Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles

DAVID JOHN GAGNE II[a]

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

AMY MCGOVERN

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

SUE ELLEN HAUPT AND RYAN A. SOBASH

National Center for Atmospheric Research, Boulder, Colorado

JOHN K. WILLIAMS

The Weather Company, Andover, Massachusetts

MING XUE

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

ABSTRACT

Forecasting severe hail accurately requires predicting how well atmospheric conditions support the development of thunderstorms, the growth of large hail, and the minimal loss of hail mass to melting before reaching the surface. Existing hail forecasting techniques incorporate information about these processes from proximity soundings and numerical weather prediction models, but they make many simplifying assumptions, are sensitive to differences in numerical model configuration, and are often not calibrated to observations. In this paper a storm-based probabilistic machine learning hail forecasting method is developed to overcome the deficiencies of existing methods. An object identification and tracking algorithm locates potential hailstorms in convection-allowing model output and gridded radar data. Forecast storms are matched with observed storms to determine hail occurrence and the parameters of the radar-estimated hail size distribution. The database of forecast storms contains information about storm properties and the conditions of the prestorm environment. Machine learning models are used to synthesize that information to predict the probability of a storm producing hail and the radar-estimated hail size distribution parameters for each forecast storm. Forecasts from the machine learning models are produced using two convection-allowing ensemble systems and the results are compared to other hail forecasting methods. The machine learning forecasts have a higher critical success index (CSI) at most probability thresholds and greater reliability for predicting both severe and significant hail.

1. Introduction

Hail causes billions of dollars in property and crop damage in the United States each year (Changnon 2009) and is a major liability for insurance companies (Brown et al. 2015) with $850 million in average annual claims. Urban sprawl and population growth in large cities such as Dallas/Fort Worth, Texas; St. Louis, Missouri; Chicago, Illinois; and Denver, Colorado, have made large amounts of property damage from hail events more likely (Rosencrants and Ashley 2015). Hail growth and melting is governed by the microphysical processes that a hailstone experiences along its trajectory through a storm. Foote (1984) and Nelson (1983) found that the amount of hail growth strongly depends on the path

---

[a] Current affiliation: National Center for Atmospheric Research, Boulder, Colorado.

Corresponding author: David John Gagne II, dgagne@ucar.edu

through the updraft and the amount of supercooled liquid water encountered. Rasmussen and Heymsfield (1987) identified how differences in ice density, relative humidity, and hail diameter can affect the rate of melting. However, hail forecasting methods have relied on a combination of coarse approximations of the expected storm environment and local hail climatology with varying degrees of success (Johns and Doswell 1992). Most existing hail forecast methods (e.g., Fawbush and Miller 1953; Moore and Pino 1990; Brimelow et al. 2002) are based on information extracted from a sounding representative of the convective environment. All of these models estimate potential updraft strength from the integrated buoyancy between the melting layer and the equilibrium level. To determine the hail size at the ground, the hail size models also include a melting component based on the height where the wet-bulb temperature reaches 0°C, which is approximately where melting would begin in hail falling through downdraft air. Shallower, cooler, and drier melting layers lead to less melting (Rasmussen and Heymsfield 1987), and larger hailstones have higher terminal fall speeds, leading to shorter transit times (Johns and Doswell 1992).

Sounding-based hail forecasting has fundamental limitations. For best results, the sounding should be representative of the storm environment, which can be problematic when observed soundings are only available at discrete locations and twice a day. This issue can be addressed by correcting the sounding based on surface observations (Moore and Pino 1990) or by utilizing model soundings near the time and location of expected hail, but this may not be representative of the updrafts that actually produce hail. Also, updraft speed from CAPE is estimated at the top of the updraft, not within the hail growth region. The largest hail may not come from the strongest part of the updraft because hail may be lofted elsewhere if the speed is too high (Foote 1984; Johns and Doswell 1992). In storms with broader updrafts, hailstones may have significant horizontal motions in their growth trajectories (Nelson 1983; Foote 1984), leading to additional growing time. Hail embryos also may not traverse the optimal growth trajectory through the updraft and thus fail to achieve the largest size possible for that updraft speed (Dennis and Kumjian 2017).

Some of the shortcomings with sounding-based hail diagnostics were previously addressed by feeding sounding information into a one-dimensional hail growth model called HAILCAST (Brimelow et al. 2002). This model creates an ensemble of updrafts based on perturbations of the temperature and moisture profile. The energy shear index estimates the updraft lifetime and the amount of entrainment. Hail embryos are released into the updraft and grow until they can no longer be sustained or the updraft collapses. A bulk melting scheme is then applied to approximate the hail size at the ground. Jewell and Brimelow (2009) found that HAILCAST generally provides a reliable hail size forecast, especially when compared to other sounding methods. Adams-Selin and Ziegler (2016) implemented an updated version of HAILCAST with many physical improvements based on observational hail studies in the WRF Model. WRF-HAILCAST, using modeled cloud liquid and ice water and vertical velocities in place of those estimated from a sounding and in spatial verification experiments, was found to forecast hail sizes within 12 mm of observed 66% of the time (Adams-Selin and Ziegler 2016). Although WRF-HAILCAST does incorporate many physical processes into its hail growth model, it does have to make some simplifying assumptions about hail growth trajectories to be computationally efficient. It is also sensitive to the sizes of the initial hail embryos and the density of the hailstones (Adams-Selin and Ziegler 2016), but that sensitivity is addressed by growing an ensemble of hail embryos with varied initial sizes.

As computational speed and storage capacities have increased in the past 20 years, research and operational groups have run real-time NWP models capable of explicitly representing deep, moist convection (Kain et al. 2008). While not all convective processes are adequately resolved, individual updrafts and their associated heat and moisture fluxes can be represented reasonably well. Because some convective processes are not fully resolved, these models are called convection-allowing models (CAMs). CAMs have shown skill over mesoscale NWP models in precipitation forecasting and in forecasting convective evolution and morphology (Weisman et al. 2008). Storms generally occur over a small subset of an area with favorable conditions, so explicit forecasts of convection help constrain the forecast severe threat area. However, errors in the initial conditions and model physics lead to spatial errors in storm placement and timing errors in convective initiation and evolution. To account for this uncertainty, CAM ensembles perturb the initial and boundary conditions, physics parameterizations, and the dynamical cores to capture the range of possible convective solutions. CAM ensembles have run in real time as part of NOAA's Hazardous Weather Testbed Spring Experiment since 2007 (Clark et al. 2012). The National Weather Service is already running deterministic CAMs and will be deploying ensembles soon (Benjamin 2014).

While CAMs cannot resolve severe hazards, such as tornadoes, hail, and high winds, storm diagnostics have shown skill in inferring hazard potential. Because storms

develop and intensify in minutes, hourly snapshots of CAM output may miss the time when the storm is at its greatest intensity. More frequent model output may capture these extremes, but this would require more computational resources. A compromise is to track the maximum value at a point over a given time period and regularly output that field. These hourly maximum fields provide a surrogate for both storm track and intensity (Kain et al. 2010). For severe weather forecasting, updraft helicity (UH), the integrated product of vertical velocity and vertical vorticity between 2 and 5 km, is a proxy for strong, rotating updrafts (Kain et al. 2008) and correlates with tornado pathlength (Clark et al. 2013). Other hourly maximum fields include updraft and downdraft speeds, simulated radar reflectivity at the −10°C level and column total graupel (CTG) mass. Storm surrogate products have some limitations. Current storm surrogates do not directly predict the occurrence of severe weather hazards. The intensity distributions of the storm surrogates are dependent on the dynamical core, grid spacing, and parameterization scheme (Kain et al. 2008), which make them more challenging for forecasters to interpret when examining mixed ensembles. Most storm surrogates are indirectly verifiable through storm reports. A forecast product that directly forecasts the probability of a severe hazard is most useful if the predictions are well calibrated and physically consistent with the model forecast.

Machine learning (ML) models integrate data from multiple sources together to produce calibrated hazard predictions. Unlike physics-based models, which derive prognostic and diagnostic equations from primitive equations, a set of assumptions, and initial and boundary conditions, ML models identify patterns in a set of data and then find a relationship between those patterns and an outcome that optimizes an error metric. In meteorology, ML and statistical models are often used to link observations or NWP model fields to hazards and other events of importance that are not resolved or described well by NWP models or our current observing systems. Examples of ML for hazard identification include storm classification (Lakshmanan et al. 2010; Gagne et al. 2009), convection initiation (Ahijevych et al. 2016), aircraft turbulence (Williams 2014), and hurricane power outages (Nateghi et al. 2014). In the hail prediction domain, Manzato (2013) used neural network ensembles to predict hail occurrence and spatial extent in Italy using sounding parameters as input, and Marzban and Witt (2001) trained a neural network on radar parameters to nowcast hail size categories. ML models trained on observed data generally can only make accurate predictions out to a few hours ahead, but when given NWP model output within a model output

statistics framework (Glahn and Lowry 1972), calibrated predictions for one or more days in advance are possible. ML and statistical models generally require a large dataset of historical NWP model runs and observations to build an accurate predictive model. These historical model runs should have similar configurations so that the ML model can identify systematic patterns and biases. Domain knowledge is still necessary to identify a reasonable set of relevant input variables for the ML models and to transform the output variables between forms that are easier for the ML models to fit and forms that are interpretable for end users.

The purpose of this paper is to describe track-based day-ahead probabilistic hail forecasts generated from ML models. The ML hail forecasts are compared with hail-related storm surrogate variables (Sobash et al. 2011), such as UH and CTG, and with HAILCAST. This evaluation tests the hypotheses that ML models will produce accurate and reliable hail forecasts, that the ML forecasts can detect more hailstorms and produce fewer false alarms than other hail size methods, and that ML model performance is more consistent across different NWP configurations than other hail size diagnostics. This work builds upon the hail size regression model described in Gagne et al. (2015), which predicted a storm's maximum hail size but could not predict hail larger than 50 mm. The ML regression model in this paper predicts the parametric distribution parameters of the radar-estimated maximum expected size of hail over the area of a storm instead of predicting the maximum hail size within the storm. Predicting the distribution of hail sizes instead of the most likely size value allows us to estimate the probability of hail exceeding any size threshold consistently with one model. The predicted spatial hail size distribution is mapped to the areas of forecast storms in a manner akin to the probability-matched mean (Ebert 2001) in order to produce spatially realistic deterministic gridded hail size forecasts that can be compared directly with other hail forecast products from CAMs. This paper is adapted from the dissertation of Gagne (2016), and a summary description of the ML hail forecast model appears in McGovern et al. (2017).

2. Methods

a. Data

The ML hail forecast method was evaluated on two CAM ensemble systems to demonstrate the robustness of the hail forecast method to differences in member configurations, data assimilation choices, training periods, and input variables. The first comes from the

TABLE 1. CAPS ensemble member physics parameterizations for 2014 and 2015. The PBL schemes tested are the MYJ (Mellor and Yamada 1982), Yonsei University (YSU; Hong et al. 2006), Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004), and Quasi-Normal Scale Elimination (QNSE; Sukoriansky et al. 2005).

