A new paradigm for medium-range severe weather forecasts: probabilistic random forest-based predictions

Aaron J. Hill\textsuperscript{a}, Russ S. Schumacher\textsuperscript{a}, and Israel Jirak\textsuperscript{b}

\textsuperscript{a} Colorado State University, Fort Collins, CO
\textsuperscript{b} Storm Prediction Center, NOAA

ABSTRACT: Historical observations of severe weather and simulated severe weather environments (i.e., features) from the Global Ensemble Forecast System v12 (GEFSv12) Reforecast Dataset (GEFS/R) are used in conjunction to train and test random forest (RF) machine learning (ML) models to probabilistically forecast severe weather out to days 4–8. RFs are trained with ~9 years of the GEFS/R and severe weather reports to establish statistical relationships. Feature engineering is briefly explored to examine alternative methods for gathering features around observed events, including simplifying features using spatial averaging and increasing the GEFS/R ensemble size with time-lagging. Validated RF models are tested with ~1.5 years of real-time forecast output from the operational GEFSv12 ensemble and are evaluated alongside expert human-generated outlooks from the Storm Prediction Center (SPC). Both RF-based forecasts and SPC outlooks are skillful with respect to climatology at days 4 and 5 with degrading skill thereafter. The RF-based forecasts exhibit tendencies to underforecast severe weather events, but they tend to be well-calibrated at lower probability thresholds. Spatially averaging predictors during RF training allows for prior-day thermodynamic and kinematic environments to generate skillful forecasts, while time-lagging acts to expand the forecast areas, increasing resolution but decreasing objective skill. The results highlight the utility of ML-generated products to aid SPC forecast operations into the medium range.

SIGNIFICANCE STATEMENT: Medium-range forecasts generated from statistical models are explored here alongside operational forecasts from the Storm Prediction Center (SPC). Human forecasters at the SPC rely on traditional numerical weather prediction model output to make medium-range outlooks and statistical products that mimic operational forecasts can be used as tools for forecasters. The statistical models relate simulated severe weather environments from a global weather model to historical records of severe weather and perform noticeably better than human-generated outlooks at shorter lead times (e.g., day 4 and 5) and are capable of capturing the general location of severe weather events 8 days in advance. The results highlight the value in these data-driven methods in supporting operational forecasting.

1. Introduction

Operational predictions of severe weather hazards (i.e., tornadoes, hail, and wind) are under the purview of the National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center (SPC), which is responsible for “timely and accurate forecasts and watches for severe thunderstorms and tornadoes over the contiguous United States” (SPC2020). The SPC uses forecast guidance from numerical weather prediction (NWP) models, including post-processed products, diagnostic parameters (e.g., storm relative helicity), as well as current observations, to issue outlooks 1–8 days in advance of the threat of severe weather; outlooks issued in the 1 and 2-day timeframe delineate threats for specific hazards whereas day 3–8 outlooks highlight risk areas of any severe hazard. Because of the tremendous societal and financial impacts of severe weather events, including 10 severe-weather attributed billion-dollar disasters in 2021 alone (NCEI2022), it is imperative that SPC forecasters receive reliable and valuable forecast information to inform their operational products and provide sufficient lead time to stakeholders of the threat of severe weather.

Deterministic and ensemble NWP model predictions of severe weather and associated hazards have improved substantially over the last decade as dynamical models have leveraged increased computing power to decrease grid spacing and increase effective resolution. Increases in model resolution have seemingly benefited short-term forecasts the most, as real-time high-resolution models are now capable of explicitly resolving parent convective storms (e.g., Done et al. 2004; Kain et al. 2008); these prediction systems are commonly referred to as convection-allowing models (CAMs). These advances provided opportunities to explicitly forecast weather hazards (e.g., tornadoes) by using proxies (e.g., updraft helicity; Sobash et al. 2011, 2016a, 2016b, Hill et al. 2021) in NWP model output, or generating calibrated guidance products (e.g., Gallo et al. 2018, Harrison et al. 2022, Jahn et al. 2022) that probabilistically depict severe weather threats like tornadoes and lightning. However, few of these methods have carried over into longer-range forecasting because of their dependence on CAMs, which are limited to the near-term (< 4 days) due to computational constraints and undesirable, rapid increases in forecast spread from small-scale errors.
As a result, SPC forecasters must rely on global prediction systems to issue day 4–8 outlooks (hereafter referred to as the medium range), including both deterministic and ensemble systems that parameterize convective processes and as a result, they have limited value for severe weather forecasting.

Efforts have been made to leverage the large global ensemble datasets to generate post-processed and calibrated forecast products for medium-range forecast events as well. Post-processed fields become increasingly valuable at these extended lead times, offering a simplistic depiction for the threat of severe weather (e.g., calibrated severe thunderstorm probabilities, Bright and Gramps 2009) that may not be contained in highly-variable deterministic model output fields or ensemble output diagnostics (e.g., ensemble mean, variance). The U.S. National Blend of Models (Hamill et al. 2017) uses quantile mapping to generate post-processed precipitation forecasts in the medium range. Additionally, analog forecasting (Lorenz 1969) has been used to forecast high-impact weather environments (e.g., Hamill and Whitaker 2006), and typically involves training a regression model (e.g., logistic regression) on past environmental states that were coincident with severe weather reports and learning how similar environments relate to severe weather frequency (e.g., CIPS Analog-Based Severe Guidance). Using these developed statistical relationships, the current atmospheric patterns are interrogated to determine comparable analogs to base a new forecast. To the authors’ knowledge, no analog techniques have been published addressing severe weather forecasting in the medium range, and a need still exists for techniques and tools to generate skillful medium-range severe weather forecast guidance.

More recently, more advanced machine learning (ML) models have emerged into the meteorology domain as a complementary and alternative option to forecast severe weather hazards. Whereas dynamical models cannot explicitly forecast hazards below their effective resolution, ML models can be trained to forecast any hazard given a sufficiently accurate observational dataset (i.e., labels) and related environmental features (i.e., predictors). ML has become a widely used technique to post-process NWP model output, for example, to generate forecasts mimicking operational center products (e.g., Herman and Schumacher 2018a, Loken et al. 2019, Hill et al. 2020, Loken et al. 2020, Sobash et al. 2020, Hill and Schumacher 2021, Schumacher et al. 2021) and forecast severe weather or hazards more generally (e.g., Gagne et al. 2014, 2017, Jergensen et al. 2020, McGovern et al. 2017, Burke et al. 2020, Flora et al. 2021, Loken et al. 2022). Others have developed ML prediction systems using observations or reanalysis datasets (e.g., Gensini et al. 2021, Shield and Houston 2022), demonstrating ML-based prediction systems capable of highlighting conducive severe weather environments, for example. Hill et al. (2020) and Loken et al. (2020) employed random forests (RFs, Breiman 2001) to generate forecasts analogous to SPC outlooks, effectively creating post-processed and probabilistic first-guess forecasts of severe weather that could be used by forecasters when generating their human-based outlooks. Loken et al. (2020) used CAM output to train an RF and derive day-1 hazard forecast probabilities whereas Hill et al. (2020) used global ensemble output to generate day 1-5 forecasts. Both studies demonstrated that RFs could produce skillful and reliable forecasts of severe weather at short lead times, and Hill et al. (2020) further highlighted that incorporating the statistical product information into the human-generated SPC forecast (i.e., through an statistical weighting procedure) yielded a better forecast at day 1 than either individual component forecast; the statistical models outperformed SPC forecasters at days 2 and 3. Therefore, it is reasonable to hypothesize that a similar prediction system devoted to forecasting severe hazards beyond day 3 would benefit SPC forecasters issuing medium-range forecasts.

