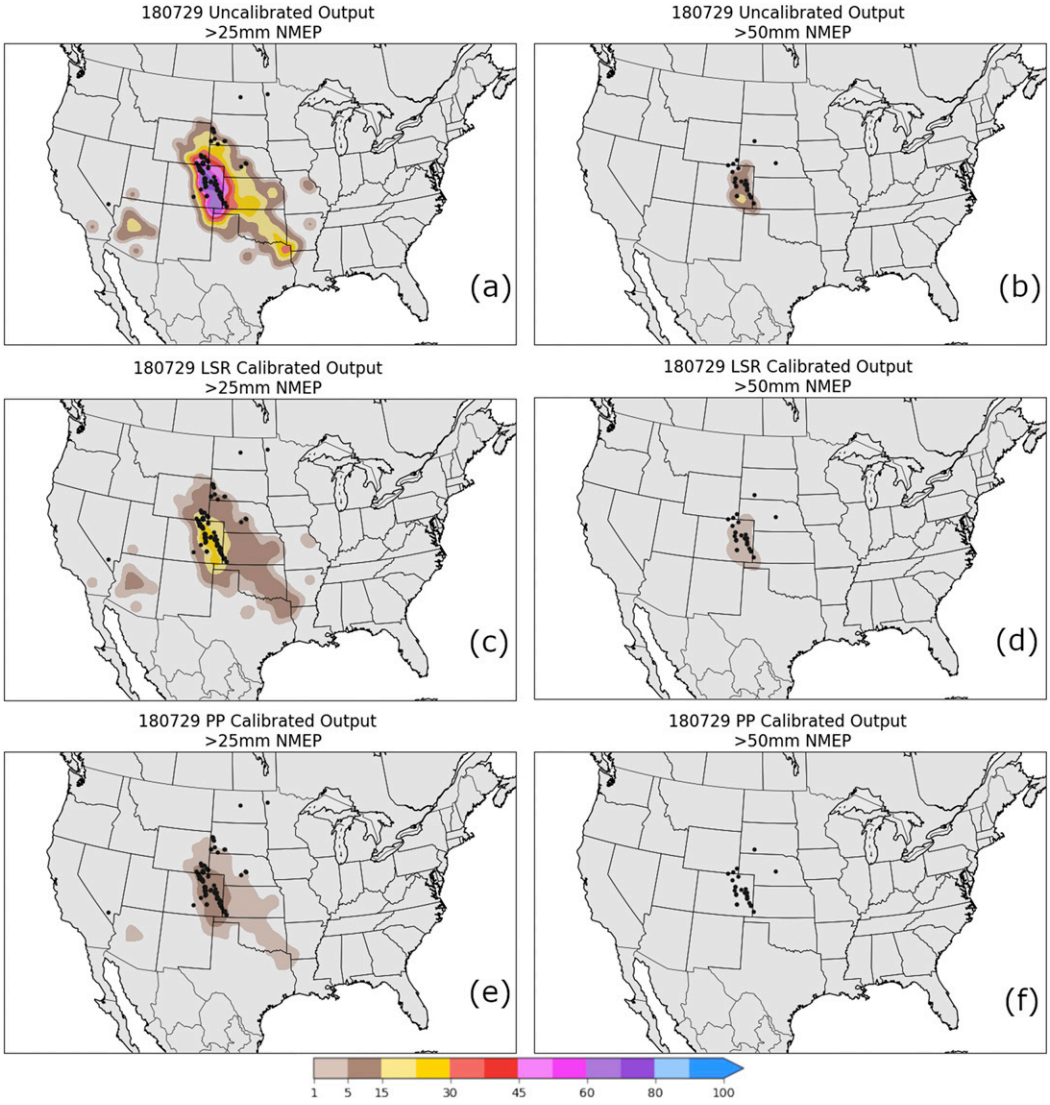

FIG. 5. As in Fig. 3, but for neighborhood maximum ensemble probabilities of hail valid on 29 Jul 2018. Predictions are produced at the (a),(c),(e) severe and (b),(d),(f) significant severe hail thresholds. Calibration is accomplished using LSRs and PP probabilities. The black dots are severe and significant severe hail reports.

values exceed 15%. Despite the overestimated hail probabilities, the uncalibrated predictions are closely collocated with the bulk of observed hail and spatially more comparable with the SPC hail outlook and PP output than in the marginal hail case.

The LSR calibrated model also outputs similar hail threat regions as the PP output and SPC hail outlook, but produces calibrated probabilities closer in magnitude to the observed severe PP output compared to the uncalibrated prediction (Fig. 5c). The calibrated probabilities do not exceed 29% over eastern Colorado, slightly lower than the maximum observed probabilities in the PP output (37%), but comparable to the hail threat in the SPC outlook. For sig-severe hail, the LSR calibrated probabilities (Fig. 5d) do not exceed 4% over portions of Colorado, again lower than the 22% maxima observed in the sig-severe PP output. In general, the LSR calibrated model predicts severe probability magnitudes closer to the PP output and SPC hail outlook, compared to the uncalibrated output, but underpredicts the sig-severe hail probabilities.

The last ML-based severe hail prediction, calibrated toward the PP output, underforecasts compared to both the PP output and SPC hail outlook (Fig. 5e). As in the marginal hail case study, the severe PP calibrated prediction exhibits very low probabilities of severe hail, not exceeding 14% in regions of eastern Colorado that observe PP probabilities of up to 22%. The sig-severe PP

TABLE 3. The contingency table metrics used to compare forecasts and observed events.