Member	2014 microphysics	2015 microphysics	2014 PBL	2015 PBL
CN	Thompson	Thompson	MYJ	MYJ
M3	Morrison	P3	YSU	MYNN
M4	Thompson	MY	QNSE	YSU
M5	Morrison	Morrison	MYNN	MYNN
M6	MY	MY	MYJ	MYJ
M7	WDM6	P3	YSU	YSU
M8	WDM6	P3	QNSE	MYJ
M9	MY	MY	MYNN	MYNN
M10	Morrison	Morrison	YSU	YSU
M11	Thompson	Thompson	YSU	YSU
M12	Thompson	Thompson	MYNN	MYNN
M13	Morrison	Morrison	QNSE	MYJ

Center for Analysis and Prediction of Storms' (CAPS) Storm-Scale Ensemble Forecast system, called the CAPS ensemble, which consists of 12 WRF-ARW models with varied combinations of microphysics and planetary boundary layer (PBL) parameterizations and perturbed initial and boundary conditions. The 2014 CAPS ensemble uses 4-km horizontal grid spacing, while the 2015 CAPS ensemble uses 3-km grid spacing. The CAPS ensemble is initialized with three-dimensional variational data assimilation (3DVAR) that includes radar and begins at 0000 UTC with hourly output for 60 h. Individual member parameterization configurations are listed in Table 1. The CAPS ensemble uses the following microphysics schemes: Thompson (Thompson et al. 2004), Morrison [with graupel settings for the rimed ice species; Morrison et al. (2009)], Milbrandt and Yau two-moment (MY; Milbrandt and Yau 2005), WRF double-moment 6-class (WDM6; Lim and Hong 2010), and Predicted Particle Properties (P3; Morrison and Milbrandt 2015; Morrison et al. 2015).

The second ensemble forecast set is from the National Center for Atmospheric Research (NCAR) ensemble (Schwartz et al. 2015), which consists of ten 3-km-grid-spacing WRF members initialized from mesoscale analyses produced using the Data Assimilation Research Testbed (DART; Anderson et al. 2009) ensemble adjustment Kalman filter data assimilation system on a 15-km grid. All members use the Thompson microphysics scheme (Thompson et al. 2004) and the Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982) PBL scheme. The ensemble began running daily in April 2015 at 0000 UTC with plans to continue running through at least summer 2017. For this project, model runs from 1 May through 31 July 2015 were used.

Hail size observations are derived from the NOAA/ National Severe Storms Laboratory Multi-Radar Multi-Sensor (MRMS) radar mosaic (Zhang et al. 2011). MRMS merges radar reflectivity from multiple radars onto a $0.01° \times 0.01°$ uniform grid with 2-min updates and performs a series of quality control procedures. The MRMS maximum expected size of hail (MESH; Witt et al. 1998; Cintineo et al. 2012) product is used for the hail size observations. MESH is a power-law relationship between the severe hail index, which is a product derived from radar reflectivity values above the melting level, and hail reports from nine hail events in 1992 in Oklahoma and Florida. The MESH relationship is calibrated such that the MESH value should exceed 75% of reported hail sizes for a given severe hail index value. Because MRMS provides more complete information about the full depth of a storm than a single radar, and because MESH indirectly accounts for melting effects through melting-layer height information and calibration to hail reports, MRMS MESH shows good spatial coverage when compared with high-resolution hail reports (Cintineo et al. 2012) and is less biased by population compared to NOAA *Storm Data* hail reports. However, MESH hail sizes exhibit a large size bias compared with high-resolution hail reports (Cintineo et al. 2012) and were not designed for identifying hail sizes below 19 mm. If the distribution of MESH sizes below 19 mm is extremely biased, it would also bias the size distribution forecasts from the ML models. Even with these known issues, MRMS MESH offers the best spatial and temporal coverage, and any size or location issues with MESH are small compared to the spatial and intensity forecast errors in the 24-h NWP predictions.

b. Storm object identification, tracking, and matching

The storm data processing procedure is summarized in Fig. 1. An object-identification method locates
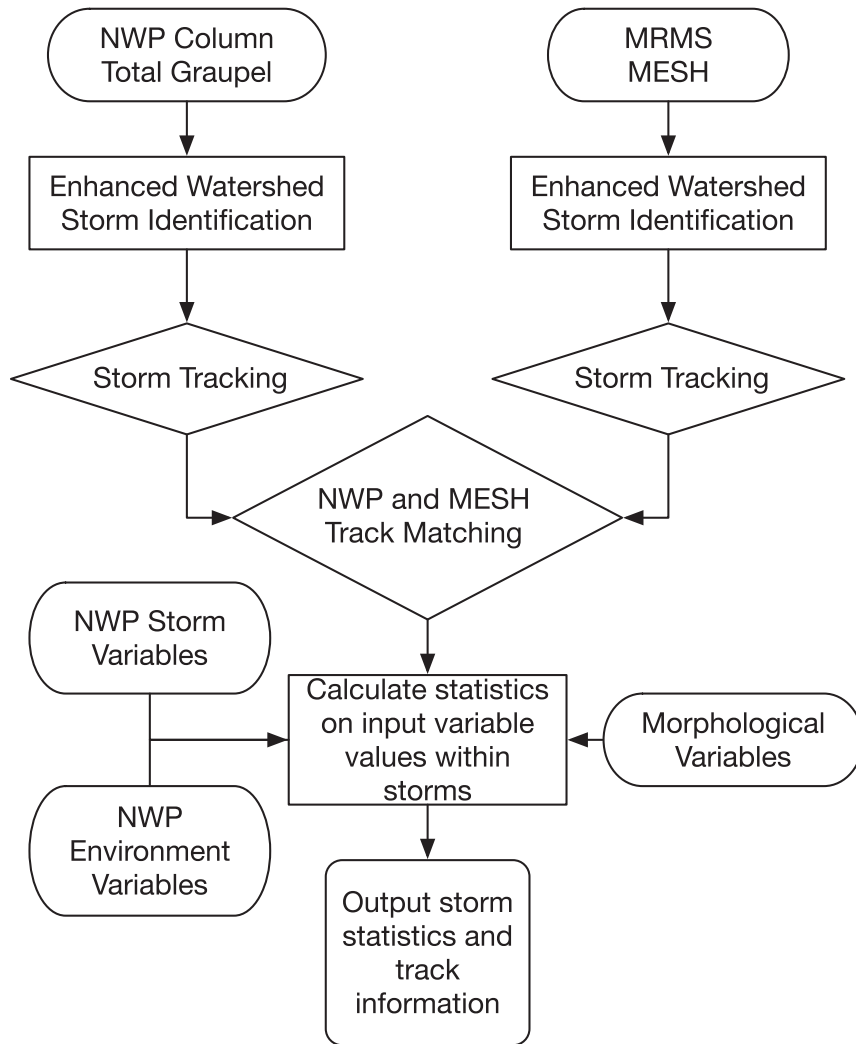
FIG. 1. Diagram of how hail forecast and observation data are preprocessed before being used in the ML model.

storms, and data extraction and forecasting processes are performed. Object-based forecasting provides the advantages of identifying relevant areas for predicting severe weather and reducing the volume of data processed. These methods exhibit parameter sensitivity, which can vary the composition of the object set. The enhanced watershed method (Lakshmanan et al. 2009) identifies potential hailstorm objects from a storm field by identifying local maxima (Fig. 2) and then growing the objects until they meet area and intensity criteria. Hourly maximum CTG mass is used as a potential hailstorm surrogate because it identifies any storm containing a large amount of graupel, including both ordinary thunderstorms and supercells. Graupel is used instead of hail because most of the CAM members' microphysics schemes feature one rimed ice species with

fixed graupel-like properties. The exceptions are the MY scheme, which has separate hail and graupel species, and the P3 scheme, which allows the particle properties to vary based on atmospheric conditions. Large CTG mass may originate from large graupel or many small graupel, but either could lead to surface hail. ML models utilize environmental information to determine the difference. A minimum intensity of $3 \, \mathrm{kg \, m^{-2}}$ captures both weak and strong storms and captures the varied distributions of graupel found with the different microphysics schemes. A minimum area of $144 \, \mathrm{km^2}$ (16 grid cells on a grid with 3-km spacing) and a maximum area of $900 \, \mathrm{km^2}$ are specified to isolate individual storm cells. The minimum area corresponds to the minimum resolvable area for a single storm cell, and the maximum area roughly corresponds to the area that a storm cell of
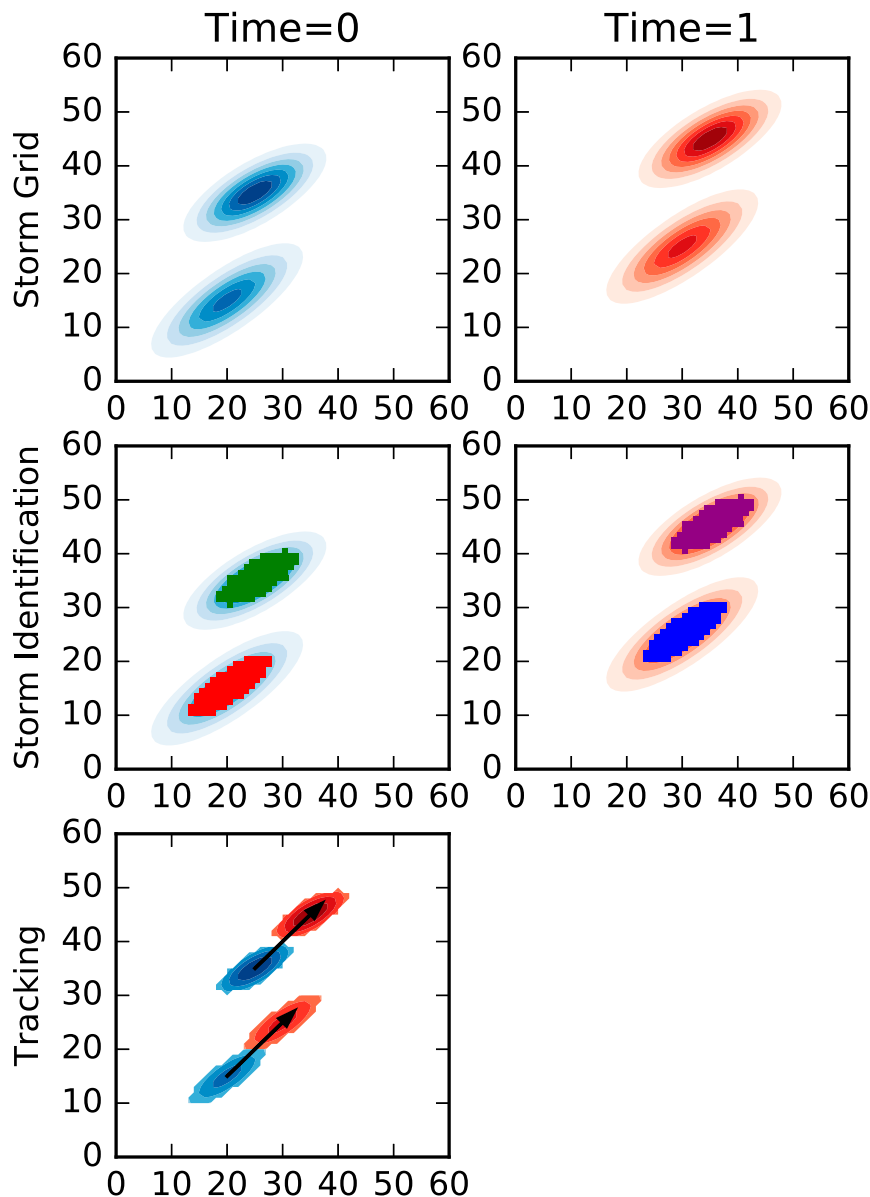
FIG. 2. An example of the enhanced watershed storm identification algorithm and the storm tracking algorithm applied to two synthetic storms. (top) Two bivariate normal distributions that have been translated northeast. (middle) The enhanced watershed determines which grid points are part of each object. (bottom) The tracking algorithm correctly links the simulated storm objects between time steps.

area 300 km^2 could traverse within an hour. Grid points outside the contiguous United States are excluded because of the lack of reliable verification data.

