Diagnosing Supercell Environments: A Machine Learning Approach

STEPHEN A. SHIELD[a] AND ADAM L. HOUSTON[a]

[a] *Department of Earth and Atmospheric Sciences, University of Nebraska–Lincoln, Lincoln, Nebraska*

(Manuscript received 15 June 2021, in final form 3 March 2022)

ABSTRACT: The importance of discriminating between environments supportive of supercell thunderstorms and those that are not supportive is widely recognized due to significant hazards associated with supercell storms. Previous research has led to forecast indices such as the energy helicity index and the supercell composite parameter to aid supercell forecasts. In this study three machine learning models are developed to identify environments supportive of supercells: a support vector machine, an artificial neural network, and an ensemble of gradient boosted trees. These models are trained and tested using a sample of over 1000 Rapid Update Cycle version 2 (RUC-2) model soundings from near-storm environments of both supercell and nonsupercell storms. Results show that all three machine learning models outperform classifications using either the energy helicity index or supercell composite parameter by a statistically significant margin. Using several model interpretability methods, it is concluded that generally speaking the relationships learned by the machine learning models are physically reasonable. These findings further illustrate the potential utility of machine learning–based forecast tools for severe storm forecasting.

SIGNIFICANCE STATEMENT: Supercell thunderstorms are a type of thunderstorm that are important to forecast because they produce more tornadoes, hail, and wind gusts compared to other types of thunderstorms. This study uses machine learning to create models that predict if a supercell thunderstorm or nonsupercell thunderstorm is favored for a given environment. These models outperform current methods of assessing if a storm that forms will be a supercell. Using these models as guidance forecasters can better understand and predict if atmospheric conditions are favorable for the development of supercell thunderstorms. Improving forecasts of supercell thunderstorms using machine learning methods like those used in this study has the potential to limit the economic and societal impacts of these storms.

KEYWORDS: Atmosphere; Convection; Storm environments; Supercells; Classification; Neural networks; Forecasting techniques; Decision trees; Machine learning; Support vector machines

1. Introduction

Forecasting convective storms is important because the severe weather they produce has far reaching impacts on society. Convective storm hazards such as flooding, tornadoes, wind, and hail can have significant economic impacts on industries including insurance (e.g., Sander et al. 2013), electrical power (e.g., Shield et al. 2021), and aviation (e.g., NTSB 2010). Supercells are responsible for a disproportionate amount of severe weather compared to other convective storm morphologies (Duda and Gallus 2010), and thus, accurate supercell forecasts are critical for decision support and hazard mitigation.

A number of studies have utilized sounding climatologies to characterize the environments supportive of severe convective weather (e.g., Rasmussen and Blanchard 1998; Rasmussen 2003; Thompson et al. 2003, 2007; Houston et al. 2008). Two composite indices developed through this prior research are used widely by forecasters: the energy helicity index (EHI; Hart and Korotky 1991; Rasmussen 2003) and the supercell composite parameter (SCP; Thompson et al. 2003, 2004, 2007; Gropp and Davenport 2018).

Following Hart and Korotky (1991), EHI combines convective available potential energy (CAPE) and storm-relative helicity (SRH) over a layer of the atmosphere typically either 0–1 or 0–3 km:

$$\mathrm{EHI} = \frac{\mathrm{CAPE} \times \mathrm{SRH}}{160\,000}, \tag{1}$$

where CAPE is in joules per kilogram (J kg^{-1}), SRH is in meters squared per second squared (m^2 s^{-2}), and the normalization factor has units that render EHI unitless. EHI has been shown to be effective in discriminating between supercell and ordinary deep convection by Rasmussen and Blanchard (1998). While EHI can be calculated over different depths, Rasmussen (2003) showed that 0–3-km EHI was better than 0–1-km EHI for discriminating between supercells and nonsupercells. In this study, the 0–3-km layer is used for SRH and CAPE is based off a surface-based parcel.