		Observed	
		Yes	No
Forecast	Yes	True positive (TP)	False positive (FP)
	No	False negative (FN)	True negative (TN)

TABLE 4. ETS and bias equations for forecast verification of the ML-based hail predictions.

Metric	Equation(s)
Bias	$\frac{TP + FP}{TP + FN}$
ETS	$\frac{TP - \text{random hits}}{TP + FN + FP - \text{random hits}}$
Random hits	$\frac{(TP + FN)(TP + FP)}{\text{total}}$

calibrated predictions (Fig. 5f) also underforecast compared to the PP output, where probabilities do not exceed 1% on this day, while the sig-severe PP probabilities reach 22% in this region. Both the severe and sig-severe PP calibrated predictions underestimate the severe hail threat in eastern Colorado, however the calibrated severe hail forecast highlights the area of greatest observed hail reports.

For the high-end severe hail case, the ML-based hail predictions focus the highest hail threat in eastern Colorado where the largest number of severe hail reports are observed. As in the marginal hail case, calibration decreases the high probability magnitude bias associated with the severe uncalibrated ML output. Also, the LSR calibrated severe hail predictions more closely resemble the observed probability magnitudes than the PP calibrated output. Last, both calibrated models underpredict the probability of sig-severe hail. As the uncalibrated model outputs relatively similar probability magnitudes compared to the observed PP output, calibration may not be necessary at the sig-severe threshold for this case.

c. *Quantitative verification*

Four metrics quantitatively verify the ML models, PP output, 1200 UTC SPC outlook, and UH proxy variable. The metrics [reliability, Brier skill score (BSS; Brier 1950), equitable threat score (ETS; Clark et al. 2010), and bias] evaluate the 24-h predictions (12–36 h) over the isotonic regression test set. ETS and bias are favorable for evaluating gridded forecasts (Hamill 1999), where ETS measures the fraction of correct forecasts to observed events and takes into account events randomly forecasted correctly. Higher ETS values are more skillful. Bias compares the frequency of forecast events to the frequency of observed events, where a bias of 1 is preferred. Both metrics require dichotomous forecasts to calculate contingency table metrics (Tables 3 and 4). The probabilistic forecasts are made deterministic at 5% interval thresholds, where forecasts equal to or greater than the threshold are predicted events. The observations for calculating the contingency table metrics are the LSRs (already binary) and PP dataset (applied thresholds similar to the forecasts).

Evaluating the forecasts using reliability and BSS, with the LSRs as observational truth, reveals that the PP dataset consistently underpredicts while the uncalibrated ML predictions overpredicts, at both size hail thresholds (Fig. 6). The 1200 UTC SPC outlook is only available at discrete intervals, however the forecasts exhibit near perfect reliability, slightly overforecasting at 45% (Fig. 6a). The severe uncalibrated hail prediction and UH proxy exhibit comparable reliability, although the uncalibrated model BSS is higher (-0.18 versus -0.4). With calibration, we expect the predictions to have similar reliability characteristics as their target datasets. As expected, the severe LSR calibrated predictions exhibit near perfect reliability up to 45%, and the severe PP calibrated predictions are comparable with the PP output. Both calibrated predictions feature higher BSSs than the uncalibrated dataset, UH proxy, and SPC outlook, although the PP output displays the highest score (0.17). At the sig-severe threshold (Fig. 6b), a high bias persists with the UH proxy and uncalibrated dataset, while the LSR calibrated output displays near perfect reliability up to about 15% before overforecasting. Additionally, the LSR calibrated model output shows a higher BSS (0.0) than the uncalibrated output (-0.02) and UH proxy (-0.73). The sig-severe PP calibrated probabilities are sufficiently low that they do not appear, but achieve a BSS of 0.01. The 1200 UTC SPC sig-severe outlook is binary and therefore does not appear on the probabilistic reliability diagram. In addition to changes in forecasting bias, calibration decreases output probability magnitudes, as the LSR calibrated model does not exceed 45% and 20% at the severe and sig-severe thresholds, respectively. The PP calibrated model predictions do not exceed 20% and 1% at the severe and sig-severe thresholds. Overall, calibration reliably maps the uncalibrated predictions toward two different target datasets and increases the BSS.

In Fig. 7, the LSR dataset serves as observations for calculating ETS and bias. Of the severe hail predictions, the LSR calibrated model outputs maxima and minima in ETS at similar forecast probabilities as the optimal PP output and 1200 UTC SPC outlook (Fig. 7a). However,