Once storm objects are identified in the models and observations, the objects are linked together through time into tracks. The tracking algorithm identifies trends within the storms and captures the full life cycle of each storm. The Hungarian method (Munkres 1957), a globally optimal matching algorithm, links tracks. The Hungarian method also forms the basis of the Thunderstorm Identification, Tracking, Analysis, and Nowcasting (TITAN) storm tracking algorithm (Dixon and Wiener 1993). Although Euclidean centroid distance works well with circular objects and frequent temporal updates, hourly maximum storm objects tend to be elongated along the axis of motion. To mitigate this issue, spatial cross-correlation motion estimation (Lakshmanan and Smith 2010) translates the centroid before matching (Fig. 2).

TABLE 2. Input variables for the CAPS ensemble ML models. The mean, maximum, minimum, median, standard deviation, skewness, 10th percentile, and 90th percentile of the gridpoint values within the boundaries of each storm object were calculated for the storm proxy and environment variables. CAPE is convective available potential energy, CIN is convective inhibition, ML is the mean layer, MU is the most unstable layer, SR is storm relative, and LCL is the lifting condensation level.

Storm proxy	Environment	Morphological	Location
CTG	MLCAPE	Area	Forecast hour
2–5-km UH	MLCIN	Eccentricity	Valid hour (UTC)
Reflectivity −10°C	MUCAPE	Major axis length	Current duration
Updraft speed	MUCIN	Minor axis length	Total duration
Downdraft speed	LCL height	Orientation	W–E storm motion
Echo-top height	0–3-km SR helicity	Extent	S–N storm motion
Precipitation	0–6-km wind shear		
Precipitable water	500-hPa temp		
Bunkers U	700-hPa temp		
Bunkers V	2-m dewpoint		
	2-m temp		
	850-hPa specific humidity		
	0–3-km lapse rate		
	700–500-hPa lapse rate		
	10-m U wind		
	10-m V wind		
	700-hPa U wind		
	700-hPa V wind		

The cross-correlation search is constrained by the storm dimensions to minimize motion estimation error. The maximum centroid distance for tracking is 24 km, which is empirically found to connect storms appropriately while minimizing the issue of tracks jumping to adjacent storms. Although tracking has been empirically tuned, not all connections will be correct, which adds some uncertainty to the tracking-related variables.

Forecast and observed storm tracks are matched assuming that the storms in similar locations and times should have a stronger connection. Matching tracks requires a metric that can account for both spatial and temporal differences and can compare different durations. This track matching distance is

$$D_{tm} = 0.5 \frac{d_s}{160\,\text{km}} + 0.3 \frac{t_s}{3\,\text{h}} + 0.1 \frac{t_d}{16\,\text{h}} + 0.1 \frac{a_d}{200\,\text{km}^2}. \quad (1)$$

Equation (1) contains the distance between the starting points d_s, the difference in start times t_s, the duration difference t_d, and the mean area difference a_d. Tight maximum distances, which are the denominators in (1), for start location difference (160 km) and start time difference (3 h) are used to limit the search area. Duration (16 h) and area (200 km²) break ties. The distances have units of kilometers, time and duration differences use hours, and area differences use square kilometers. If any components exceed their maximum, the track pairing is excluded from consideration. The scaling values are chosen empirically based on inspection of matches and an examination of the percentage of matched storms in the

training data. Each forecast track is paired with the closest observed track, so multiple forecast tracks can be matched with the same observed track. This approach accounts for inherent spatial and temporal errors in storm placement. The main limitation is that small changes in the weights could result in different matches. The choice of matching function parameters is inherently subjective, but some combination of space, time, and storm properties should provide reasonable results.

c. Machine learning procedure

After storm tracking and matching has occurred, other meteorological variables are extracted from each forecast storm. This approach extracts statistics about the distribution of meteorological variable values from grid cells within the extent of a storm. These statistics are the mean, standard deviation, minimum, maximum, median, 10th and 90th percentiles, and skew. The full list of input variables for the CAPS ensemble is listed in Table 2, and the full list for the NCAR ensemble is in Table 3. Storm variables are extracted from model grids at the same forecast time as the storm, and environment variables are extracted from the previous hour to sample the environment before the storm traverses that area. The CAPS and NCAR ensembles use different archive and postprocessing systems, so some variables are available in one system but not the other.

The ML hail forecast procedure is summarized in Fig. 3. The goal of the ML algorithm is to use the meteorological variables within each storm footprint to

TABLE 3. As in Table 2, but for the NCAR ensemble ML models. SB stands for surface based.

Storm proxy	Environment	Morphological	Location
CTG	SBCAPE	Area	Forecast hour
2–5-km UH	SBCIN	Eccentricity	Valid hour (UTC)
Composite reflectivity	MUCAPE	Major axis length	Current duration
Updraft speed	Precipitable water	Minor axis length	Total duration
Downdraft speed	LCL Height	Orientation	W–E storm motion
Thompson hail size max	0–3-km SR helicity	Extent	S–N storm motion
Thompson hail size surface	0–6-km wind shear		
0–3-km UH	0–1-km wind shear		
2–5-km min UH			
10-m wind speed			
0–1-km vorticity			

forecast hail occurrence and a parametric approximation of the observed spatial MESH distribution. To distinguish the ML forecast product from the output of NWP model microphysics schemes, the output of the ML models will be named as a forecast MESH distribution instead of a forecast hail size distribution. If a forecast storm track is matched with a MESH track, then hail is assumed to have occurred. The spatial MESH distribution is modeled as a gamma distribution fit to the MESH values using maximum likelihood estimation. Fitting a parametric distribution to the spatial MESH values reduces the dimensionality of the prediction problem and preserves the proportional relationships between small and large hail areas within each storm. The gamma distribution probability density function takes the following form:

$$f(x) = \frac{(x/\beta)^{\alpha-1}\exp(-x/\beta)}{\beta\Gamma(\alpha)} - x_0, \quad \alpha, \beta > 0, \quad (2)$$

where Γ is the gamma function, α is the shape parameter, x_0 is the location parameter, and β is the scale parameter. Unlike a bulk microphysics scheme, which estimates the moments of a parametric hail or graupel size distribution describing the sizes within each grid cell, this MESH distribution describes the radar-estimated hail sizes over the entire area of a storm. An example of a MESH object and the gamma distribution fit to its MESH distribution is shown in Fig. 4. The shape parameter affects the skewness, with small shape parameter values causing more skew and larger values leading to less skew. The location parameter determines the minimum value of the distribution and is fixed at 6 mm. The scale parameter adjusts the width of the gamma distribution and has the same units as the modeled quantity (Wilks 2011).

We use ML models from the Scikit-learn package version 0.16.1 (Pedregosa et al. 2011). A stacked model approach is used with a classifier model predicting whether or not hail occurs, along with a regression model that predicts the MESH distribution parameters. A random forest (RF; Breiman 2001) model is used as the hail occurrence classifier. RF is an ensemble of decision trees in which each decision tree is grown from a resampled version of the original training set, and in which a different random subset of input variables is evaluated for inclusion into the tree at each branch in the growing process. These sources of randomness ensure the independence of each tree. The RF classifier uses a configuration that weights each training example by the inverse of its class frequency, has 500 trees, and 10 minimum samples at each leaf node. The maximum number of features sampled at each split is chosen from the square root of the total number of features: 20, 30, or 50 features based on which choice maximizes the area under the receiver operating characteristic (ROC) curve (Mason 1982) through a grid search with fivefold cross validation.

Multitask learning (Caruana 1997) is used by RF and an elastic net to predict the shape and scale parameters of the MESH distribution simultaneously. During training, the models choose parameters to minimize the total error across all outputs. Multitask learning provides additional information about the fitting process and maintains correlations among the output values. If separate ML models predict the shape and scale parameter values independently, then the predicted values will be uncorrelated. The spatial MESH distribution parameters are log transformed and normalized before being fit by the ML models in order to capture the log-linear relationship between the shape and scale parameters, as shown in Fig. 5, and to ensure that both parameters are weighted equally by the ML loss functions. Both RFs and elastic net regression (Zou and Hastie 2005) support multitask learning and are used in this experiment. The elastic net is a linear regression model with additional terms in its optimization function that penalize large weights. A default RF, called random forest, has 500 trees, minimum samples at the split node
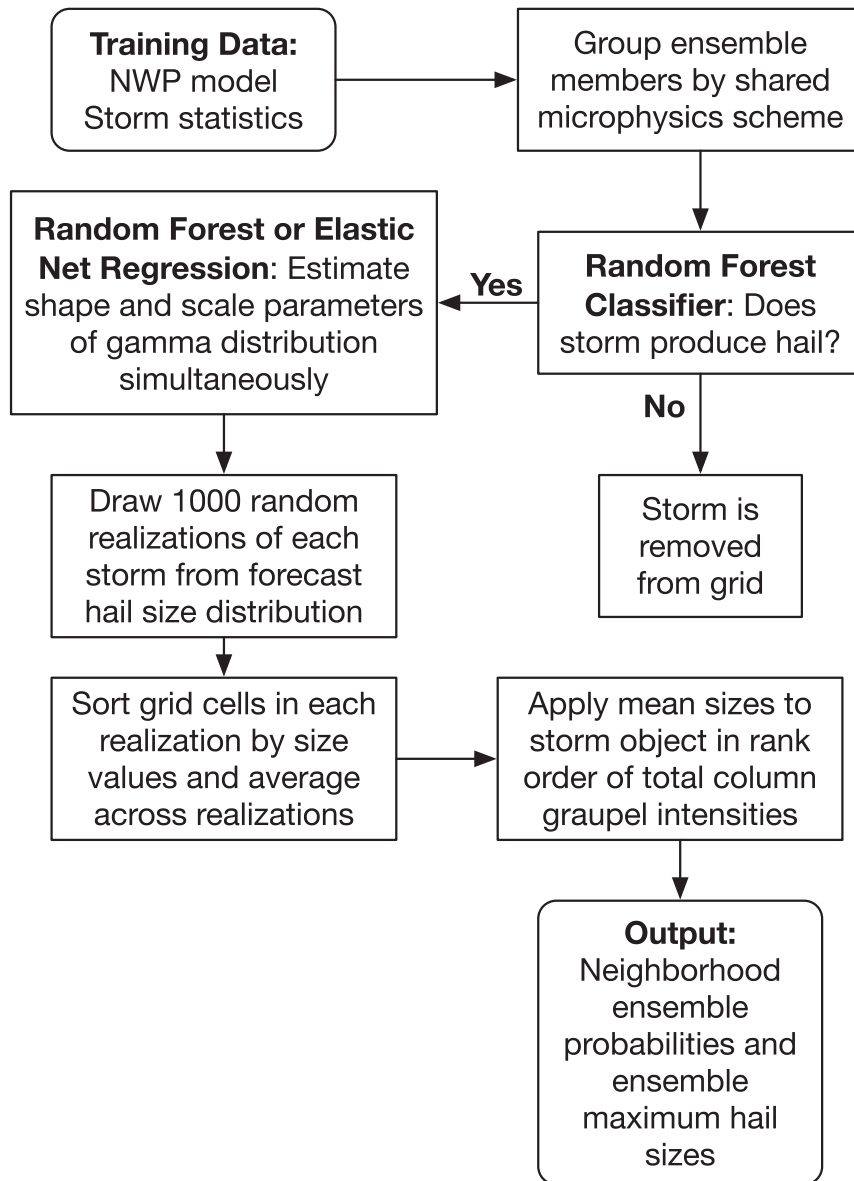
```
┌─────────────────────┐        ┌─────────────────────┐
│   Training Data:     │        │   Group ensemble    │
│    NWP model         │───────▶│  members by shared   │
│  Storm statistics    │        │ microphysics scheme  │
└─────────────────────┘        └─────────────────────┘
                                          │
                                          ▼
┌─────────────────────┐        ┌─────────────────────┐
│ Random Forest or Elastic     │   Random Forest     │
│ Net Regression: Estimate     │   Classifier: Does  │
│ shape and scale parameters◀──│  storm produce hail?│
│ of gamma distribution    Yes └─────────────────────┘
│ simultaneously       │                  │ No
└─────────────────────┘                   ▼
          │                     ┌─────────────────────┐
          ▼                     │     Storm is        │
┌─────────────────────┐        │     removed         │
│  Draw 1000 random   │        │    from grid        │
│  realizations of each │      └─────────────────────┘
│  storm from forecast │
│  hail size distribution │
└─────────────────────┘
          │
          ▼
┌─────────────────────┐        ┌─────────────────────┐
│ Sort grid cells in each     │  Apply mean sizes to │
│ realization by size   │─────▶│  storm object in rank│
│ values and average    │      │  order of total column│
│ across realizations   │      │  graupel intensities │
└─────────────────────┘        └─────────────────────┘
                                          │
                                          ▼
                               ┌─────────────────────┐
                               │     Output:         │
                               │  Neighborhood       │
                               │  ensemble           │
                               │  probabilities and  │
                               │  ensemble           │
                               │  maximum hail       │
                               │  sizes              │
                               └─────────────────────┘
```