Building upon the work of Hill et al. (2020), this study trains and develops RFs to forecast severe thunderstorm hazards in the medium range (i.e., days 4–8). The RF infrastructure of the Colorado State University Machine Learning Probabilities (CSU-MLP) prediction system (e.g., Hill et al. 2020, Schumacher et al. 2021) is used herein to explore medium-range severe weather predictions and determine their utility in relation to operational forecasts. Feature engineering (e.g., Loken et al. 2022) is also explored (i.e., how to organize predictors from the dynamical model output used to train the RF models) to determine if medium-range ML-based forecast skill is impacted by the way predictors are gathered and trained with observations. All RF-based forecasts are evaluated alongside corresponding SPC outlooks to illustrate the relative value of incorporating the statistical guidance into forecast operations.

2. Methods

a. Random Forests

RFs combine the simplicity of the decision tree architecture with the robustness of ensemble tree methods. Individual decision trees are constructed beginning with the root node (i.e., top of the tree) where a subset of training examples (i.e., instances of severe weather or no severe weather observation) is extracted from the full training set via bootstrapping and a feature is selected that best splits the examples, i.e., the feature best describes separation of severe weather events from non-events. Successive nodes are similarly constructed along the branches of the tree until a maximum tree depth is reached or a minimum number
of training samples needed to split a node is breached, ending that particular branch in a terminal or “leaf” node. The leaf node either contains all the same observation types (e.g., all examples are severe weather events) or a mixture. As new inputs are supplied to a decision tree, e.g., from real-time forecast output, the tree can be traversed to a leaf node using the inputs, producing a deterministic or probabilistic prediction of severe weather from the single tree. The aggregation of all decision tree predictions from a forest of decision trees provides a probabilistic prediction for the threat of severe weather.

RF-based forecasts in this context are analogous to the SPC outlooks at days 4–8, i.e., the probability of severe weather occurrence within 40 km (or 25 mi) of a point over the daily 24-hour period defined by 1200–1200 UTC, and forecast products are constructed to mimic those produced by SPC (e.g., Fig. 1). For RF training, severe weather observations are encoded onto a grid by interpolating severe storm reports from the SPC Severe Weather Database (SPC 2022b) to NCEP grid 4 (0.5 degree grid spacing); the same grid is used to define RF features, further discussed in subsection 2.a.1. Any severe weather report during the 24-hour forecast window is interpolated to the 0.5 degree grid using a 40 km neighborhood, resulting in at least one grid point encoded for each severe weather report; the 0.5 degree grid has approximately 55 km grid spacing. Severe weather reports are defined as tornadoes of any EF rating, hail exceeding 1” (2.54 cm) in diameter, and convective wind speeds exceeding 58 mph (93 km/h). Training examples are encoded as 0 (no severe), 1 (non-significant severe), or 2 (significant severe) across the CONUS; the significant severe designation is a specific class of tornadoes eclipsing F2 or EF2 strength, hail exceeding 2” (5.08 cm) in diameter, and convective wind gusts exceeding 74 mph (119 km/h). Thus, the RFs are tasked with a prediction of 3 classes. However, because the current version of SPC day 4–8 outlooks do not include significant severe probability contours, CSU-MLP RF significant severe forecasts will not be formally evaluated in this work – readers are referred to the work of Hill et al. (2020) who go in-depth on significant severe forecast skill for day-1 forecasts.

1) Predictors

Meteorological features surrounding the encoded observations used for training are obtained from the Global Ensemble Forecast System version 12 (GEFSv12) Reforecast dataset (hereafter GEFS/R; Hamill et al. 2022; Zhou et al. 2022), a 5-member daily 0000-UTC initialized ensemble system that utilizes the finite volume cubed sphere (FV3) dynamical core. Reforecasts date back to 1 January 2000 and extend forward to the end of 2019. Variables with known or presumed relationships to severe weather are extracted from the GEFS/R, including convective available potential energy (CAPE), precipitable water (PWAT), and bulk vertical wind shear (e.g., SHEAR500); a full list of variables is provided in Table I. Due to inconsistent resolution in the GEFS/R dataset, e.g., near-surface variables have higher resolution (0.25 degree grid) than upper-tropospheric variables (0.5 degree grid), each relevant meteorological output field is interpolated to the 0.5 degree grid, which also aligns the simulated reanalysis meteorological environments with the encoded observations. Additionally, latitude, longitude, and Julian day are used as static features, coincident with the encoded severe weather report location.

Meteorological features are assembled in a forecast-point relative framework in which raw variables are gathered both at the observation training example point and over a pre-defined radius around the point. The pre-defined radius is set to 3 for all models trained in this manner. Additionally, the GEFS/R has 3-hourly temporal resolution across the 24 h forecast windows, allowing for both spatial and temporal sampling of the environment surrounding a severe weather report that occurred within the window; a depiction of the assembly method is provided in Fig. 2. The raw variables accumulated in space and time represent the ensemble median for that given point; previous work demonstrated the superiority of using the ensemble median as opposed to the mean or outliers of the ensemble distribution (Herman and Schumacher 2018a). In other words, a 2-dimensional time series is constructed for each meteorological variable at every grid point across the forecast window.

Two alternative feature assembly methods are also applied. Previous research employing the CSU-MLP prediction system has nearly universally applied the forecast-point relative framework to assemble features (e.g., Herman and Schumacher 2018a,b; Hill et al. 2020; Schumacher et al. 2021). However, Hill and Schumacher (2021) showed that spatially averaging the features (e.g., 1-D time series) yielded improved RF-based excessive rainfall forecasts, and they hypothesized that characterizing the environment at a particular time with a single mean value reduced noise during model training. Similarly, Loken et al. (2022) demonstrated that improved RF forecast skill could be achieved by using the ensemble mean at each spatial point rather than all individual members at each point because it reduced training noise. While Hill and Schumacher (2021) and Loken et al. (2022) both used CAM inputs to train RFs, which may have inherently more noise than a global model, the same spatial averaging procedure of Hill and Schumacher (2021) is employed here (e.g., Fig. 3) to explore if the medium-range RF predictions generated here suffer from noisy inputs and whether skill could be improved by removing that noise.

Another unique opportunity presents itself when spatially-averaging predictors: the total number of fea-
Fig. 1. (a) CSU-MLP day 4 forecast initialized 27 March 2022, valid for period ending 1200 UTC 31 March 2022, and corresponding (b) day 4 SPC outlook. Forecast probabilities of any severe hazard are shaded, and the circle icons are SPC local storm reports of tornadoes (red), hail (green), and wind (blue). Forecast BSS and report coverage are provided in the lower right and left corners of each panel, respectively.

Table 1. Meteorological features used for RF training and forecasts

| Symbol | Variable Description | Variable Type |
|--------|----------------------|---------------|
| APCP   | 3-hourly accumulated precipitation | thermodynamic |
| CAPE   | Convective available potential energy | thermodynamic |
| CIN    | Convective inhibition | thermodynamic |
| PWAT   | Precipitable water | thermodynamic |
| T2M    | 2-m temperature | thermodynamic |
| Q2M    | 2-m specific humidity | thermodynamic |
| U10    | 10-m latitudinal horizontal wind speed | kinematic |
| V10    | 10-m longitudinal horizontal wind speed | kinematic |
| MSLP   | Mean sea level pressure | kinematic |
| UV10   | 10-m wind speed | kinematic |
| SHEAR500 | Surface to 500 hPa bulk vertical wind difference | kinematic |
| SHEAR850 | Surface to 850 hPa bulk vertical wind difference | kinematic |