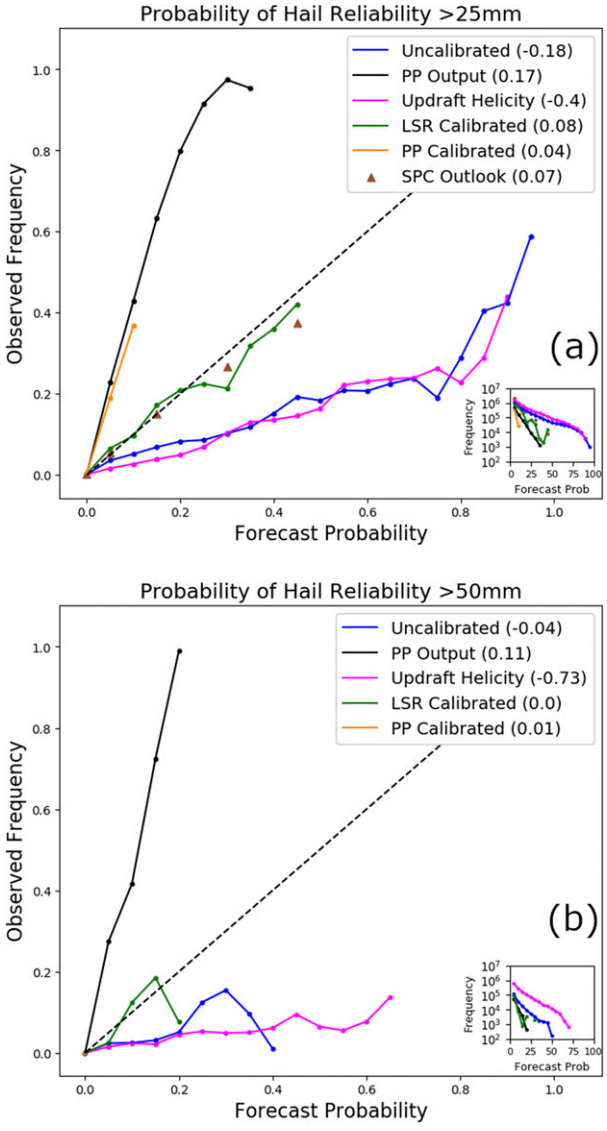

FIG. 6. Reliability diagrams of the full period (forecast hours 12–36) predictions. Data plotted include the ML predictions, SPC outlook, practically perfect output, and UH proxy over the isotonic regression test dataset. Predictions are verified at the (a) severe and (b) significant severe hail thresholds. The SPC outlook at the significant severe threshold is binary and therefore not plotted. Included in the legend is the Brier skill score for each prediction.

the PP output achieves a higher skill score (0.32), compared to the LSR calibrated (0.12). The PP calibrated probabilities also achieve an ETS value of 0.12, but at a lower forecast probability (around 5%). Both the uncalibrated output and UH proxy achieve higher skill than the PP output past 20% forecast probability, however the datasets also provide the highest bias (Fig. 6b) with the uncalibrated model outputting lower bias below 70%. Although the calibrated models produce high biases at low forecast probabilities, the biases are lower than the SPC outlook across all forecast probabilities. Of all the severe hail predictions, the LSR calibrated model best resembles the PP output and SPC hail outlook in terms of skill.

Calculations of ETS and bias, again compared to the LSRs, at the sig-severe threshold show that the uncalibrated model outputs the highest ETS skill (0.03) of the ML predictions, while maintaining overall lower bias than the UH proxy (Figs. 7c,d). Both calibrated predictions have zero skill above 20%, however the LSR calibrated model bias is most comparable to the PP output. As with the reliability diagram, the binary sig-severe SPC outlook does not appear on the probabilistic diagram. The UH field displays the greatest ETS skill above 20%, but maintains the highest bias. Similar to case studies, the uncalibrated ML model performs better at the sig-severe threshold. The differences in bias between the uncalibrated output and UH proxy, at both size thresholds, indicates that the ML algorithm better focuses on hail threat regions before calibration. At the sig-severe threshold, the uncalibrated model displays the greatest skill of the different ML predictions, however the LSR calibrated model produces bias values closer to the PP output.

In addition to the LSRs as observations, ETS and bias are calculated using the PP dataset to verify the hail predictions against a probabilistic field (Fig. 8). The severe PP calibrated prediction achieves similar ETS skill as the SPC hail outlook at 5%, however decreases in ETS and bias values above the 5% probability threshold (Fig. 8a). The severe LSR calibrated model produces the highest ETS skill below 30%. Above 30% the SPC outlook achieves a higher score. For bias, the LSR calibrated model predicts lower values than the SPC hail outlook across all probabilities (Fig. 8b). The severe hail uncalibrated model and UH proxy display comparable skill, but again the uncalibrated output achieves lower bias. Verification of the sig-severe hail predictions show that the UH proxy achieves the highest ETS value (0.04) (Fig. 8c), but continues to produce the highest bias of all the datasets, similar to previous examinations (Fig. 8d). At the sig-severe threshold, the uncalibrated model outputs the greatest ETS skill of the ML predictions, but the LSR calibrated model produces lower bias values. Overall, verifying against the PP dataset indicates that the severe LSR calibrated model produces similar skill as the SPC hail outlook, but calibration decreases ETS skill at the sig-severe threshold. Despite decreases in ETS skill, calibration lowers bias values at both size thresholds, compared to the PP dataset.

Due to the subjective success of the LSR calibrated model from the previous verification metrics, model timing information is explored. Only the severe hail