FIG. 3. ML hail size forecasting procedure diagram.

of 10, and sampling square root of the total number of features. An optimized RF, called random forest CV, has 500 trees; fivefold cross-validated grid searching of the maximum number of features from the square root, 30, 50, and 100 features; and minimum samples at a splitting node from 5, 10, and 20 samples. The elastic net determines the ratio between the ridge and lasso terms from an independent validation set and normalizes the input features.

An RF classifier forecasts the probability of hail occurring, and RF and elastic net forecast the MESH size distribution for each storm in the testing set. To compare the ML model forecasts with other hail forecasting methods, the ML MESH distribution is translated into a deterministic set of MESH values overlaid onto the original storm in the NWP model (Fig. 3). The resulting grid can then be treated like any other diagnostic product and different gridded products, such as storm surrogate ensemble probabilities and ensemble maximum fields, can be generated. These products are used by forecasters at NOAA's Storm Prediction Center to create day-ahead hail outlooks, so presenting the forecasts in a familiar format helps build trust in the product and simplifies incorporating the information into a forecaster's existing workflow. If a forecast storm has a probability of hail occurrence less than 50%, it is
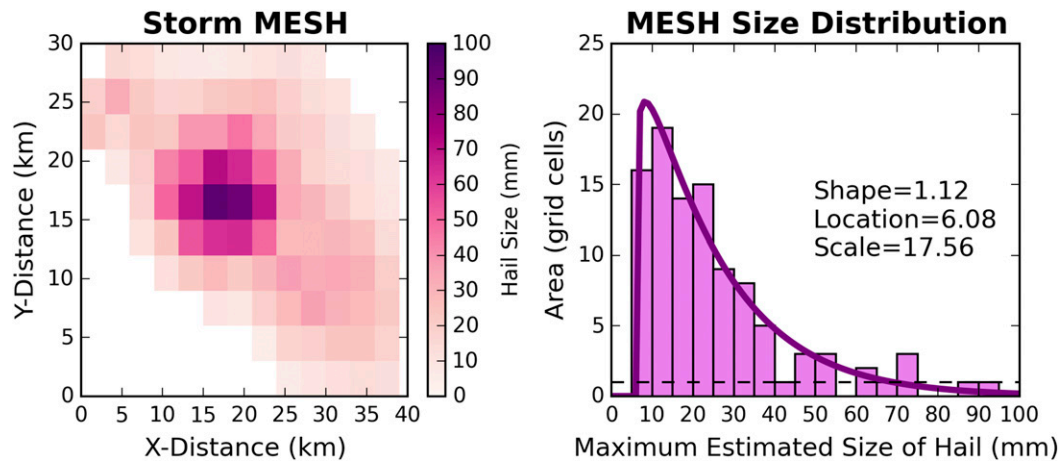
FIG. 4. An example MESH object and the resulting discrete and parametric MESH size distributions.

removed from the grid. Weighting each training example by the inverse of the class frequency (examples from the no-hail class are weighted less than examples from the hail class) ensures that the 50% threshold maximizes the separation between the prediction distributions of the two classes. A deterministic threshold is used to retain or remove individual forecast hailstorms because we did not want to decrease the range of forecast hail sizes artificially by combining the hail occurrence probability with the spatial MESH distribution. Because the gamma distribution has no upper limit, we utilized Monte Carlo sampling to generate hail size values for each storm grid point. For each grid point occupied by a forecast storm, 1000 random samples are drawn from the forecast spatial MESH distribution [Eq. (2)], corresponding to 1000 statistical realizations of the MESH spatial field. The MESH values within each statistical realization are sorted by size. The sorted sizes are then associated with each grid cell occupied by the forecast storm in order of CTG intensity. For example, the set of largest sampled MESH sizes are associated with the largest CTG grid point. Within each set of MESH samples, the mean and percentiles of the MESH sizes are then calculated and output as gridded deterministic fields. The mean forecast MESH size field is used for further evaluation and comparison with other hail forecast methods.

Storm-surrogate ensemble probabilities (Sobash et al. 2011) of hail at different sizes are created from the ML models and are compared with similar probabilities derived from thresholding the storm surrogate variables and HAILCAST. A storm surrogate probability forecast creates a smooth probability field from a deterministic forecast that mimics a convective outlook from the Storm Prediction Center, which is the intended product for which these hail forecasts would provide

guidance. The 3-km grid for each ensemble member is subsampled into a 42-km grid, and each point on the 42-km grid is marked with a 1 if at least one 3-km grid point within a 42-km radius over a 24-h period exceeds a specified intensity threshold. The 42-km radius has been chosen based on the Storm Prediction Center convective outlook criteria of at least one report within 40 km of a point (Hamill et al. 2005) and to be divisible by the 3-km grid spacing. A Gaussian filter with a standard deviation of 42 km is then applied to the coarse binary grid and spreads the probability mass to surrounding grid points to reflect the estimated spatial uncertainty. Larger standard deviations for the Gaussian filter result in a greater probability of detection but also decrease the sharpness and increase the false alarm ratio. Storm-surrogate neighborhood probabilities are also calculated for HAILCAST and storm surrogate variables, including simulated radar reflectivity at the $-10°C$ level, UH, and CTG. UH thresholds of 75 and $150\,\mathrm{m^2\,s^{-2}}$, a reflectivity threshold of $60\,\mathrm{dB}Z$ (used for both 25- and 50-mm hail), and CTG thresholds of 25 and $50\,\mathrm{kg\,m^{-2}}$ were used to discriminate 25- and 50-mm-diameter hail, respectively. These thresholds were chosen based on previous experience with storm surrogate parameters and sensitivity studies for severe weather forecasting, such as Sobash et al. (2016).

d. Evaluation

The training data for the ML models were pooled so as to maximize the size of the dataset while preserving the temporal independence of the evaluation data. For the CAPS ensemble, each ML model is trained on a pool of ensemble members sharing the same microphysics parameterization scheme. The 2014 CAPS ensemble output is used to train the ML models, and the 2015 ensemble is used for evaluation. The training set consists
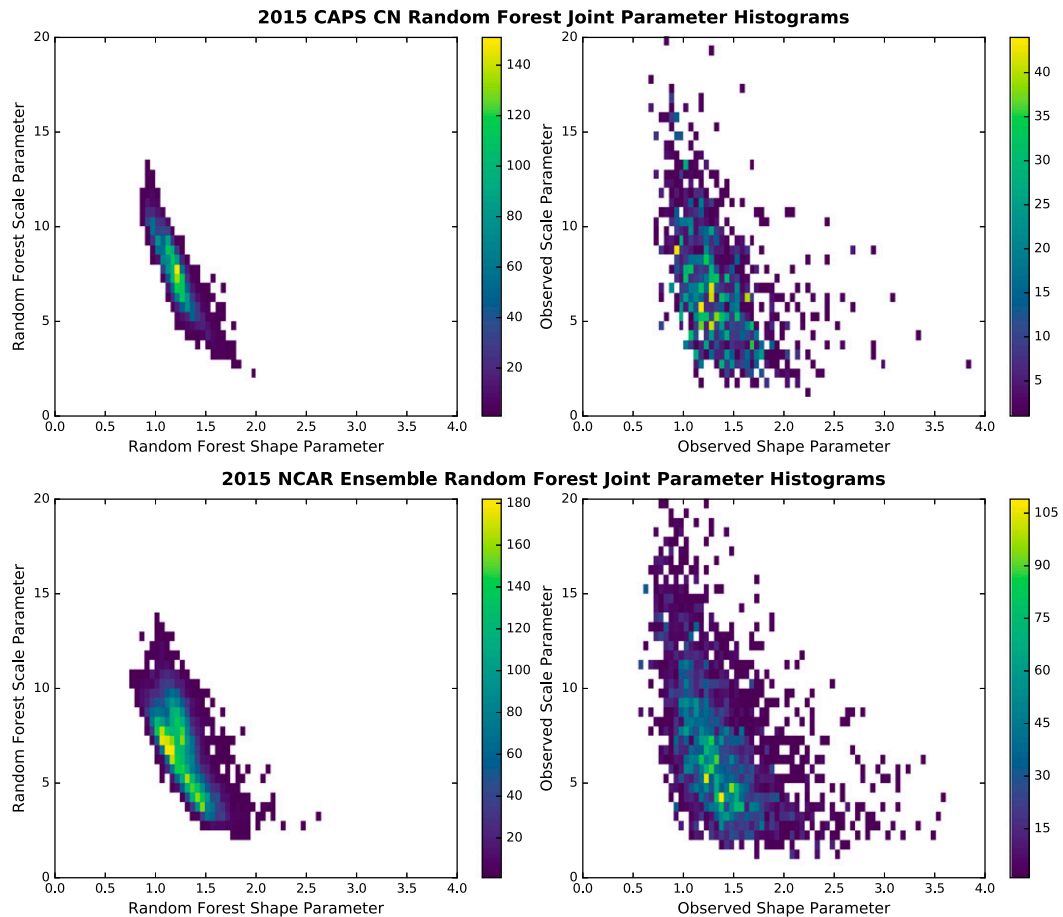
FIG. 5. Joint distribution of the shape and scale parameters for the (top) CAPS and (bottom) NCAR ensemble control member RF forecast and observations. The other CAPS and NCAR ensemble members exhibited similar relationships.

of 20 model runs on weekdays between 6 May and 6 June 2014. Because the 2015 ensemble used the P3 microphysics scheme in place of WDM6, the P3 members are evaluated using models trained on the Milbrandt and Yau members. The testing dataset includes 18 runs on weekdays from 12 May to 5 June 2015. One run with missing MESH observation data during that period is excluded from the evaluation.

Training and evaluation for the NCAR ensemble is performed cyclically with a new round of training performed every 2 weeks and testing performed over the subsequent 2 weeks. Since each member uses the same parameterizations and is equally likely, all members are pooled into the training dataset for the ML models. The forecasting period runs from 15 May through 30 July 2015. Because the CAPS and NCAR ensembles have different training and evaluation periods, the evaluation statistics cannot be used to determine whether one ensemble configuration is better than the other.