tures is significantly reduced, decreasing the computational time needed to train the RFs. For example, the original CSU-MLP feature assembly method (forecast-point relative method described above) contains \( N = m p (2r + 1)^2 \) features, where \( m \) is the number of meteorological variables (\( m = 12 \)), \( p \) is the number of forecast hours in the window (\( p = 9 \)), and \( r \) is the number of grid points used surrounding each training example (\( r = 3 \)); \( N \) is 5292 for the traditional CSU-MLP methods and 108 for the spatial-averaging procedure. The reduced number of features motivates an additional exploration into how preceding environments (i.e., the day(s) prior leading up to an event) may contribute to RF predictions and subsequent skill. For instance, it is well understood that moisture return from the Gulf of Mexico into the Great Plains can often precede a severe weather event, providing deeper boundary layer moisture and ample CAPE to support deep convection and severe weather (Johns and Doswell [1992]) – can the RFs learn to better predict severe weather by considering how the atmosphere becomes ‘primed’ before cyclogenesis in the lee of the Rockies? Furthermore, ensemble spread increases as a function of lead time, so considering GEFS/R ensemble median meteorological predictors on prior days for a day 6 forecast, for example, may yield additional confidence in a forecast outcome than considering only predictors spanning the day 6 forecast window. Therefore, three experiments are conducted that use the preceding 1, 2, and 3 days of features – hereafter referred to as p1, p2, and p3, respectively. For simplicity, only the surrounding environment near the training example is sampled, rather than employing a trajectory analysis to sample the upstream environment; such a method could be explored in future work. Using four total days of features (i.e., 3 preceding days and 1 valid over the forecast window) only
requires 396 features, an order of magnitude less than the traditional CSU-MLP method. For brevity, and since this method evaluation is exploratory in nature, the spatial averaging methodology, using 0–3 preceding days (i.e., p0, p1, p2, and p3), is only applied for day-6 model training and corresponding forecast evaluation.

Time-lagging methods have also been shown to be effective at creating forecast spread, yielding improved forecasts over their deterministic components and are competitive with multi-model and initial-condition perturbation ensembles at convective scales (e.g., Jirak et al. 2018, Wade and Jirak 2022). One specific benefit from time-lagging is no additional computation is required since the forecasts are already created and can be combined relatively easily to compute ensemble statistics. In this work, time-lagging is used to artificially increase the GEFS/R ensemble size,
ideally yielding a more representative ensemble median when assembling features. Both 10- and 15-member time-lagged ensembles are experimented with, which use the previous and 2 previous reforecast initializations, respectively, with features from each initialization valid over the same forecast window. As a result, the 10- and 15-member time-lagged RF models use 1 and 2 less days for training, discussed in the next subsection.

2) Training, Validation, and Testing

While the GEFS/R has 20 years of daily forecasts, only ~9 years are using for training the medium-range forecast models. This decision was largely to facilitate a comparison between day 1–3 models developed by Hill et al. (2020) and companion models trained with the GEFS/R, which are not explored in this work. Daily initializations of the GEFS/R from 12 April 2003–11 April 2012 are used to assemble predictors and severe weather reports are aggregated and encoded for this same period. It should be noted that since the first initialization is 12 April 2003, the first training examples are valid for 1200–1200 UTC 15–16 April 2003. 4-fold cross validation is employed over the 9 year period to select an optimal model as well as avoid overfitting the RFs. Testing of the trained RFs is conducted 2 October 2020–1 April 2022; the operational GEFSv12 was implemented in early October 2020.

RFs are trained for distinct regions of the country that represent somewhat unique regional climatologies of severe weather (Fig. 4). For each region, an RF is cross validated and the optimal model is selected based on minimizing the Brier Score across the four folds. After a fold is chosen, hyperparameters are varied to tune the skillful RF; skill of the trained RFs is more sensitive to the cross validation fold than the hyperparameters varied. Hyperparameters varied included the number of trees in the forecast and minimum number of samples needed to split a node. Entropy was set as the splitting criterion and a random selection of the features were evaluated at each node, equal to the square root of the total number of features. RFs trained with the alternative predictor assembly methods undergo the same cross-validation and hyperparameter tuning procedure.

Forecasts across the testing set are made using the GEFSv12 operational prediction system that consists of 21 ensemble members and 0.5 degree resolution. Forecasts are made with each region, optimized RF and the severe weather probabilities are stitched together with a smoothing function to limit discontinuities at regional borders and create a CONUS-wide forecast. As with training, only the 21-member ensemble median is used to generate real-time features, which feed into all RF versions to generate predictions. It should be noted that the GEFS/R dataset could have been used for testing between 2012 and 2019, but realtime forecasts are generated with the operational GEFSv12 and the authors felt the most appropriate evaluation should use products that SPC forecasters would be using in the future.

b. Verification

Traditional methods used to quantify probabilistic prediction skill – e.g., Brier Skill Score (BSS), area under the receiver operating characteristic curve (AUROC), and reliability diagrams – are employed to evaluate the CSU-MLP and SPC forecast. Observations of severe weather are obtained from the SPC archive of National Weather Service local storm reports since the SPC severe weather database was not updated through 2021 at the time the analysis was conducted. For a more direct comparison between continuous RF-generated forecasts and discrete SPC probabilities, the RF probabilities are discretized to resemble the outlooks issued by SPC forecasters. Discretization converts all RF probabilities within a probabilistic bin to the midpoint of the bin (e.g., Schumacher et al. 2021) – i.e., all probabilities below the 15% minimum SPC probability contour are set to 7.5, probabilities between 15% and 30% to 22.5%, and probabilities above 30% are set to 65%. The discretization procedure is also applied to the SPC contours. Additional discretized contours (e.g., a 5% contour from the RF forecasts) are introduced in the verification section to elucidate factors influencing RF forecast skill. While the continuous RF probabilities could be evaluated alongside interpolated SPC contours (e.g., Hill et al. 2020), which have been shown to be more skillful than discrete contours (Herman et al. 2018), the limited number of possible SPC contours at days 4–8 reduces the utility of interpolation; the discrete 15% and 30% SPC contours are retained for verification. All medium-range SPC outlooks evaluated herein were issued at ~0900 UTC each day and the shapefiles are converted to a gridded domain with ArcGIS as in Herman et al. (2018) as well as upscaled to NCEP Grid 4 for comparison to the RF-based forecasts.

The Brier Score (BS) is a measure of the mean squared error between binary events and probabilistic forecasts. The BS can be converted to a skill score, the BSS, by comparing the BS of a forecast to a reference climatology (e.g., Schumacher et al. 2021) – i.e., all probabilities below the 15% minimum SPC probability contour are set to 7.5, probabilities between 15% and 30% to 22.5%, and probabilities above 30% are set to 65%. The discretization procedure is also applied to the SPC contours. Additional discretized contours (e.g., a 5% contour from the RF forecasts) are introduced in the verification section to elucidate factors influencing RF forecast skill. While the continuous RF probabilities could be evaluated alongside interpolated SPC contours (e.g., Hill et al. 2020), which have been shown to be more skillful than discrete contours (Herman et al. 2018), the limited number of possible SPC contours at days 4–8 reduces the utility of interpolation; the discrete 15% and 30% SPC contours are retained for verification. All medium-range SPC outlooks evaluated herein were issued at ~0900 UTC each day and the shapefiles are converted to a gridded domain with ArcGIS as in Herman et al. (2018) as well as upscaled to NCEP Grid 4 for comparison to the RF-based forecasts.

The Brier Score (BS) is a measure of the mean squared error between binary events and probabilistic forecasts. The BS can be converted to a skill score, the BSS, by comparing the BS of a forecast to a reference climatology (e.g., Schumacher et al. 2021) – i.e., all probabilities below the 15% minimum SPC probability contour are set to 7.5, probabilities between 15% and 30% to 22.5%, and probabilities above 30% are set to 65%. The discretization procedure is also applied to the SPC contours. Additional discretized contours (e.g., a 5% contour from the RF forecasts) are introduced in the verification section to elucidate factors influencing RF forecast skill. While the continuous RF probabilities could be evaluated alongside interpolated SPC contours (e.g., Hill et al. 2020), which have been shown to be more skillful than discrete contours (Herman et al. 2018), the limited number of possible SPC contours at days 4–8 reduces the utility of interpolation; the discrete 15% and 30% SPC contours are retained for verification. All medium-range SPC outlooks evaluated herein were issued at ~0900 UTC each day and the shapefiles are converted to a gridded domain with ArcGIS as in Herman et al. (2018) as well as upscaled to NCEP Grid 4 for comparison to the RF-based forecasts.