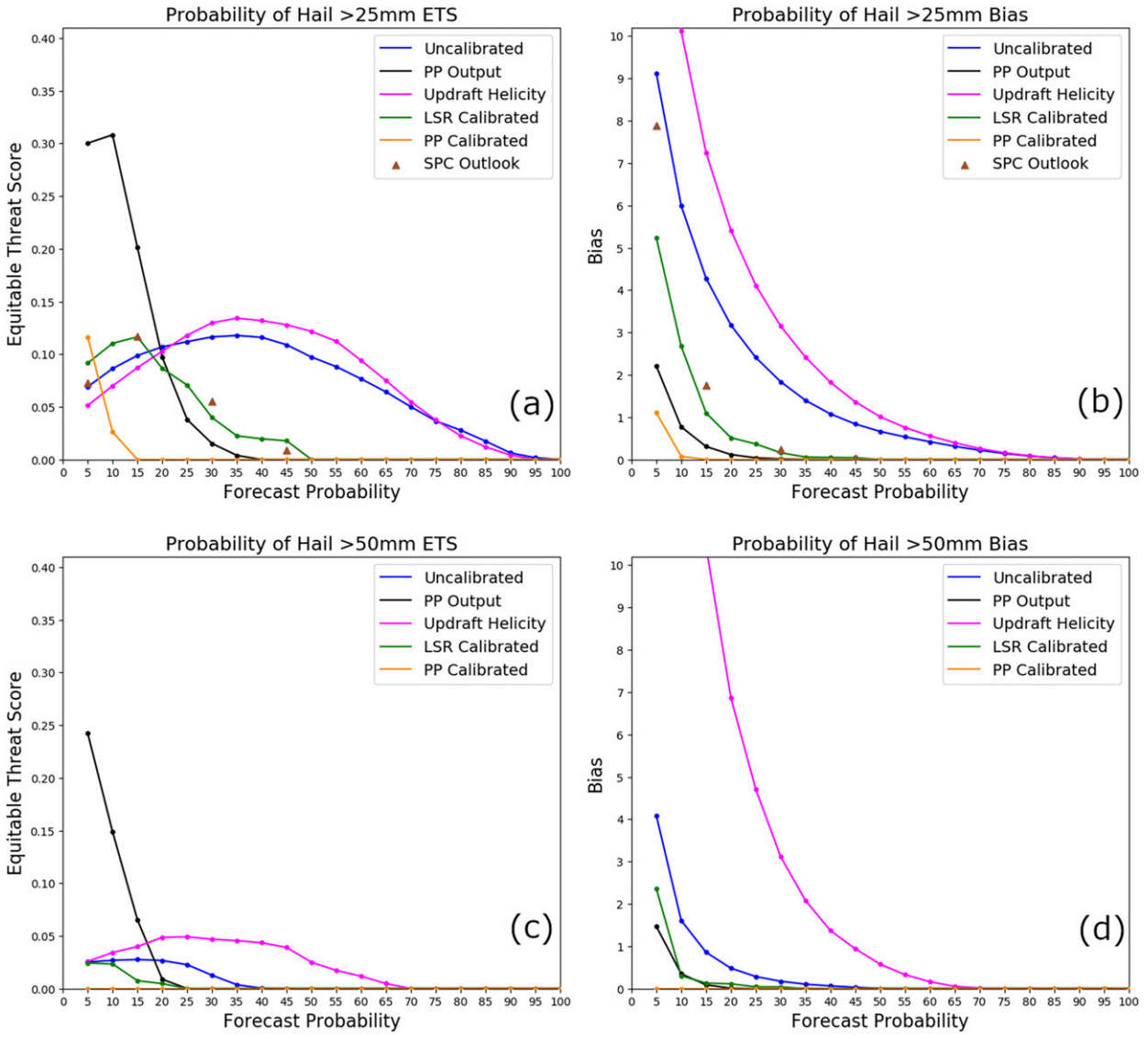

FIG. 7. ETS and bias of the full period (forecast hours 12–36) predictions. Data plotted include the ML predictions, SPC outlook, practically perfect output, and UH proxy, with local storm reports as observations. ETS and bias are calculated at the (a),(b) severe and (c),(d) significant severe hail thresholds. The SPC outlook at the significant severe threshold is binary, and therefore not plotted.

calibrated predictions are examined over the different forecast periods (1700–2100, 1900–2300, 2100–0100, and 1200–1200 UTC the next day) because of the relative rarity of sig-severe hail over short time periods. A similar analysis at the sig-severe threshold could be enlightening, however a much larger dataset would be needed. The calibrated predictions evaluated using ETS against the LSRs (Fig. 9a) and PP output (Fig. 9c) show that the 2100–0100 UTC and 24-h forecast periods achieve the highest ETS values of all the periods. Additionally, bias calculations when compared to the LSRs (Fig. 9b) and PP dataset (Fig. 9d) show similar values among the different time periods, although the 1900–2300 UTC prediction against the PP dataset shows lower bias at higher probabilities. In general, the later time periods outperform the earlier periods, but large biases are apparent across all time periods regardless of the observational dataset.

4. Discussion and summary

A random forest (RF) machine learning (ML) method for day-ahead hail prediction, based on that of G17, predicts severe hail probabilities with data from HREFv2 numerical model forecasts and observations from the Maximum Expected Size of Hail (MESH) dataset.