The forecast methods are evaluated using ROC curves, performance diagrams, reliability diagrams, and attributes diagrams, as well as metrics related to all of those diagrams. The formulas for every score are shown in Table 4. ROC curves (Mason 1982) divide a probabilistic forecast into a set of binary forecasts by varying the decision threshold. The goal of a ROC curve is to show how well a forecasting method discriminates between two outcomes. The probability of detection (POD) for each decision threshold is plotted against the probability of false detection (POFD). Skilled forecasts will consistently have a higher POD than POFD for each decision threshold. The amount of skill is quantified by the area under the ROC curve (AUC). An AUC higher than 0.5 is considered skilled. The performance diagram (Roebber 2009) is a variation of the ROC curve in which the POFD is replaced with the success ratio (SR), which is the additive inverse of the false alarm ratio (FAR). In addition to these scores, the frequency bias (FB) is shown as rays expanding outward from the origin and

TABLE 4. Scores used to evaluate the hail forecasts. TP is a true positive, FP is a false positive, FN is a false negative, and TN is a true negative. The number of forecasts is N, the number of forecasts for a given probability threshold k is n_k, and the number of probability thresholds is K.

Score	Formula
Probability of detection	$POD = \dfrac{TP}{TP + FN}$
Probability of false detection	$POFD = \dfrac{FP}{FP + TN}$
Success ratio	$SR = \dfrac{TP}{TP + FP}$
Critical success index	$CSI = \dfrac{TP}{TP + FP + FN}$
Frequency bias	$FB = \dfrac{TP + FP}{TP + FN}$
Area under ROC	$AUC = \displaystyle\int_{POFD=0}^{1} POD\, d POFD$
Forecast probability	p_k
Observed relative frequency	$\bar{o}_k$
Climatological probability	$\bar{o}$
Reliability	$REL = \dfrac{1}{N}\displaystyle\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2$
Resolution	$RES = \dfrac{1}{N}\displaystyle\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2$
Uncertainty	$UNC = \bar{o}(1 - \bar{o})$
Brier score	$BS = REL - RES + UNC$
Brier skill score	$BSS = \dfrac{RES - REL}{UNC}$

the critical success index (CSI) as contours that maximize in the top-right corner of the plot. These statistics are all independent of the number of true negatives, so they are more sensitive to how well the positive event is predicted. This feature is particularly important for rare event forecast evaluation.

Probabilistic forecasts are also assessed for their reliability and resolution. Reliability is the difference between the forecast probability and the relative frequency of the forecast event occurring conditional on that probability being forecast. Resolution is the difference between the climatological probability of an event and the relative frequency of the event occurring for a given forecast probability. Information about reliability and resolution is contained within the Brier score (BS; Brier 1950), which can be decomposed into reliability, resolution, and uncertainty components (Murphy 1973). These components can be reorganized to calculate the Brier skill score (BSS), which is the difference between reliability and resolution divided by the uncertainty. The reliability diagram displays the observed relative frequency versus the forecast probability. If the forecast probability is greater than the observed relative frequency, then the model is overforecasting. The attributes diagram (Hsu and Murphy 1986) is a reliability diagram that also shows the climatological probability and the regions where the resolution is larger than the reliability and thus the BSS is positive. The attributes diagram also contains an inset showing the frequency of forecasts made at each probability threshold, which can be used to assess the sharpness, or variance across probabilities, of the forecasts. Reliability diagrams can also be used to evaluate the predictions of regression models by discretizing the forecast values and determining the mean observed value for a given range of forecast values.

3. Results

a. Individual hailstorm forecast evaluation

Probabilistic forecasts of hail occurrence from the RF classifier are evaluated for the CAPS and NCAR ensembles. The top panel in Fig. 6 shows the ROC curve and reliability diagram for the CAPS ensemble members. Both diagrams evaluate probabilities in increments of 0.05 from 0 to 1. The observed relative frequency in the reliability diagrams refers to the percentage of hailstorms observed when a certain probability was forecast. The member ROC curves cluster by microphysics. The ML forecasts for Thompson members possess the lowest AUCs (about 0.73), and P3 members have higher POD at low probability thresholds and AUCs around 0.78. In the reliability diagram, all members display a consistent underforecasting bias, which may be caused by the change in grid spacing between the 2014 and 2015 CAPS ensembles and any resulting changes in the forecast variable distributions. At low forecast probabilities, the Thompson microphysics members display a higher observed relative frequency than other members, resulting in a lower BSS. For the NCAR ensemble (Fig. 6, bottom), the RF shows AUCs of about 0.70. The reliability diagram indicates that the probability forecasts are reliable but have a small overforecasting bias. The CAPS ensemble members show more variability in performance, which is expected from using multiple microphysics schemes. The differences in scores have not been evaluated for statistical significance.

Reliability diagrams for the shape and scale parameters in the CAPS and NCAR ensembles indicate how closely the forecast values correspond to the observed values on average (Fig. 7). The CAPS ensemble exhibits poor calibration with an underforecasting bias for small values of the shape and scale parameter and an overforecasting bias for large values. All of the ensemble members show similar issues. The NCAR ensemble reliability curves exhibit better calibration than the CAPS ensemble but still show biases at extreme parameter

2015 CAPS Ensemble Hail Yes-No Evaluation

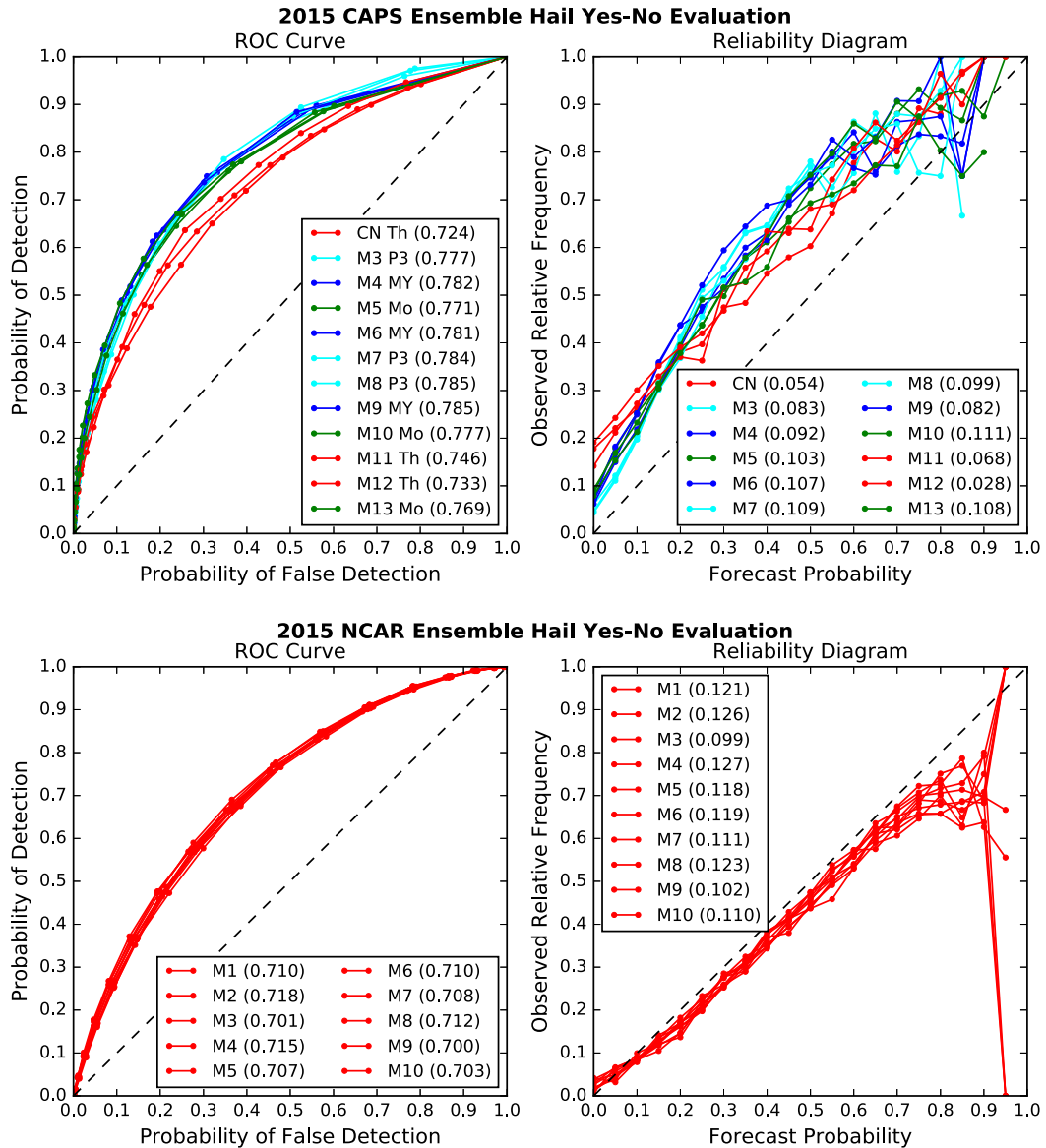**2015 NCAR Ensemble Hail Yes-No Evaluation**

FIG. 6. ROC curve and reliability diagram of RF hail occurrence forecasts from each member of the 2015 CAPS and NCAR ensembles. Members are colored by microphysics scheme. The AUC for each member is indicated in the ROC curve legend, and the BSS for each member is in the reliability diagram legend.

values. Poor sharpness in the shape and scale parameter forecasts results in forecast MESH distributions that underestimate the probability of very large hail when conditions are favorable and overestimate the probability of large hail when conditions are not favorable.

b. Full-period forecast evaluation

Storm surrogate probabilities were calculated over the 24-h period from 1200 to 1200 UTC for the ML models and raw storm surrogate variables over the testing sets. These probabilities are analogous to Storm Prediction Center convective outlooks and allow for a

more direct comparison between the two types of forecast methods.