The Brier Score (BS) is a measure of the mean squared error between binary events and probabilistic forecasts. The BS can be converted to a skill score, the BSS, by comparing the BS of a forecast to a reference climatology BS. BS ranges from -∞ to 1, with scores closer to 0 meaning the forecast skill is indistinguishable from the reference climatology (BS\text{ref}):

\[
BSS = 1 - \frac{BS_{\text{ref}}}{BS_{\text{ref}}}.
\]
The reference climatology used is a spatially and temporally smoothed long-term climatology of any severe weather reports across the CONUS. SPC severe weather database reports from 1990 to 2019 are gridded and Gaussian smoothers with 15-day temporal and 120-km spatial filters are used to create daily climatologies (e.g., Fig. 5) for the 30-year period (e.g., Brooks et al. 2003; Sobash and Kain 2017; Schumacher et al. 2021). BSSs are computed in aggregate (i.e., considering all forecast points for all days) and spatially (i.e., considering all forecasts for a particular point in space) to characterize SPC and RF forecast skill.

AUROC (e.g., Marzban 2004) measures forecast resolution and in this work the prediction system’s ability to discriminate severe weather from non-severe weather at various probabilistic thresholds. AUROC values of 0.5 suggest no discrimination, 0.7-0.8 is considered good, 0.8-0.9 great, and values $>0.9$ exceptional (Mandrekar 2010). Reliability diagrams are also used to characterize the relative forecast probability against the observed frequency of events, highlighting forecasts calibration at the various continuous probability contours and the discrete 15% and 30% SPC contours. Finally, the percent of forecast area covered by observations is computed to evaluate consistent biases in contour size. If a 15% contour frequently contains more than 30% fractional coverage of observations (i.e., the next contour interval), that issued contour is considered too small. Alternatively, if the fractional coverage is less than 15%, the contour is too large. Fractional coverage has been used extensively to evaluate probabilistic ML-based forecasts against human-generated outlooks (e.g., Erickson et al. 2019; Hill et al. 2020; Erickson et al. 2021; Hill et al. 2021; Schumacher et al. 2021).

Finally, the statistical relationships identified by the RFs are inspected to glean additional insights about the features that the models rely on to make predictions and how those relationships align with our physical understanding of severe weather forecasting. Feature importances (FIs) are computed and evaluated to assess the use of predictor information in each tree of the forest. Specifically, the "Gini importance" metric (e.g., Pedregosa et al. 2011; Herman and Schumacher 2018b; Whan and Schmeits 2018; Hill et al. 2020) is used to quantify the FIs. Each feature is assigned an importance value based on the number of times it is used to split a decision tree node, weighted by the number of training examples at the split (Friedman 2001). The importances are then summed over all splits and trees in the forest and can be aggregated to characterize temporal or spatial importance patterns (e.g., Hill et al. 2020). The higher the importance, the more value the RFs place on that predictor to make predictions. While other FI techniques exist (e.g., McGovern et al. 2019), the Gini importance metric is used here as an initial glance and not a holistic interrogation of FIs of the developed RFs; a follow-on manuscript is being prepared to fully interrogate model FIs with sufficient breadth and depth.

3. Verification Period Overview

a. Frequency of Severe Weather

Tornado event frequency highlights an active 1.5 years across the southeast U.S. (Fig. 6), likely attributable to two fall–winter seasons in the dataset; the climatological frequency in tornado activity across the southeast US has two peaks, one in the early fall and the other late winter/early spring (e.g., Horgan et al. 2007; Guyer and Dean 2010; Dixon et al. 2011; Gensini and Ashley 2011; Cintineo et al. 2012; Smith et al. 2012; Gensini et al. 2020). With only one full spring season of severe weather in 2021, limited tornadoes were reported across the Great Plains (e.g., Fig. 6). Unsurprisingly, hail reports are
primarily confined to the Great Plains and wind reports are more uniform across the U.S. (Fig. 5), east of the Rocky Mountains, compared to the other two hazards. Across the mid-Atlantic up into New England, multiple high-impact weather events produced numerous wind reports (Fig. 3c,d). These reports were anomalous compared to the long-term severe weather climatology (Fig. 5).

b. Frequency of Forecasts over extended period (day 4–8)

Day 4–7 15% forecasts from the CSU-MLP system and outlooks from SPC were issued across areas that experienced frequent severe weather, including the southeast US and to some extent the southern Great Plains out to day 7 (Fig. 7); 30% forecast contours are qualitatively similar, but omitted for brevity. The RF-based forecasts were issued much more frequently, however, compared to SPC. At day 7, where SPC issued only a handful of 15% contours across the CONUS (Fig. 7g), the RF issued forecasts in some point locations as many as 10 times (2% of the days; Fig. 7i). Moreover, the RFs forecast areas of severe weather across larger areas of the CONUS, covering nearly all states east of the Rockies at day 5 (Fig. 7h) when SPC limited their day 5 outlooks primarily south of 37° N (i.e., the northern border of Oklahoma).

Despite the relatively short verification period, and only one spring forecast season in the verification dataset, there is clearly seasonality to the issuance of SPC medium-range outlooks and CSU-MLP forecasts. SPC issued more 15% contours in the spring and fall seasons than summer and winter (Fig. 8). This pattern is not necessarily surprising given the climatological "double peak" of severe weather across the CONUS in the spring and fall (e.g., Smith et al. 2012). In contrast, the CSU-MLP had a nearly uniform distribution of 15% contours across the months for days 4–6 (Fig. 8), with perhaps a slight peak in frequency across the summer months – 15% contours at days 7 and 8 were more common between January and August. SPC issued only a handful of 30% contours in the months of March and October (Fig. 8c) whereas the CSU-MLP issued a number of higher probability contours through the spring months, primarily at days 4 and 5 (Fig. 8d), highlighting the RF-based system’s confidence in forecasting both predictable and less predictable (i.e., warm season) severe weather regimes.

4. Forecast Verification

RF and SPC forecast skill is first evaluated in aggregate across the entire verification period. Control RF (i.e., CSU-MLP) forecast skill decreases with increasing lead time, demonstrating significantly better skill than SPC outlooks at days 4 and 5 and equally near-climatological skill beyond day 5 (Fig. 9i). As lead time increases, the RFs are confronted with learning how GEFS/R environments (specifically, the median environments) associate with severe weather as forecast variability and ensemble variance also increases. The limited number of GEFS/R ensemble members likely prohibits a proper depiction of all future atmospheric states and forces the RFs to "learn" relationships between severe weather events and simulated environments that may not be conducive to severe weather. As a result, the ability of the RFs to discriminate events from non-events similarly decreases with increasing lead time, with AUROC falling to ~0.5 by day 8 (Fig. 9j); the AUROC is highest for day 4 at 0.62. However, none of the AUROC values eclipse the 0.7 mark denoting good resolution (Mandrekar 2010). On the other hand, when 5% probability contours are included in the RF-forecast discretization process and retained in the AUROC calculation, reliability is increased substantially to >0.07 at day 4 and is significantly larger than 0 (i.e., climatology) at day 8. Meanwhile, the AUROC surpasses 0.8 at day 4 and resolution is nearly ‘good’ at day 8. While the higher probability 15 and 30% contours do not have significant skill beyond day 5 (or good resolution at any forecast lead time), the adjusted skill metrics resulting from including 5% probability contours demonstrates that low-probability forecast contours may have tremendous value for forecasting severe weather out to day 8.