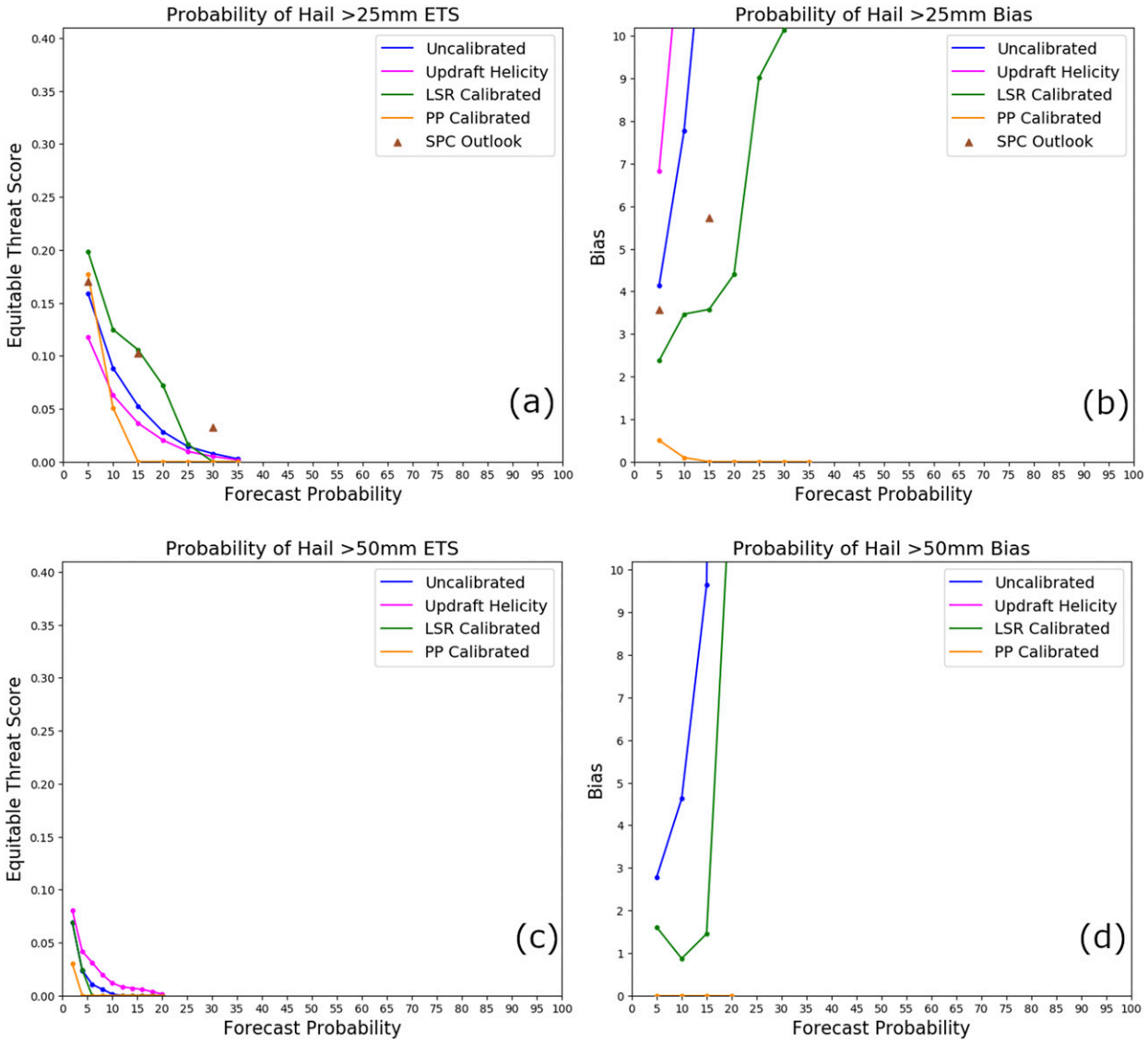

FIG. 8. As in Fig. 7, but where ETS and bias are calculated with practically perfect probabilities as observations. ETS and bias are evaluated at the (a),(b) severe and (c),(d) significant severe hail thresholds.

An added isotonic regression (IR) model calibrates the RF severe hail predictions toward the local storm reports (LSRs) and SPC's practically perfect (PP) output. The resulting hail predictions, including an uncalibrated ML model as well as the LSR and PP calibrated models, were verified through two case studies, reliability diagrams, Brier skill score (BSS), and plots of equitable threat score (ETS) and bias.

The marginal and high-end severe hail studies used to evaluate the HREFv2 ML hail predictions indicate that the uncalibrated severe hail predictions are spatially similar to the SPC hail outlook, but exhibit a high spatial bias compared to the severe PP dataset. The severe uncalibrated model also produces a high magnitude bias compared to both verification datasets. The significant severe (sig-severe) uncalibrated case study forecasts do not exhibit as high spatial or magnitude biases, however a high magnitude bias persists over the IR test set at both hail size thresholds, evidenced by the reliability diagrams and plots of ETS and bias.

The high magnitude bias (and spatial bias compared to the severe PP dataset) may be due in part from the different training and verification datasets used, where MESH observations are more numerous than LSRs. Verifying against MESH may decrease the uncalibrated ML model bias, however the LSRs are a common operational verification dataset. Beyond verification datasets, the member classification thresholds for "filtering