The performance diagrams in Fig. 8 show the POD and success ratio at 5% probability intervals. Lower probability thresholds will have higher POD and appear in the upper portion of the diagram. Because POD decreases faster than FAR for all models, we focus our analysis on the differences in FAR over a range of probability thresholds. For the 25-mm-diameter hail threshold, the ML methods all have lower FAR values than every other model at all probability thresholds (Fig. 8). The FAR for UH is nearly as low as the
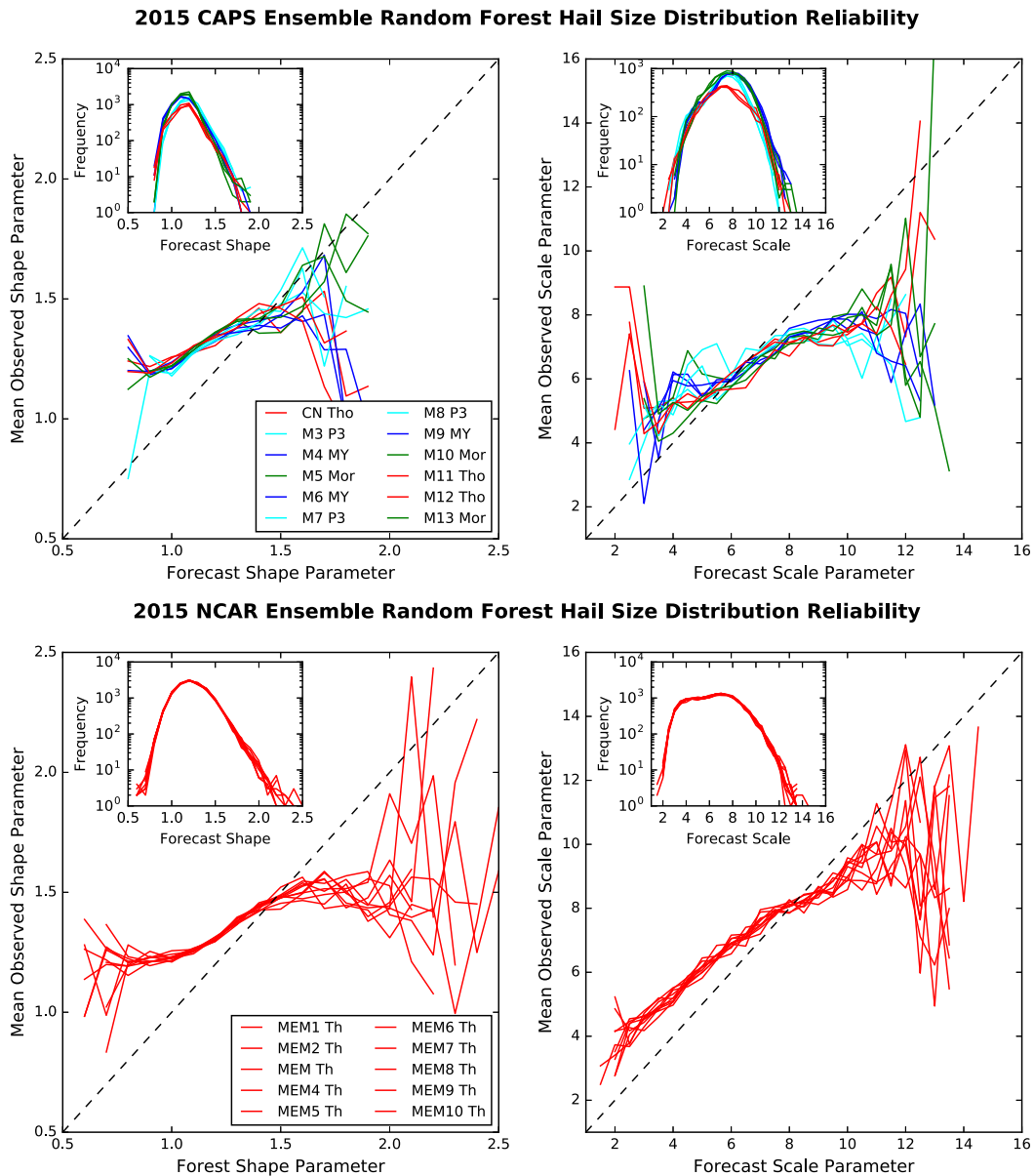
FIG. 7. Reliability diagrams showing the (top) CAPS and (bottom) NCAR ensemble mean RF observed hail size distribution parameters for each forecast parameter value.

ML models. For the 50-mm hail threshold, all models have a much higher FAR, but their relative rankings stay the same (Fig. 8). At higher probability thresholds (near the bottom of the diagram), the ML models have much lower FAR compared with the other methods but are only able to detect a small percentage of events. The NCAR ensemble performance diagrams highlight the difference in false alarms for each method at high probability thresholds. UH has a lower POD than total graupel for 25-mm hail but also has a lower FAR for most probability thresholds (Fig. 8). The ML

methods are able to maintain a lower FAR than UH for all thresholds. The RFs maintain a higher CSI consistently (Fig. 8). The elastic net underperforms compared to the RFs, and the performance of UH is comparable.

The attributes diagrams in Fig. 9 indicate the reliability and sharpness of each method for predicting the probability of 25- and 50-mm hail over a 40-km area, unlike the reliability diagrams in Fig. 6, which evaluated the probability of hail occurring for each potential hail storm object. The ML methods and UH are all reliable

CAPS Ensemble Performance
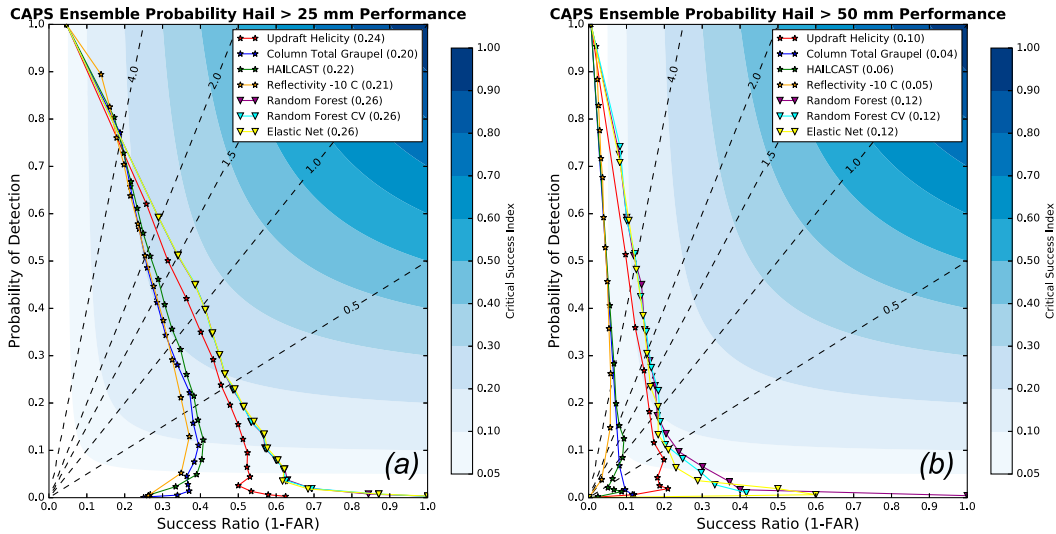
NCAR Ensemble Performance
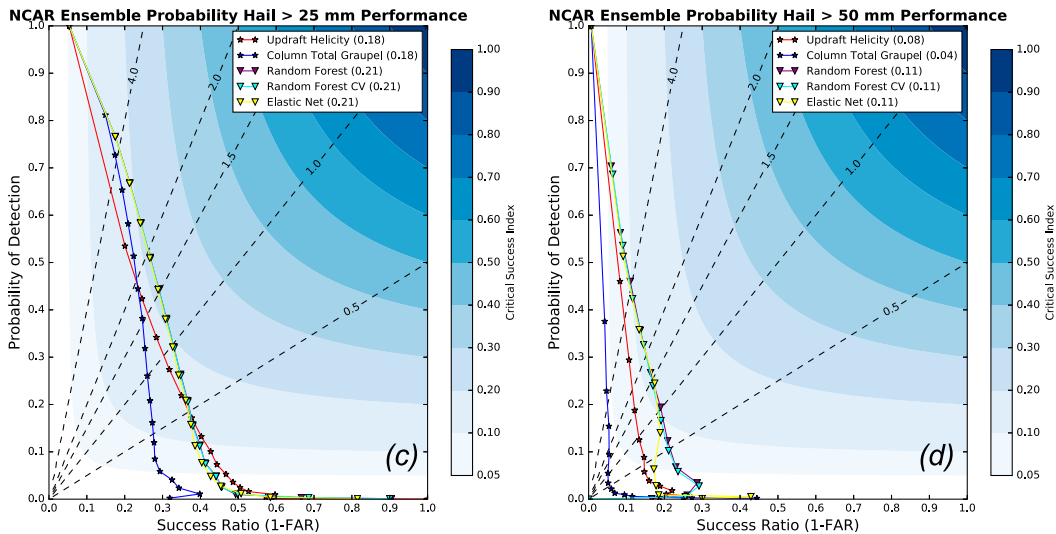
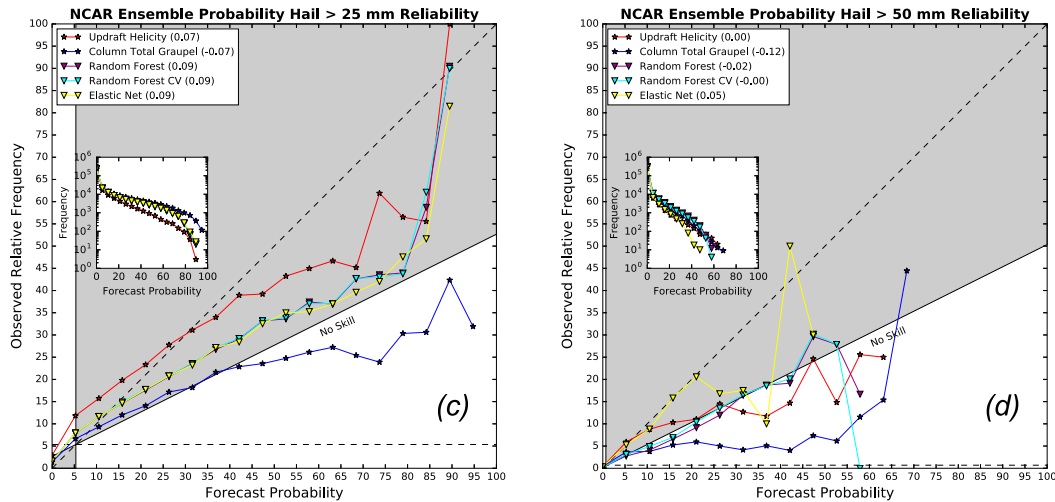FIG. 8. Performance curves for each (top) CAPS and (bottom) NCAR ensemble mean storm-surrogate probability for 25- and 50-mm hail. The maximum CSI for each curve is in the legend.

at the 25-mm threshold (Fig. 9) while the other storm surrogate methods are all overconfident. The overconfidence improves their AUC, because they detect more events at a given probability threshold, but hurts their reliability. All of the methods are sharp but HAILCAST has the largest number of forecasts at high probabilities. Every method is overconfident and less sharp at the 50-mm threshold (Fig. 9). The ML methods have less of an overconfidence bias than reflectivity and greater sharpness than CTG. HAILCAST is more overconfident than the ML models but still produces sharp forecasts. In addition to being better discriminators

for the NCAR ensemble, the ML methods also produce more reliable probabilities. While an overconfidence bias is apparent at both hail size thresholds, the ML methods have higher BSSs for 25- (Fig. 9) and 50-mm-diameter hail (Fig. 9). Both UH and the ML models are underconfident at low probability thresholds and then trend toward overconfidence at higher thresholds while still showing skill. The elastic net produces the most reliable probabilities for hail sizes of 50 mm at the expense of some sharpness compared with UH (Fig. 9). The RFs are overconfident with 50-mm-diameter hail probabilities.