The skill and resolution of forecasts derived from alternatively-trained RFs is also assessed to determine if medium-range prediction skill can be improved by learning to associate severe weather events with features in the days leading up to events (p0, p1, p2, and p3 experiments), reducing the impact of noisy predictors, and increasing the representative sample of atmospheric states in the underlying GEFS/R ensemble (t10 experiment) used to assemble associated features. Aggregated BSSs are computed over the same verification period for the p0, p1, p2, p3, and t10 experimental forecasts at the day 6 lead time (Fig. 10a) with the 5% contour included for comparison against the best control CSU-MLP forecasts (e.g., Fig. 9j). The forecast skill for all predictor-averaged models is statistically indistinguishable from the control CSU-MLP RF model (Fig. 10a). These BSSs alone imply that reasonably skillful day-6 forecasts can be derived by simply considering how the atmosphere evolves locally before a high-impact severe weather event. Moreover, the relatively equal skill amongst forecasts suggests that the raw GEFS/R predictors used in the control model (e.g., at each point in space) do not add significant value in training and perhaps exhibit less noise than their CAM-model counterparts (e.g., Hill et al. 2021; Loken et al. 2022). The time-lagged model forecasts exhibit significantly less skill than the control and predictor-averaged model forecasts, but their resolution is significantly better (Fig. 10b) suggesting that the time-lagged RF model is issuing probability contours that are larger than the corresponding predictor-averaged model
Fig. 5. (a) Daily climatological any severe hazard probability (%) centered on 21 May and (b) mean climatological hazard probability (%) for any severe weather hazard.

Fig. 6. Frequency of (a) tornado, (b) hail, (c) wind, and (d) all severe hazard reports across the verification period.

forecast contours (i.e., improved resolution), but sacrificing skill (not shown). Furthermore, the predictor-averaged model forecasts all improve upon the resolution of the con-
control system, with p0 containing the highest resolution (Fig. 10b).

An assessment of spatial forecast skill underscores the frequency biases of the RF-based and SPC forecasts. The SPC outlooks feature patchy areas of positive skill associated with instances of severe weather that were correctly highlighted days in advance (Figs. 11a,b); in other words, when SPC forecasters do issue outlooks, they do so quite...
skillfully, particularly at day 4. However, areas of slightly negative skill in the southern Great Plains and mid-Atlantic suggest there were missed opportunities to forecast high-impact events 4 and 6 days in advance. Fewer SPC forecasts at day 6 (Fig. 7e) further limits the extent of positive forecast skill (Fig. 11b) compared to day 4. It is possible that forecasters had little confidence in the forecast evolution to warrant a 15% or 30% outlook contour at these lead times for various high-impact weather events, but as lead time decreased (i.e., days 1–3), confidence increased and forecasts were issued at multiple probabilistic thresholds. Furthermore, the limited verification dataset likely creates localized areas of high/low skill, and increasing the length of verification to multiple years would help to clarify SPC forecast skill in the medium range.
The spatial skill of control RFs similarly emphasizes the high-frequency forecast bias with more expansive and smoothed areas of positive and negative BSSs (Fig. [11]c,d). For brevity and since the predictor-averaged models display similar skill, only the control forecast spatial skill is considered here. Prominent areas of negative forecast skill in the control RF forecasts across the Great Lakes is a symptom of the "forecast anywhere" nature of the RFs regardless of the climatological report frequency. BSS differences between SPC outlooks and RF forecasts (Fig. [11]e,f) further illustrate the complexities of verifying these forecasts on a short period, but also clearly demonstrate that issuing more probabilistic contours in the medium-range yields better skill. This facet is perhaps most notable across the upper Midwest and mid-Atlantic, where the RF forecasts at days 4 and 6 were notably better than SPC as a result of SPC not issuing many outlooks (Figs. [7]e). The CSU-MLP spatial forecast skill at days 4 and 6 is improved when the 5% probability contour is included (Fig. [12]), which amplifies areas of positive and negative skill. In some instances, including the 5% forecast contour reverses areas of negative skill (c.f. Figs. [11] and [12] in Georgia) or replaces neutral skill with positive skill as in the Great Plains (c.f. Figs. [11] and [12] in North and South Dakota).

Forecast skill is also assessed by computing the fractional coverage of observations within the forecast contours. While the SPC contours are defined as single probabilistic levels and not a discrete range (e.g., 15–30%), it is reasonable to suggest that the fractional coverage of observations for a probabilistic contour should not exceed the next probabilistic value, otherwise a higher contour would be warranted. Fig. [13] shows the fractional coverage of observations, in which the objective is to be within the green and red horizontal bars for a particular probabilistic threshold. When below the green horizontal line, a forecast contours are believed to be too large on average, and when above the red line, forecast areas are too small. SPC and CSU-MLP control forecasts are well calibrated at the 15% threshold at almost all forecast days; SPC forecasts are potentially too small at day 8, but a small sample size limits a complete analysis for that lead time (Fig. [13]a). On the other hand, the day 4 and 5 30% outlooks from SPC are typically smaller than the CSU-MLP control forecasts, which appear generally well-calibrated prior to day 7 (Fig. [13]b). Unsurprisingly, the CSU-MLP 5% probability contours are also well-calibrated at all forecast lead times (not shown).

For a complete depiction of calibration, reliability diagrams are constructed for all SPC and control RF forecasts (Fig. [14]). Reliability curves that fall above or below the perfect-reliability line, dashed black lines in both panels of Fig. [14] are said to under-forecast or over-forecast severe weather events, respectively. All SPC outlooks are generally well-calibrated prior to day 8, whereas RF forecasts generally have an under-forecast bias above the 15% probability threshold but achieve reliability for lower thresholds. Day 7 and 8 control RF forecasts lose reliability quickly after 15% (Fig. [14]b), plummeting effectively to no skill above 30% and no resolution (i.e., below the red dashed line) at the highest probabilities considered – SPC maintains skill at day 7 but underforecasts severe weather events at day 8 (Fig. [14]b). While reliability for days 4–6 forecasts above 30% is considerably more variable (Fig. [14]a), owing to smaller sample sizes (e.g., inset figure in Fig. [14]), forecast skill still hovers near perfect reliability. Analysis of reliability for the alternative predictor assembly methods reveals that predictor-averaging maintains reliability relative to the control forecasts up to slightly higher probabilistic thresholds (e.g., 25%) for the day 6 forecasts whereas the day 6 110 experiments overforecast observed events (not shown), aligning with the aggregate BSS statistics (Fig. [10]h). This analysis underscores the utility of the

![Brier Skill Score](image1.png)

![Area under the ROC](image2.png)

**Fig. 10.** Aggregate (a) Brier Skill Score and (b) Area under the ROC (AUROC) curve for CSU-MLP and experimental (p0, p1, p2, p3, and t110) RF forecasts. Error bars are computed from 100 bootstrap resamples of the forecast distributions and represent the 95% confidence interval. BSS and AUROC are computed using 5, 15, and 30% probability contours.
continuous RF probabilities as a forecast tool at medium-range lead times, and also stresses the difficulty of accurately capturing severe weather threat areas multiple days in advance (i.e., most RF-based forecasts underforecast the observed event probabilities), even for skillful statistical models.

a. Example Forecasts

Example forecasts are provided that display some of the skill and resolution attributes of the RF outlooks described previously (Fig. 15). All day 4–8 example forecasts are valid for the 24-hour period ending 1200 UTC 16 December 2021. This particular event featured a compact short-wave trough with strong low- and mid-level wind
12. As in Figs. 11c,d, but BSS is computed with a discretized 5% contour.