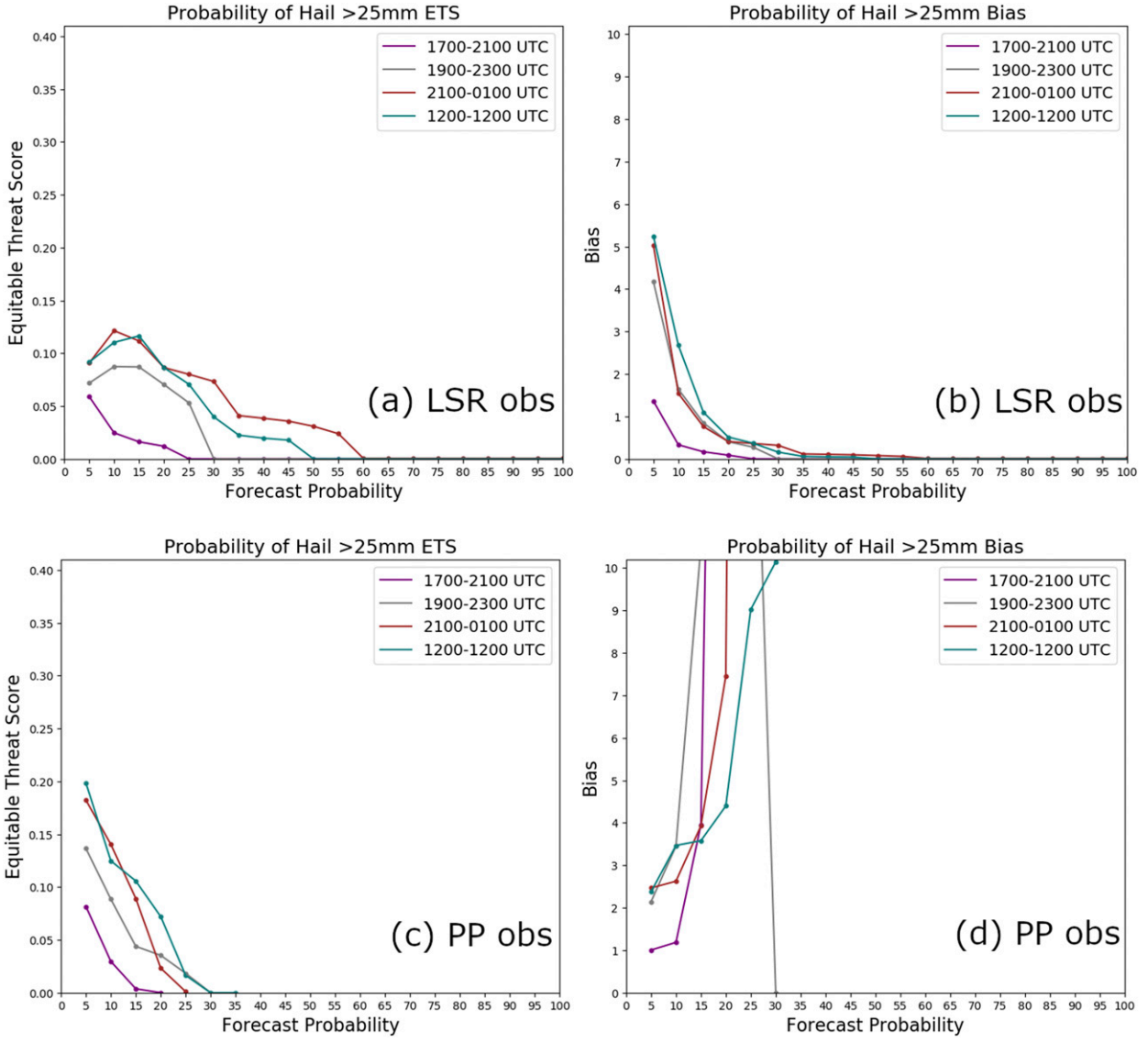

FIG. 9. Plots of ETS and bias of the LSR calibrated output at varying forecast periods. Predictions are verified at the severe hail threshold using the (a),(b) local storm reports and separately (c),(d) practically perfect probabilities as observations.

out'' storms as hail producing may contribute to the high size bias. A lower threshold classifies more storms as hail producing, leading to more nonzero hail sizes output from the RF regression model. Changing the skill metric, such as BSS, that determines the filtering thresholds may increase the thresholds. In addition, training both the classification and regressions model on a larger dataset could increase the filtering threshold, while also decreasing the scale parameter errors previously mentioned. Finally, changing the object-tracking algorithm thresholds, like increasing the MAXUVV and MESH thresholds, or decreasing the areal intensity used for creating storm tracks could decrease the number and spatial extent of predicted storm tracks, however this may also lead to greater misses. Even though the uncalibrated ML model exhibits a high magnitude bias, the severe hail proxy variable (updraft helicity) displays higher bias values and a lower BSS than the uncalibrated output at both size thresholds.

Adding calibration reduces the high magnitude bias associated with the uncalibrated ML severe hail predictions, across both calibrated datasets. For both case studies, the LSR calibrated model most closely resembles the operational day-1 SPC hail outlook and PP output at the severe hail threshold. However, the sig-severe uncalibrated ML model outputs similar probabilities, both in magnitude and spatially, as the two verification datasets. Calibrating the sig-severe ML

model decreases the probability magnitudes below the observed PP output. Although the qualitative case studies display decreases in calibrated sig-severe probabilities, the reliability diagrams calculated over the IR test set show that calibration can reliably map the RF hail predictions toward either the PP or LSR data, and increase BSS, at both size thresholds. The LSR calibrated model achieves a higher BSS, and is more reliable at certain forecast probabilities, than the 1200 UTC SPC outlook, which has access to 12 h more data than the HREFv2 model runs.

Increases in reliability are important when forecasters consider using automated guidance (Hoffman et al. 2013; Karstens et al. 2018). Indicating to forecasters that the LSR calibrated dataset is reliable compared to the LSRs, a common operational verification dataset, could increase trust in the guidance. In addition to increased reliability, the severe LSR calibrated predictions exhibit lower bias than the SPC outlook, while producing higher ETS values across most forecast probabilities, against both the PP dataset and LSRs. Although both the uncalibrated ML output and UH proxy variable produce higher biases than PP dataset or SPC outlook, calibrating the UH field toward the LSRs would not produce the same results as the LSR calibrated ML model. Calibrating the UH field would only be appropriate in environments where hail results from strongly rotating storms. Conversely, all storms modes can be skillfully predicted by the RFs because of the multiple different storm and environment fields input to the models.

In addition to day-ahead forecast skill, ETS and bias calculations verify timing information from the LSR calibrated model against both the LSR and PP dataset. The 1200–1200 and 2100–0100 UTC time periods displayed the highest skill across both verification datasets. The poorer performance at early times likely results from the higher prevalence of severe weather, including severe hail, during the afternoon and evening hours. The decrease in performance may indicate that the LSR calibrated ML model exhibits an overforecasting bias during early time periods.