CAPS Ensemble Reliability
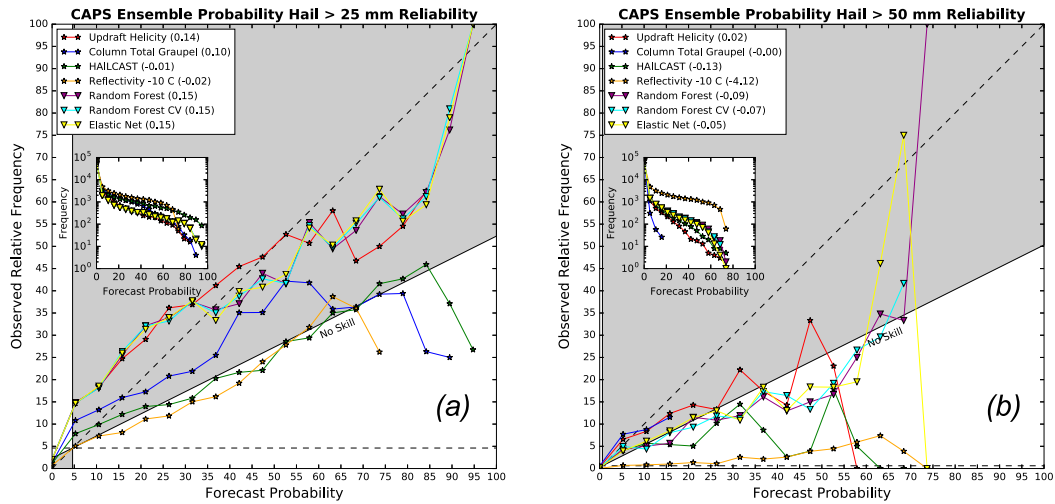
NCAR Ensemble Reliability

Fig. 9. Reliability curves for each (top) CAPS and (bottom) NCAR ensemble mean storm-surrogate probability of 25- and 50-mm hail. The BSS for each curve is shown in the legend. Points in the gray-shaded areas contribute positively to the BSS, and points outside the gray areas contribute negatively.

The spatial distribution of probabilistic hail forecasts from the 2015 CAPS ensemble over all evaluated runs is shown in Fig. 10. The chosen 10% probability threshold displays the maximum spatial extent of the forecasts at a threshold commonly used by forecasters to assess hail risk. RF and UH capture the two observed maxima in 25-mm hail frequency in the Texas Panhandle and across northeast Colorado. Overall, the RF and UH methods are comparable in their spatial forecast frequencies. The frequency maxima for HAILCAST and CTG are displaced eastward compared to the observed maximum. HAILCAST and CTG also produce large numbers of hail forecasts along the Gulf coast whereas

the RF and UH have none in that area. The RF captures the extent of the 50-mm reports (Fig. 10) while the other methods produce more 50-mm hail forecasts to the east of the observed maximum. RF and UH present two local maxima in Texas and northeast Colorado whereas CTG and HAILCAST have one maximum in north-central Texas.

The spatial distribution of hail forecasts from the NCAR ensemble is displayed in Fig. 11. Most of the observed hail reports occur along the high plains east of the Rocky Mountains with a smaller secondary maximum in the Southeast. For 25-mm hail, the ML methods capture the full extent of the high plains hail
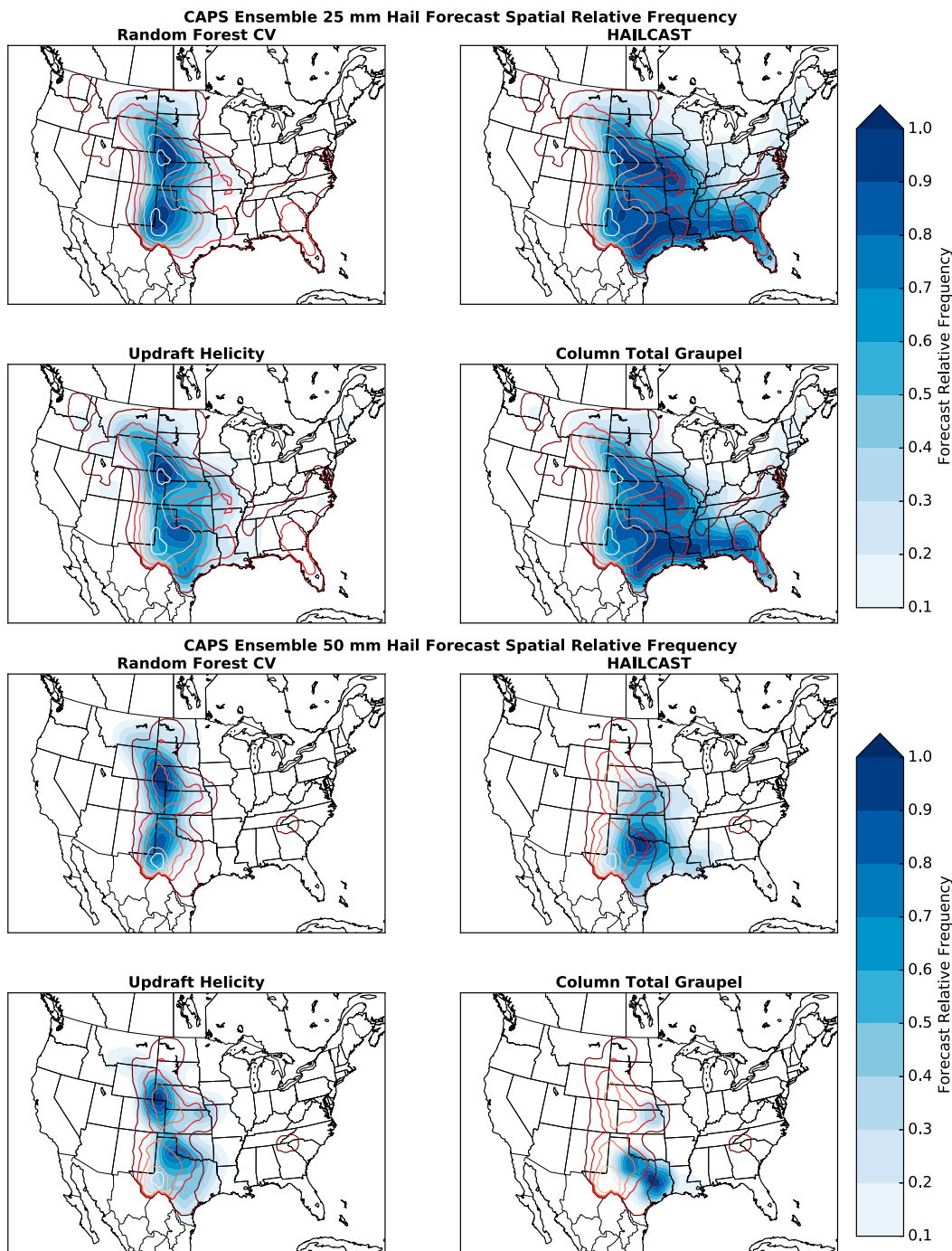
FIG. 10. The 2015 CAPS ensemble spatial distributions of 10% hail forecasts from select models at the 25- and 50-mm hail thresholds for storm-surrogate probability forecasts over the period from 1200 UTC the day the model is initialized to 1200 UTC the following day. The blue filled contours are forecast relative frequencies, and the red contours are observed relative frequencies at 20% increments starting at 10%.

observations while slightly underforecasting in the Southeast. UH follows a similar pattern. CTG greatly overforecasts hail frequency in the Southeast and underforecasts hail in the high plains with an eastward

bias in the maximum. An eastward bulge in eastern Nebraska from overnight MCSs is also visible in both the forecast and the observed relative frequencies from all models. The ML models overforecast the frequency
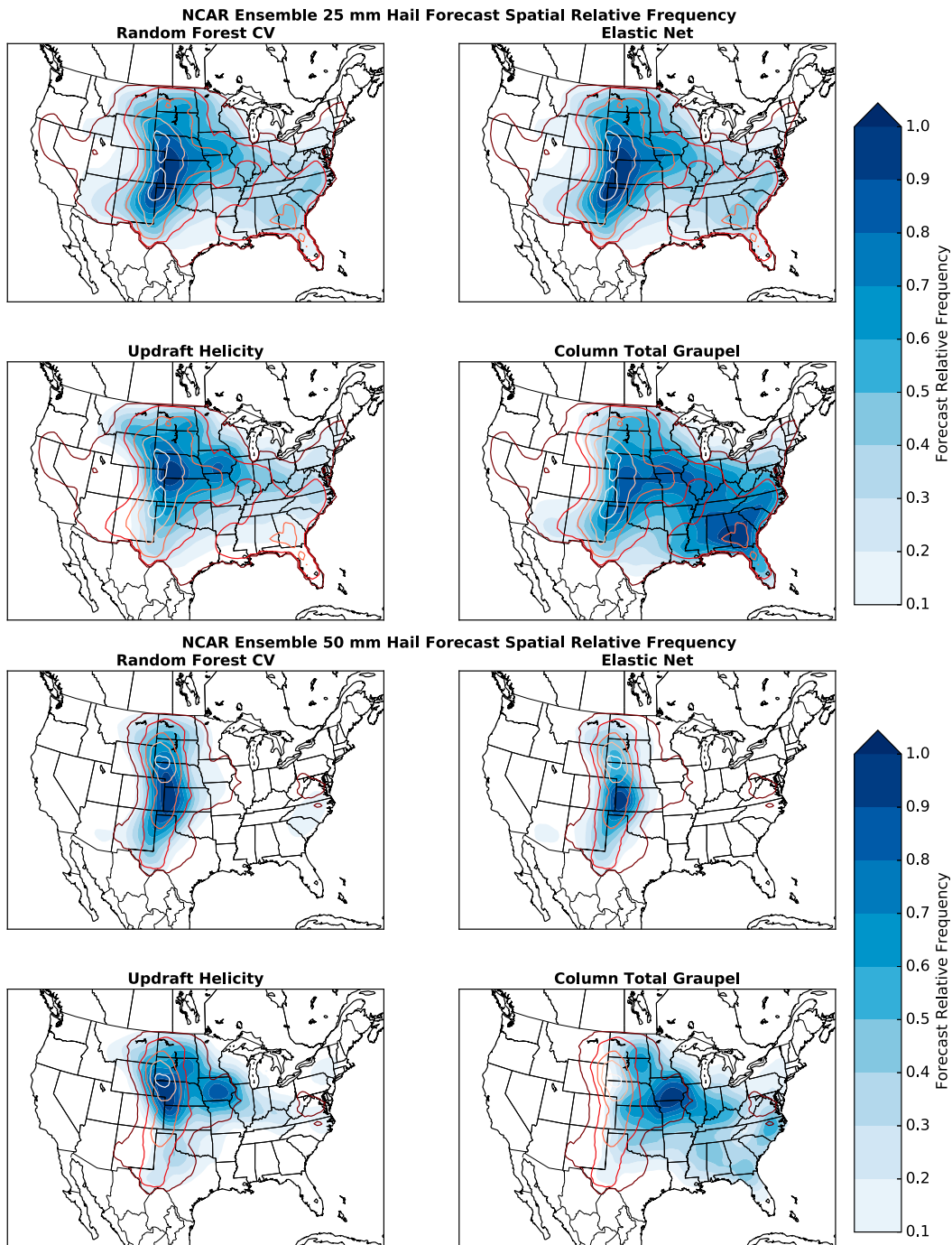
FIG. 11. As in Fig. 10, but for the NCAR ensemble. The spatial frequencies cover the period from 15 May through 31 Jul 2015.

of hail in the high plains while CTG does the same over the eastern half of the United States. The ML methods and UH have the most misses along the Gulf Coast. For 50-mm hail, the RF captures the full extent of the hail events in the high plains while the elastic net is more

concentrated in central Kansas (Fig. 11). The UH forecasts concentrate their presence in western Nebraska with a secondary maximum in Iowa, while CTG features the local maximum in Iowa but not in Nebraska.