13. Spatial coverage of local storm reports in SPC and CSU-MLP forecasts for all forecast days and the (a) 15% and (b) 30% probability contours. Green horizontal lines denote the bottom probability of the forecast "bin", whereas red horizontal lines denote the top of the bin, i.e., the next probability contour value.

fields. Robust low-level warm air advection with low 60s dewpoints contributed to an atmosphere primed for severe thunderstorms and a highly anomalous severe-weather event for mid-December; a robust discussion of the mete-
Medium-range forecasts from NWP models depicted the shortwave trough days in advance, but did not have a good handle on the instability parameter space even at day 4. SPC forecasters decided to issue a 5% severe hazard risk at day 3, noting the impressive kinematic support for damaging wind gusts. Increasing forecaster confidence in a high-impact severe weather event with decreasing lead time resulted in a moderate categorical risk being issued at day 1, with a 45% probability of severe wind and a large area of significant severe wind delineated; a 10% probability of tornadoes also accompanied the SPC day-1 forecast.

The CSU-MLP control forecasts (Fig. 15) depicted a severe weather threat area across the upper Mississippi valley eight days in advance (Fig. 15e). By day 6, a 15% probability contour was introduced in the forecasts with a 30% contour added in subsequent, shorter lead-time forecasts (Fig. 15a-c) that mostly encircled severe weather reports for the event. Forecast skill scores (BSS) gradually increased from 0.04 at day 8 to 0.19 at day 5 with a slight decrease back to 0.15 at day 4. In short, the probabilistic RF-based guidance showed substantial skill out to day 8 for this particular case, and the progression of forecasts from day 8 to 4 showcases the utility of the forecast system to highlight areas that may experience severe weather.

Another example is provided to reinforce the similarities between the experimental forecast systems for one particular case. On this day, the 24-hour period ending 1200 UTC 14 August 2021, numerous wind reports were recorded across Ohio, Pennsylvania, West Virginia, and numerous other mid-Atlantic states (Fig. 16). All RF-based forecasts have a broad 5% contour across this region, and erroneously westward into Indiana, Illinois, and Missouri. None of the forecasts suggest greater than 15% probability of severe weather, despite dense sets of wind reports in two corridors across the northeast. The 15% contour in the control system seems subjectively well-positioned (BSS=0.1076), but the p0-model forecast inaccurately extends the 15% contour westward (BSS=0.1064). The p1-model forecast contracts the 15% contour back eastward, but also eliminates an area in New York and Pennsylvania that experience wind reports (BSS falls to 0.0979). The p2- and p3-model forecasts further contract the 15% contour (BSSs of 0.059 and 0.0441, respectively), leaving the 5% area relatively unchanged; the time-lagged model is nearly identical to the p3-model forecast. This example illustrates the rather subtle differences between the experiments that renders the control system and flow-dependent RF models objectively similar. A more comprehensive case-study evaluation would be needed to characterize these subtle forecast differences over the entire forecast period, which as the event neared. See note from SPC forecaster Andrew Lyons: https://twitter.com/TwisterKidMedia/status/1471585397444874337?s=20&f=tw

---

3 Day 3 discussion available at https://www.spc.noaa.gov/products/outlook/archive/2021/day3otlk_20211213_0830.html

4 SPC forecasters used the CSU-MLP forecasts during this event, highlighting their value in upgrading SPC outlooks as the event neared. See note from SPC forecaster Andrew Lyons: https://twitter.com/TwisterKidMedia/status/1471585397444874337?s=20&f=tw

---
is beyond the scope of this experimental exploration but is an active area of research. Additionally, the computational savings of the spatial-averaging models, particularly in training the RFs, may support a continued investigation alongside the CSU-MLP control forecast system into their utility as an operational tool.

**b. Feature Importances**

To better understand what the RFs have learned about severe weather prediction from the training process, FIs are aggregated by meteorological variable and region for the day 4, 6, and 8 CSU-MLP control models (Fig. [17]). Consistent with day 1-3 models developed by Hill et al. (2020), CAPE, CIN, MSLP, SHEAR500, and SHEAR850 are the most importance predictors for the day 4 models.
Fig. 16. RF-based forecasts of any severe hazard at day 6 lead time from (a) trad, (b) p0, (c) p1, (d) p2, (e) p3, and (f) tl10. All forecasts are initialized 0000 UTC 8 August 2021 and valid 1200 UTC 13 August 2021 – 1200 UTC 14 August 2021. NWS local storm reports for wind, hail, and tornadoes are included as blue, green, and red circles, respectively. Observation coverage and BSS are included in lower left and right corners of each panel, respectively.

(Fig. 17a); CAPE is also less important in the East region as MSLP and SHEAR850 increases in importance, consistent with high-shear low-CAPE environments that are more prevalent in the southeast U.S. (Sherburn and Parker 2014). As lead time increases, CAPE becomes less important in the Central region (Fig. 17b,c), being replaced by Q2M. Q2M also becomes more important in the West region, but CIN replaces CAPE as the most important predictor (e.g., Fig. 17d). In the East, CIN also replaces CAPE as the most important predictor.

FIs of the p1 and p2 models are explored to further understand how the RFs leverage features in the day(s) leading up to severe weather events. FIs for both models peak during the day 6 forecast window (i.e., forecast hours...
132–156), but they also ramp up from the days leading up to the event (Fig. 18), suggesting the local meteorological environment is being used by the RFs in predictions. Slight differences in FIs exist by regional model as well. For example, the West region p1 and p2 models have secondary peaks near forecast hours 120–123 (Fig. 18a), approximately 00–03 UTC the day before the forecast window, and p2 has a tertiary peak near forecast hour 99 (Fig. 18b). Not only are the prior days variables being used, but there is a notable cyclical nature that matches the diurnal climatology of severe weather (e.g., Hill et al. 2020). In contrast, FIs in the Central and East regions do not have the same cyclic pattern, but rather have a nearly constant ramp up in FIs (e.g., orange and yellow bars in Fig. 18a). However, since these FIs are a summation over all meteorological predictors, it is not clear what aspects of the environment prior to a severe weather event are being learned by the RFs to make day 6 predictions.

To further clarify the day-prior FIs, features and FIs are separated into thermodynamic and kinematic subgroups (refer to Table 1) for the p1 models (Fig. 19). In the West region models, which exhibited a strong cyclic FI pattern, the thermodynamic variables (e.g., CAPE, Q2M) are the primary contributor to the day-prior FI secondary peak, with a sharp increase at forecast hour 123 (Fig. 19a) – the kinematic variable FIs in the West p1 model have a more subtle cyclic pattern, but still peak during the day 6 forecast window. In the Central region model, the FIs have a broad and uniform peak from forecast hours 135–147 and a smaller contribution compared to the West region in the day prior period (Fig. 19b). The East region models lean on day-prior thermodynamic predictors slightly more than the Central region models, with a longer ramp-up of FIs from forecast hours 123–141 (Fig. 19c), but the kinematic FIs for both the Central and East models are markedly smaller than the thermodynamic variable contributions. A full explanation for these FI patterns is reserved for future work, but as an initial assessment, the FIs highlight unique regional and meteorological relationships learned by the experimental models that were exploited by the RFs to make severe weather predictions. Changes to the local environment that preceeded severe weather events
clearly influenced the RFs during training, but to what extent those variables contributed to forecast probabilities is not discernible from Gini importances alone.

5. Summary and Discussion

Nine years of reforecasts from the GEFSv12 reforecast dataset are used along with historical records of severe weather to construct a novel RF prediction system capable of explicitly and probabilistically predicting severe weather at 4–8 day lead times, i.e., the medium range. Human forecasts issued by the SPC are evaluated alongside the RF-based predictions to assess the operational utility of the ML forecasts. A handful of experiments are also conducted to explore whether forecasts could be improved through feature engineering and expanding the GEFSv12 ensemble size. The main conclusions are as follows:

1. RF forecasts have more skill and higher resolution than the human-based outlooks, which is partly a reflection of the continuous probabilities of the RFs and their ability to issue lower probability contours more frequently that add considerable skill and resolution to the forecast system.

2. The CSU-MLP forecasts tend to underforecast the occurrence of severe weather in the medium range at probabilistic contours above 15% whereas SPC forecasts are calibrated prior to day 8.