Similar to the LSR calibrated model, the PP calibrated model decreased the uncalibrated model output bias. However, the PP calibrated predictions were extremely conservative in predicting both hail size thresholds. The low forecasting bias likely results from the PP dataset underforecasting bias. The Gaussian smoother applied to the LSR locations smooths the probabilistic dataset, causing the PP dataset to underforecast the LSRs by design. Calibration further decreases the overall probability magnitudes of the RF ML hail predictions, leading to an enhanced low bias with the PP calibrated model. Comparatively, the LSR calibrated model calibrates directly to the LSRs, similar to the PP output, and therefore does not include a pre-existing bias. The lack of underforecasting bias associated with the LSR calibrated target dataset most likely causes the LSR calibrated output to better resemble the PP output, than the PP calibrated. The low forecast skill and reliability of the PP calibrated model, compared to the LSR calibrated model, indicates that the LSR calibrated model may be more trusted in an operational setting.

Beyond reliability, model interpretation can be important in increasing forecaster acceptance of automation (McGovern et al. 2019). Burke (2019), which uses two separate RF regression models to predict the shape and scale parameters, investigated permutation variable importance (Lakshmanan et al. 2015) for interpretation of the ML-based hail prediction algorithm. The regression method differences could influence the produced variable importance ranks, therefore further exploration applied to the single RF regression model method is needed.

Overall, the LSR calibrated ML-based hail guidance using the HREFv2 dataset provides increased reliability and skill over the uncalibrated ML output and SPC hail outlook, indicating that the forecasts may provide operationally useful predictions. All of the ML models overforecast hail in terms of spatial extent, but this could benefit forecasters preparing forecasts in terms of highlighting potential hail risk areas and addressing storm placement uncertainty. The severe hail predictions were particularly skillful after calibration, which is encouraging for this initial application of the technique to the HREFv2 dataset. Preliminary results indicate that calibrating the ML NMEP forecasts before smoothing produces predictions with decreased spatial bias and false alarms. The next iteration of calibrated severe hail forecasts will include this step to further increase the skill of the LSR calibrated output. Even without this update, the calibrated predictions produce output that, compared to the uncalibrated predictions, are more comparable to the PP probabilities while outperforming the SPC hail outlook. However, any improvements of the calibrated models over the SPC hail outlook may be within the ranges of forecast uncertainty. Subjective ratings from the 2019 HWT SFE indicate that the LSR calibrated hail guidance performs similarly to a SPC hail guidance product, with a slightly higher mean score. In general, producing reliable forecasts with comparable skill as SPC hail outlooks has the potential to increase operational forecaster trust in the LSR calibrated automated predictions.

Acknowledgments. This work was primarily supported by the Joint Technology Transfer Initiative (JTTI) Grant

NA16OAR4590239 provided by NOAA, and supplemented with funding provided by NOAA JTTI Grant NA18OAR4590371. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. Computing was primarily executed on the University of Texas Advanced Computing Center (TACC) Stampede supercomputer. The authors thank Timothy Supinie and Chris Cook for data support, Ryan Lagerquist for implementation feedback, and Jonathon Labriola for contribution of ideas. We also thank Steven Weiss, Israel Jirak, and the participants of the 2018 HWT SFE for useful assessments and remarks. We thank David Harrison and Christopher Karstens for providing the gridded SPC outlook data. Finally, we thank the anonymous reviewers for their feedback, which helped improve this manuscript.

REFERENCES

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1.

Aligo, E., B. Ferrier, J. Carley, E. Rogers, M. Pyle, S. Weiss, and I. Jirak, 2014: Modified microphysics for use in hifgh-resolution NAM forecasts. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 16A.1, https://ams.confex.com/ams/27SLS/webprogram/Paper255732.html.

Bosart, L. F., 1989: Automation: Has its time really come? *Wea. Forecasting*, **4**, 271–271, https://doi.org/10.1175/1520-0434(1989)004<0271:AHITRC>2.0.CO;2.

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

——, and P. Spector, 1992: Submodel selection and evaluation in regression. The X-random case. *Int. Stat. Rev.*, **60**, 291–319, https://doi.org/10.2307/1403680.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Burke, A., 2019: Using machine learning applications and HREFv2 to enhance hail prediction for operations. M.S. thesis, School of Meteorology, University of Oklahoma, 102 pp., https://hdl.handle.net/11244/320425.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous United States. *Wea. Forecasting*, **27**, 1235–1248, https://doi.org/10.1175/WAF-D-11-00151.1.

Clark, A. J., W. A. Gallus, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

——, and Coauthors, 2016: Preliminary findings and results—Spring forecasting experiment 2016.NOAA, 50 pp., https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE_2016_preliminary_findings_final.pdf.

Davis, C., and F. Carr, 2000: Summary of the 1998 workshop on mesoscale model verification. *Bull. Amer. Meteor. Soc.*, **81**, 809–820, https://doi.org/10.1175/1520-0477(2000)081<0809:SOTWOM>2.3.CO;2.

Durran, D. R., and J. A. Weyn, 2016: Thunderstorms do not get butterflies. *Bull. Amer. Meteor. Soc.*, **97**, 237–243, https://doi.org/10.1175/BAMS-D-15-00070.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Gagne, D. J., 2016: Coupling data science techniques and numerical weather prediction models for high-impact weather prediction. Ph.D. thesis, University of Oklahoma, 204 pp., https://shareok.org/handle/11244/44917.

——, 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. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

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.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, https://doi.org/10.1175/2007MWR2410.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

Herman, G. R., and R. S. Schumacher, 2018a: 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.