4. Discussion

ML storm-based hail forecasts were evaluated and compared with other diagnostic hail forecasting methods to determine what value may be added to raw CAM ensemble output by these approaches. The different forecasting methods were evaluated using spring and summer 2015 runs of the CAPS and NCAR CAM ensembles and validated against gridded radar-estimated hail sizes. The ML models showed skill in discriminating between forecast storms that produced hail and those that did not but were underdispersive with the hail size forecasts. Additional calibration of the raw ML forecasts improved the sharpness of the size forecasts for the NCAR ensemble. For 24-h hail outlooks, the ML methods were better able to identify hail threat areas while minimizing false alarms compared with HAILCAST and other storm-surrogate methods. Of the existing storm surrogates, UH provided the best indicator of large hail. This result is expected given that UH is a good proxy for the strong, rotating updrafts found in supercells. The skill of ML hail forecasts is more sensitive to the choices made during preprocessing than to the choice of the ML model. Any statistical or ML model will produce less skilled forecasts if NWP model configuration changes result in large alternations to the distributions of important input variables because the correlation with the forecast output variable will have changed. The NCAR ensemble ML algorithms benefited from having a larger amount of training data due to the longer training period and the ability to aggregate across 10 diverse but similarly configured ensemble members.

The verification statistics and spatial maps showed that the hail forecasting methods exhibit more pronounced forecasting biases in particular areas. The ML models and UH work very well for discriminating hail in plains supercells, particularly storms in the high plains and just east of the Rockies. Both of these methods take into account shear and updraft intensity. CTG and the 2015 HAILCAST are more influenced by CAPE and less by shear, so they produce hail in any storm that has a strong updraft. For a given CAPE environment, the amount of shear has a significant impact on the amount of hail growth (Dennis and Kumjian 2017). The ML model uses the LCL height to account for relative humidity effects on melting and gives lower probabilities to areas with low LCL heights, but it is too aggressive with lowering probabilities in the Southeast and misses too many storms. In addition, CTG does not account for melting, and HAILCAST melts hail in cloud based on hailstone temperature, cloud temperature, and collisions with liquid water and below cloud level based on the

mean wet-bulb temperature in that layer. HAILCAST and CTG do not fully account for the melting that occurs with storms in the Southeast and tend to overpredict hail there. Because of the updraft issue, HAILCAST was significantly updated following the 2015 HWT Spring Experiment to account for hail growth occurring more on the edge of updrafts than in the center; these improvements are discussed by Adams-Selin and Ziegler (2016). The hailstorms in the Southeast in May–July tend to be pulse thunderstorms (Smith et al. 2012), which lack rotating updrafts, so hail methods calibrated to plains supercells that make up the bulk of hail events will not detect them as well. Using a lower decision threshold may help capture some of these cases that are currently missed with the 50% decision threshold. Alternatively, regionally calibrated decision thresholds may capture different environments more accurately.

While constraining ML hail forecasts to areas where the NWP model produces storms does produce better forecasts, there are situations in which this approach will struggle compared with ingredients-based methods. If the CAM forecast contains errors with the placement, timing, and evolution of storms, then the hail forecasts dependent on those storm forecasts will also contain those errors. These struggles tend to occur in scenarios where large-scale forcing is weaker, leading to convection initiation and evolution being governed by poorly observed mesoscale effects. When an area receives multiple days of convection, errors in forecasting the diurnal cycle of convection lead to additional spatial and temporal biases. Hail forecasts based solely on environmental parameters will tend to have better coverage in these situations and will detect hail threat areas that storm-based methods may miss.

Further improvements to ML model forecasts are possible with some additions to the current modeling framework. Additional diagnostic variables describing conditions within the hail growth region and the melting layer near the ground may provide stronger links between those processes and the forecast hail size. A more thorough sensitivity study of the forecast and observed storm object identification and matching procedure could lead to stronger correlations between the different weather variables and radar-estimated hail size. Longer convection-allowing model archives with more consistent configurations over multiple severe weather seasons could allow the ML models to capture a larger sample of extreme hail events. The evaluation of hail forecast accuracy conditioned on the storm environment could assist forecasters with knowing when to trust certain models more. Probabilities from the MESH distributions themselves could be interrogated by a forecaster in an interactive viewer to identify which

areas have storms with a higher probability of very large hail. Creating and evaluating these visualizations is an area of future research. The ML hail forecasts were produced using the Hagelslag (https://github.com/djgagne/hagelslag) open-source Python package developed by the lead author. Data used for this project are available upon request.

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, doi:10.1175/MWR-D-16-0027.1.

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data-mining technique. *Wea. Forecasting*, **31**, 581–599, doi:10.1175/WAF-D-15-0113.1.

Anderson, J., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, doi:10.1175/2009BAMS2618.1.

Benjamin, S., 2014: From the RAPv3/HRRRv2 deterministic era to the NARRE/HRRRE ensemble era. *2014 Warn-On-Forecast and High Impact Weather Workshop*, Norman, OK, National Severe Storms Laboratory, http://www.nssl.noaa.gov/projects/wof/documents/workshop2014/.

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

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

Brimelow, J. C., G. W. Reuter, and E. R. Poolman, 2002: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, **17**, 1048–1062, doi:10.1175/1520-0434(2002)017<1048:MMHSIA>2.0.CO;2.

Brown, T. M., W. H. Pogorzelski, and I. M. Giammanco, 2015: Evaluating hail damage using property insurance claims data. *Wea. Climate Soc.*, **7**, 197–210, doi:10.1175/WCAS-D-15-0011.1.

Caruana, R., 1997: Multitask learning. *Mach. Learn.*, **28**, 41–75, doi:10.1023/A:1007379606734.

Changnon, S. A., 2009: Increasing major hail losses in the U.S. *Climatic Change*, **96**, 161–166, doi:10.1007/s10584-009-9597-z.

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, doi:10.1175/WAF-D-11-00151.1.

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

——, J. Gao, P. T. Marsh, T. Smith, J. S. Kain, J. Correia Jr., M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, doi:10.1175/WAF-D-12-00038.1.

Dennis, E. J., and M. R. Kumjian, 2017: The impact of vertical wind shear on hail growth in simulated supercells. *J. Atmos. Sci.*, **74**, 641–663, doi:10.1175/JAS-D-16-0066.1.

Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A radar-based methodology. *J. Atmos. Oceanic Technol.*, **10**, 785–797, doi:10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2.

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, doi:10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Fawbush, E. F., and R. C. Miller, 1953: A method for forecasting hailstone size at the earth's surface. *Bull. Amer. Meteor. Soc.*, **34**, 235–244.

Foote, G. B., 1984: A study of hail growth utilizing observed storm conditions. *J. Climate Appl. Meteor.*, **23**, 84–101, doi:10.1175/1520-0450(1984)023<0084:ASOHGU>2.0.CO;2.

Gagne, D. J., II, 2016: Coupling data science techniques and numerical weather prediction models for high-impact weather prediction. Ph.D. dissertation, University of Oklahoma, 185 pp., http://hdl.handle.net/11244/44917.

——, A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Oceanic Technol.*, **26**, 1341–1353, doi:10.1175/2008JTECHA1205.1.

——, ——, ——, M. Coniglio, J. Correia Jr., and M. Xue, 2015: Day-ahead hail prediction integrating machine learning with storm-scale numerical weather models. *27th Conf. on Innovative Applications of Artificial Intelligence*, Austin, TX, AAAI, 3954–3960, http://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/view/9724.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203: TUOMOS>2.0.CO;2.

Hamill, T. M., R. S. Schneider, H. E. Brooks, G. S. Forbes, H. B. Bluestein, M. Steinberg, D. Meléndez, and R. M. Dole, 2005: The May 2003 extended tornado outbreak. *Bull. Amer. Meteor. Soc.*, **86**, 531–542, doi:10.1175/BAMS-86-4-531.

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, doi:10.1175/MWR3199.1.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, doi:10.1016/0169-2070(86)90048-8.

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, doi:10.1175/2009WAF2222230.1.

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

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

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

Lakshmanan, V., and T. Smith, 2010: An objective method of evaluating and devising storm-tracking algorithms. *Wea. Forecasting*, **25**, 701–709, doi:10.1175/2009WAF2222330.1.

——, K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.*, **26**, 523–537, doi:10.1175/2008JTECHA1153.1.

——, K. L. Elmore, and M. B. Richman, 2010: Reaching scientific consensus through a competition. *Bull. Amer. Meteor. Soc.*, **91**, 1423–1427, doi:10.1175/2010BAMS2870.1.

Lim, K.-S. S., and S.-Y. Hong, 2010: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models. *Mon. Wea. Rev.*, **138**, 1587–1612, doi:10.1175/2009MWR2968.1.

Manzato, A., 2013: Hail in northeast Italy: A neural network ensemble forecast using sounding-derived indices. *Wea. Forecasting*, **28**, 3–28, doi:10.1175/WAF-D-12-00034.1.

Marzban, C., and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610, doi:10.1175/1520-0434(2001)016<0600:ABNNFS>2.0.CO;2.

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

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.*, https://doi.org/10.1175/BAMS-D-16-0123.1, in press.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys.*, **20**, 851–875, doi:10.1029/RG020i004p00851.

Milbrandt, J. A., and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. Part I: Analysis of the role of the spectral shape parameter. *J. Atmos. Sci.*, **62**, 3051–3064, doi:10.1175/JAS3534.1.

Moore, J. T., and J. P. Pino, 1990: An interactive method for estimating maximum hailstone size from forecast soundings. *Wea. Forecasting*, **5**, 508–525, doi:10.1175/1520-0434(1990)005<0508:AIMFEM>2.0.CO;2.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests. *J. Atmos. Sci.*, **72**, 287–311, doi:10.1175/JAS-D-14-0065.1.

——, G. Thompson, and V. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007, doi:10.1175/2008MWR2556.1.

——, J. A. Milbrandt, G. H. Bryan, K. Ikeda, S. A. Tessendorf, and G. Thompson, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part II: Case study comparisons with observations and other schemes. *J. Atmos. Sci.*, **72**, 312–339, doi:10.1175/JAS-D-14-0066.1.

Munkres, J., 1957: Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.*, **5**, 32–38, doi:10.1137/0105003.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, doi:10.1023/B:BOUN.0000020164.04146.98.

Nateghi, R., S. Guikema, and S. M. Quiring, 2014: Power outage estimation for tropical cyclones: Improved accuracy with simpler models. *Risk Anal.*, **34**, 1069–1078, doi:10.1111/risa.12131.

Nelson, S. P., 1983: The influence of storm flow structure on hail growth. *J. Atmos. Sci.*, **40**, 1965–1983, doi:10.1175/1520-0469(1983)040<1965:TIOSFS>2.0.CO;2.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Rasmussen, R. M., and A. J. Heymsfield, 1987: Melting and shedding of graupel and hail. Part I: Model physics. *J. Atmos. Sci.*, **44**, 2754–2763, doi:10.1175/1520-0469(1987)044<2754:MASOGA>2.0.CO;2.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, doi:10.1175/2008WAF2222159.1.

Rosencrants, T. D., and W. S. Ashley, 2015: Spatiotemporal analysis of tornado exposure in five US metropolitan areas. *Nat. Hazards*, **78**, 121–140, doi:10.1007/s11069-015-1704-z.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.

Smith, B. T., R. L. Thompson, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154, doi:10.1175/WAF-D-11-00116.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme

phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, doi:10.1175/WAF-D-10-05046.1.

——, C. S. Schwartz, G. S. Romine, K. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, doi:10.1175/WAF-D-15-0138.1.

Sukoriansky, S., B. Galperin, and V. Perov, 2005: Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. *Bound.-Layer Meteor.*, **117**, 231–257, doi:10.1007/s10546-004-6848-4.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, doi:10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

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, doi:10.1175/2007WAF2007005.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, doi:10.1007/s10994-013-5346-7.

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, doi: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, doi:10.1175/2011BAMS-D-11-00047.1.

Zou, H., and T. Hastie, 2005: Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc.*, **67B**, 301–320, doi:10.1111/j.1467-9868.2005.00503.x.