3. Using spatially-averaged GEFS/R features yielded similarly skillful forecasts as the traditional CSU-MLP method while also allowing for prior-day meteorological information to inform the forecasts; the models learned to associate the buildup of thermodynamically- and kinematically-favorable environments with next-day severe weather events. Additionally, the similar skill amongst models suggests that spatiotemporally-displaced GEFS/R predictors are not particularly noisy but do not provide tremendous value to the RFs.

4. Time-lagging the GEFSv12 reforecasts to produce larger initial ensembles for RF training degraded forecast skill but increased forecast resolution by generating larger areas of low probability forecasts.

5. Feature importances revealed relationships known to be important for severe weather forecasting, providing confidence in the RF forecasts.
6. The performance of the RF-based forecasts alongside the human-generated outlooks demonstrate their utility and potential value as a guidance tool in the human forecast process.

The comparisons between RF-based predictions and human forecasts provided in this work have some important caveats to consider. SPC forecasters have often employed specific philosophies in generating day 4–8 outlooks [Steven Weiss, personal communication] that likely limit the number of forecasts issued and hamper more skillful human-based medium-range forecasts. First, SPC forecasters are tasked with forecasting the probability of “organized severe thunderstorms”, not necessarily severe weather reports, so they will never outline a high CAPE, low shear event in the medium range despite a high likelihood of thunderstorms. Second, SPC forecasters are very concerned with continuity and ramping up to an event. For example, forecasters may opt to introduce a forecast area in a day 3 or 2 outlook rather than days 4–8 because they may not want to highlight a threat area that has to be shifted, enlarged, or removed altogether in subsequent outlooks – forecasters are hesitant to add outlook areas when confidence is too low, and they can add it in on the next forecast shift. Relatedly, SPC forecaster perception is that NWS Weather Forecast Offices do not like when SPC removes
or reduces severe weather probabilities because it affects public messaging of severe weather. As a result, it is often common for SPC day 4–8 outlooks to be relatively small (e.g., Fig. 13) and infrequent, particularly when atmospheric predictability for severe weather wanes in the warm season (e.g., Fig. 8). As confidence increases in a severe weather threat area, the probabilities can be increased and area expanded in day 1–3 outlooks. These and other internal constraints, along with the relative dearth of useful NWP model guidance, naturally restrict SPC forecast skill at longer lead times. On the other hand, the ML-guidance generated from the CSU-MLP could significantly aid SPC by increasing confidence and consistency to provide more lead time to operational partners and end users to the threat of severe weather.

The results presented highlight some ML success against baseline forecasts and also a number of unique avenues that could be explored moving forward to enhance and improve both ML-based guidance and the SPC human-based forecasts, as well as increase interpretability of the ML ‘black box’. While the feature assembly experiments (e.g., p2, t10) did not yield forecasts that surpassed the skill of the traditional CSU-MLP system, the simplification of features could be exploited to include other ensemble diagnostic or summary metrics (e.g., mean, high or low member values) that characterize ensemble spread into the medium range. The meteorological predictors could also be varied in any of the ML configurations to explore which predictors add the most value, or objective methods (e.g., permutation importance) could be used to reduce feature redundancy and select a more optimal subset of features. It will also be vitally important that alternative interpretability metrics (e.g., tree interpreter [Saabas 2016], Shapely additive values [SHAP; Shapley 2016], accumulated local effect [ALE; Apley and Zhu 2020]) are employed to interrogate how the RFs make predictions; this exploration is underway and will be the focus of a follow-on manuscript. Additionally, the added benefit of the ML system against the underlying GEFS model could be quantified more explicitly. Traditionally, ML-based forecasts have been measured against the very model that generates the ML predictors, with demonstrated success improving upon the raw dynamical models (e.g., [Herman and Schumacher 2018a]). In this instance, with notable 2-m dry and low-instability biases in the GEFSv12 system [Manikin et al. 2020], it would be informative to quantify the value added by the ML system to correct for these biases when making severe weather predictions.

Equipped with calibrated statistical products and expert human knowledge, SPC forecasters may be able to increase medium-range outlook skill by using the CSU-MLP RF forecasts and delineating lower-probability threat areas. Furthermore, by incorporating these types of robust, skillful statistical guidance products into the forecast process, SPC forecaster confidence in a forecast outcome may increase. Being able to unveil how and why the ML models are issuing probabilities in a certain area will provide additional confidence to forecasters to rely on the products as a forecast tool. We expect that the usefulness of the CSU-MLP prediction system and others like it is not necessarily limited to the medium-range either, and the applicability to subseasonal or seasonal predictions of severe weather is planned for future investigation. Additionally, with continued effort from the meteorological community to explain AI/ML methodologies (e.g., Chase et al. 2022) with comprehension, and to make these tools more common in academic settings, there will be more opportunities to pursue new and improved forecast methodologies. Finally, it is crucial that constant communication exists between ML developers and SPC forecasters to generate products that SPC operations finds useful and valuable. One such avenue includes continued participation and development of these ML products in the Hazardous Weather Testbed Spring Forecast Experiment [Clark et al. 2021].

---

4Internal SPC surveys have suggested these biases exist and are reducing forecaster confidence in deterministic Global Forecast System and GEFSv12 forecasts.
Acknowledgments. This work is supported by the Joint Technology Transfer Initiative and NOAA Grant NA200AR4590350. We would like to thank SPC forecasters for their invaluable perspectives about these forecast products and continued collaboration to develop cutting edge medium-range guidance products.

Data availability statement. All RF-based forecasts are available upon request from the Colorado State University researchers, and will be made available in the near future in an online repository. SPC outlooks are available via a public archive at https://www.spc.noaa.gov/. The GEFS/R dataset and GEFSv12 forecasts are publicly available from Amazon AWS at https://registry.opendata.aws/noaa-gefs/.

References
Apley, D. W., and J. Zhu, 2020: Visualizing the effects of predictor variables in black box supervised learning models. *J. Roy. Stat. Soc.*, **82**, 1059–1086, https://doi.org/10.1111/rssa.12377.

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

Bright, D. R., and J. S. Grams, 2009: Short range ensemble forecast (SREF) calibrated thunderstorm probability forecasts: 2007-2008 verification and recent enhancements. *Fourth Conference on the Meteorological Applications of Lightning Data*, Phoenix, AZ, 63.

Brooks, H. E., C. A. Doswell, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the united states. *Weather and Forecasting*, **18**, 626 – 640, https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2.

Burke, A., N. Snook, D. J. G. II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning–based probabilistic hail predictions for operational forecasting. *Weather and Forecasting*, **35**, 149 – 168, https://doi.org/10.1175/WAF-D-19-0105.1.

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.

Clark, A. J., and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bulletin of the American Meteorological Society*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Dixon, P. G., A. E. Mercer, J. Choi, and J. S. Allen, 2011: Tornado risk analysis: Is dixie alley an extension of Tornado alley. *Bull. Amer. Meteor. Soc.*, **92**, 433–441, https://doi.org/10.1175/2010BAMS3102.1.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the weather research and forecasting (wrf) model. *Atmos. Sci. Lett.*, **5**, 110–117.

Erickson, M., J. Kastman, B. Albright, S. Perfater, J. Nelson, R. Schumacher, and G. Herman, 2019: Verification results from the 2017 HMT–WPC flash flood and intense rainfall experiment. *J. Appl. Meteor. Climatol.*, **584**, 2591–2600.

Erickson, M. J., B. Albright, and J. A. Nelson, 2021: Verifying and redefining the weather prediction center’s excessive rainfall outlook forecast product. *Wea. Forecasting*, **36**, 325–340.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system. *Monthly Weather Review*, in press, https://doi.org/10.1175/MWR-D-20-0194.1.

Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **1189–1232*, https://doi.org/10.1214/asn1013203451.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840.