——, and ——, 2018b: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

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

Hoffman, R. R., M. Johnson, J. M. Bradshaw, and A. Underbrink, 2013: Trust in automation. *IEEE Intell. Syst.*, **28**, 84–88, https://doi.org/10.1109/MIS.2013.24.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

——, K.-S. S. Lim, Y.-H. Lee, J.-C. Ha, H.-W. Kim, S.-J. Ham, and J. Dudhia, 2010: Evaluation of the WRF double-moment 6-class microphysics scheme for precipitating convection. *Adv. Meteor.*, **2010**, 1–10, https://doi.org/10.1155/2010/707253.

Janjić, Z. I., 1990: The step-mountain coordinate: Physical package. *Mon. Wea. Rev.*, **118**, 1429–1443, https://doi.org/10.1175/1520-0493(1990)118<1429:TSMCPP>2.0.CO;2.

——, 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta Hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592–1609, https://doi.org/10.1175/2009WAF2222230.1.

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 Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., P9.137, https://ams.confex.com/ams/26SLS/webprogram/Manuscript/Paper211729/2012_SLS_SSEO_exabs_Jirak_final.pdf.

——, 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 Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

Johns, R. H., and C. A. Doswell, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612, https://doi.org/10.1175/1520-0434(1992)007<0588:SLSF>2.0.CO;2.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010a: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

——, and Coauthors, 2010b: Assessing advances in the assimilation of radar data and other mesoscale observations within a collaborative forecasting–research environment. *Wea. Forecasting*, **25**, 1510–1521, https://doi.org/10.1175/2010WAF2222405.1.

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.

——, 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.

Labriola, J., N. Snook, Y. Jung, B. Putnam, and M. Xue, 2017: Ensemble hail prediction for the storms of 10 May 2010 in south-central Oklahoma using single- and double-moment microphysical schemes. *Mon. Wea. Rev.*, **145**, 4911–4936, https://doi.org/10.1175/MWR-D-17-0039.1.

——, ——, ——, and M. Xue, 2019: Explicit ensemble prediction of hail in 19 May 2013 Oklahoma City thunderstorms and analysis of hail growth processes with several multimoment microphysics schemes. *Mon. Wea. Rev.*, **147**, 1193–1213, https://doi.org/10.1175/MWR-D-18-0266.1.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol.*, **32**, 1209–1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

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. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

——, R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 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.

Melick, C. J., I. L. Jirak, J. Correia, A. R. Dean, and S. J. Weiss, 2014: Exploration of the NSSL Maximum Expected Size of Hail (MESH) product for verifying experimental hail forecasts in the 2014 Spring Forecasting Experiment. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 76, https://ams.confex.com/ams/27SLS/webprogram/Paper254292.html.

Moller, A. R., C. A. Doswell, M. P. Foster, and G. R. Woodall, 1994: The operational recognition of supercell thunderstorm environments and storm structures. *Wea. Forecasting*, **9**, 327–347, https://doi.org/10.1175/1520-0434(1994)009<0327:TOROST>2.0.CO;2.

Murillo, E., and C. Homeyer, 2019: Severe hail fall and hail storm detection using remote sensing observations. *J. Appl. Meteor. Climatol.*, **58**, 947–970, https://doi.org/10.1175/JAMC-D-18-0247.1.

Niculescu-Mizil, A., and R. Caruana, 2005: Predicting good probabilities with supervised learning. *Proc. 22nd Int. Conf. on Machine Learning*, Bonn, Germany, ACM, 625–632, https://doi.org/10.1145/1102351.1102430.

Ortega, K., 2018: Evaluating multi-radar, multi-sensor products for surface hailfall diagnosis. *Electron. J. Severe Storms Meteor.*, **13** (1), http://ejssm.org/ojs/index.php/ejssm/article/view/163.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Schaefer, J. T., J. J. Levit, S. J. Weiss, and D. W. McCarthy, 2004: The frequency of large hail over the contiguous United States. *14th Conf. on Applied Climatology*, Seattle, WA, Amer. Meteor. Soc., 3.3, https://ams.confex.com/ams/pdfpapers/69834.pdf.

Schwartz, C., and R. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

Skamarock, W. C., and J. B. Klemp, 2008: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J. Comput. Phys.*, **227**, 3465–3485, https://doi.org/10.1016/j.jcp.2007.01.037.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Snellman, L. W., 1977: Operational forecasting using automated guidance. *Bull. Amer. Meteor. Soc.*, **58**, 1036–1044, https://doi.org/10.1175/1520-0477(1977)058<1036:OFUAG>2.0.CO;2.

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. *Wea. Forecasting*, **31**, 811–825, https://doi.org/10.1175/WAF-D-15-0152.1.

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

Svaldi, A., 2018: Damage from last year's massive front range hail storm cost $2.3 billion—$900 million more than first estimated. *The Denver Post*, 7 May 2018, https://www.denverpost.com/2018/05/07/2017-front-range-hail-storm-damage-cost/.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Wilson, C. J., K. Ortega, and V. Lakshmanan, 2009: Evaluating multi-radar, multi-sensor hail diagnosis with high resolution hail reports. *25th Conf. on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., P2.9, https://ams.confex.com/ams/89annual/techprogram/paper_146206.htm.

Wilson, K. A., P. L. Heinselman, C. M. Kuster, D. M. Kingfield, and Z. Kang, 2017: Forecaster performance and workload: Does radar update time matter? *Wea. Forecasting*, **32**, 253–274, https://doi.org/10.1175/WAF-D-16-0157.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88d. *Wea. Forecasting*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, https://doi.org/10.1175/2011BAMS-D-11-00047.1.