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

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL–WRF ensemble forecasts. *Weather and Forecasting*, **33**(2), 443 – 460, https://doi.org/10.1175/WAF-D-17-0132.1. URL https://journals.ametsoc.org/view/journals/wefo/33/2/waf-d-17-0132_1.xml.

Gensini, V., and W. Ashley, 2011: Climatology of potentially severe convective environments from the north american regional reanalysis. *Electronic Journal of Severe Storms Meteorology*, **6**.

Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the united states using ERA5 proximity soundings. *Weather and Forecasting*, **36**, 2143 – 2160, https://doi.org/10.1175/WAF-D-21-0056.1.

Gensini, V. A., A. M. Haberlie, and P. T. Marsh, 2020: Practically Perfect Hindcasts of Severe Convective Storms. *Bulletin of the American Meteorological Society*, **101**, E1259–E1278, https://doi.org/10.1175/BAMS-D-19-0321.1.

Guyer, J. L., and A. R. Dean, 2010: Tornadoes Within Weak CAPE Environments Across the Continental United States. *25th Conference on Severe Local Storms*, 1.5.

Hamill, T. M., E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. national blend of models for statistcal postprocessing of probability of precipitation and deterministic precipitation amount. *Monthly Weather Review*, **145**, 3441 – 3463, https://doi.org/10.1175/MWR-D-16-0331.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, **134**, 3209 – 3229, https://doi.org/10.1175/MWR3237.1.

Hamill, T. M., and Coauthors, 2022: The reanalysis for the global ensemble forecast system, version 12. *Mon. Wea. Rev.*, **150**, 59 – 79, https://doi.org/10.1175/MWR-D-21-0023.1.

Harrison, D. R., M. S. Elliot, I. L. Jirak, and P. T. Marsh, 2022: Utilizing the high-resolution ensemble forecast system to produce calibrated probabilistic thunderstorm guidance. *Weather and Forecasting*, **37**, 1103 – 1115, https://doi.org/10.1175/WAF-D-22-0001.1.

Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of storm prediction center convective outlooks. *Weather and Forecasting*, **33**, 161 – 184, https://doi.org/10.1175/WAF-D-17-0104.1.
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.

Herman, G. R., and R. S. Schumacher, 2018b: “dendrology” in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Monthly Weather Review*, **146**, 1785 – 1812.

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

Hill, A. J., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Wea. Forecasting*, **36**, 1693–1711, https://doi.org/10.1175/WAF-D-21-0026.1.

Hill, A. J., C. C. Weiss, and D. C. Dowell, 2021: Influence of a portable near-surface observing network on experimental ensemble forecasts of deep convection hazards during VORTEX-SE. *Wea. and Forecasting*, **36**, 1141 – 1167, https://doi.org/10.1175/WAF-D-20-0237.1.

Horgan, K. L., D. M. Schultz, J. E. Hales Jr., S. F. Corfidi, and R. H. Johns, 2007: A five-year climatology of elevated severe convective storms in the united states east of the rocky mountains. *Wea. Forecasting*, **22**, 1031 – 1044.

Jahn, D. E., I. Jirak, A. Wade, and J. Milne, 2022: Storm mode and tornado potential determination using statistical moments of updraft helicity distribution. 27th Conf. on Numerical Weather Pred., Houston, TX, Amer. Meteor. Soc.

Jirak, I., B. Roberts, B. Gallo, and A. Clark, 2018: Comparison of the HRRR time-lagged ensemble to formal CAM ensembles during the 2018 HWT spring forecasting experiment. 29th Conference on Severe Local Storms, Stowe, VT, poster 123.

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

Kain, J. S., and Coauthors, 2008: Severe-weather forecast guidance from the first generation of large-domain convection-allowing models: Challenges and opportunities.

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

Loken, E. D., A. J. Clark, and A. McGovern, 2022: Comparing and interpreting differently-designed random forests for next-day severe weather hazard prediction. *Wea. and Forecasting*, https://doi.org/10.1175/WAF-D-21-0138.1.

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

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences*, **26**, 636 – 646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

Mandrekar, J. N., 2010: Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, **5**, 1315–1316, https://doi.org/10.1097/JTO.0b013e3181ec173d.

Manikin, G., A. Bentley, L. Dawson, and S. Shields, 2020: Model evaluation group (MEG) GEFSv12 evaluation overview. [Available online at https://www.emc.ncep.noaa.gov/users/meg/gefsv12/ppt/MEG_4-23-20_GEFSv12_Overview.pptx], accessed 3 August 2022.

Marzban, C., 2004: The roc curve and the area under it as performance measures. *Weather and Forecasting*, **19**, 1106 – 1114, https://doi.org/10.1175/825.1.

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

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergenson, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100**, 2175—2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

NCEI, 2022: NOAA National Centers for Environmental Information (NCEI) U.S. billion-dollar weather and climate disasters (2022). [Available online at https://www.ncdc.noaa.gov/billions/], accessed 19 January 2022, https://doi.org/10.25921/stkw-7w73.

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

Saabas, A., 2016: Random forest interpretation with scikit-learn. [Available online at https://blog.datadive.net/random-forest-interpretation-with-scikit-learn/], accessed 13 July 2022.

Schumacher, R. S., A. J. Hill, M. Klein, J. Nelson, M. Erickson, and G. R. Herman, 2021: From random forests to flood forecasts: A research to operations success story. *Bull. Amer. Meteor. Soc.*, **102**, E1742–E1755.

Shapley, L. S., 2016: A value for n-person games. *Contributions to the Theory of Games (AM-28), Volume II*, T. editor, Ed., Princeton University Press, chap. 8, 307–318.

Sherburn, K. D., and M. D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-cape environments. *Weather and Forecasting*, **29**, 854–877, https://doi.org/10.1175/WAF-D-13-00041.1.

Shield, S. A., and A. L. Houston, 2022: Diagnosing supercell environments: A machine learning approach. *Weather and Forecasting*, in press, https://doi.org/10.1175/WAF-D-21-0098.1.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous united states. part i: Storm classification and climatology. *Weather and Forecasting*, **27**, 1114–1135.

Sobash, R. A., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Weather and Forecasting*, **32**, 1885 – 1902, https://doi.org/10.1175/WAF-D-17-0043.1.
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. *Weather and Forecasting*, 26, 714 – 728, https://doi.org/10.1175/WAF-D-10-05046.1.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A Comparison of Neural-Network and Surrogate-Severe Probabilistic Convective Hazard Guidance Derived from a Convection-Allowing Model. *Wea. Forecasting*, 35, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Sobash, R. A., G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016a: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Weather and Forecasting*, 31, 1591 – 1614, https://doi.org/10.1175/WAF-D-16-0073.1.

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

SPC, 2020: About the SPC. [Available online at https://www.spc.noaa.gov/misc/aboutus.html], accessed 28 April 2022.

SPC, 2022a: Dec 15, 2021 0600 UTC day 1 convective outlook. [Available online at https://www.spc.noaa.gov/products/outlook/archive/2021/day1otlk_20211215_1200.html], accessed 6 May 2022.

SPC, 2022b: Storm Prediction Center WCM page. [Available online at https://www.spc.noaa.gov/wcm/], accessed 12 January 2022.

Wade, A. R., and I. L. Jirak, 2022: Exploring hourly updating probabilistic guidance in the 2021 spring forecasting experiment with objective and subjective verification. *Weather and Forecasting*, 37, 699 – 708, https://doi.org/10.1175/WAF-D-21-0193.1.

Whan, K., and M. Schmeits, 2018: Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Mon. Wea. Rev.*, 146, 3651–3673.

Zhang, F., N. Bei, R. Rotunno, C. Snyder, and C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, 64, 3579–3594.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, 60, 1173–1185.

Zhou, X., and Coauthors, 2022: The development of the NCEP global ensemble forecast system version 12. *Weather and Forecasting*, https://doi.org/10.1175/WAF-D-21-0112.1.