SCP incorporates most unstable parcel convective available potential energy (MUCAPE) along with effective-layer SRH (ESRH) as well as the effective bulk wind difference (EBWD; Thompson et al. 2007). Effective layers are designed to represent the inflow layer for supercells and the calculations follow Thompson et al. (2007). EBWD is defined as the magnitude of the vector difference in the wind between the bottom of the effective inflow layer and the point halfway to the equilibrium level of the most unstable parcel (Thompson et al. 2007). The formulation of SCP used here follows the Storm Prediction Center mesoanalysis calculation which also incorporates the

Corresponding author: Stephen A. Shield, stephen.shield@huskers.unl.edu

DOI: 10.1175/WAF-D-21-0098.1

most unstable parcel convective inhibition (MUCIN) based on the work of Gropp and Davenport (2018):

$$\text{SCP} = \frac{\text{MUCAPE}}{1000} \times \text{EBWD} \times \frac{\text{ESRH}}{50} \times \text{MUCIN}, \quad (2)$$

where MUCAPE has units of joules per kilogram (J kg^{-1}), ESRH has units of meters squared per second squared ($\text{m}^2\ \text{s}^{-2}$), and

$$\text{EBWD} = \begin{cases} 0; \text{EBWD} < 10 \\ \text{EBWD}/20; 10 < \text{EBWD} < 20 \\ 1; \text{EBWD} > 20 \end{cases},$$

where EBWD has units of meters per second (m s^{-1}), and

$$\text{MUCIN} = \begin{cases} -40/\text{MUCIN}; \text{MUCIN} < -40 \\ 1; -40 > \text{MUCIN} \end{cases},$$

where MUCIN has units of joules per kilogram (J kg^{-1}). All normalization factors have units that render SCP unitless.

While composite indices like EHI or SCP are useful tools for diagnosing supercell storm potential, there are some potential shortcomings.

First, they incorporate a limited number of variables. Other variables have been shown to indicate if an environment supports supercells. Supercells are more often observed when CAPE in the lowest 3 km above the level of free convection (LFC) is larger (Rasmussen and Blanchard 1998). Upper-tropospheric storm-relative wind has been found to be slightly higher in supercell environments (Rasmussen and Blanchard 1998), and effective inflow-layer storm-relative wind has also been shown to be higher in supercell environments (Peters et al. 2020). Mixed-layer lifting condensation level (MLLCL) height is lower and 0–1-km shear magnitude is higher in environments that support supercells than environments with nonsupercell storms (Thompson et al. 2002a). Thompson et al. (2002b) as well as Gropp and Davenport (2018) contend that a supercell is less likely as CIN magnitudes increase while Rasmussen and Blanchard (1998) found that supercells were more likely in environments with larger CIN magnitudes. These hypotheses each have theoretical support since CIN magnitudes that are too large would tend to prevent the sustenance of deep convection required for storms to evolve into supercells. In contrast, low CIN magnitudes could favor widespread deep convection that interferes with discrete supercell organization.

Second, while these composite indices are based on simple equations that are easily calculated, understood, and trusted by forecasters they may fail to capture crucial relationships. For example, an environment may have a large EHI value because of extreme CAPE and small but nonzero SRH, but the minimal SRH makes rotating storms less likely than in an environment that may have the same EHI value but more SRH and less CAPE. Similarly, SCP is not based on a proven physical process but on the observations of supercells which leads to short comings in certain situations. For example, since each term in SCP is multiplied together a low value of SCP would result from an environment with high shear and low cape, although supercells occur in such environments (Sherburn and Parker 2014).

Machine learning has the ability to generate predictions incorporating a variety of inputs based on relationships learned from historical data. As a result of this capability machine learning is becoming more widely used in the atmospheric sciences including for convective weather applications. These applications include using machine learning for storm classification (e.g., Haberlie and Ashley 2018; Jergensen et al. 2020), nowcasting the risk of hazards from convective storms (e.g., Cintineo et al. 2014; Lagerquist et al. 2017; Cintineo et al. 2018; Gagne et al. 2019; Cintineo et al. 2020; Lagerquist et al. 2020), and postprocessing output from convection-allowing numerical weather prediction models (e.g., Loken et al. 2020; Sobash et al. 2020; Flora et al. 2021). Machine learning has also been used to generate convective hazard outlooks and tornado warnings similar to those issued by the National Weather Service (e.g., Hill et al. 2020; Steinkruger et al. 2020). Research using machine learning to directly diagnose environments supportive of supercells has been more limited. Nowotarski and Jensen (2013) utilized single variable self-organizing maps to identify differences between environments for nonsupercells, nontornadic supercells, weakly tornadic supercells, and significantly tornadic supercells. While their results were promising, they were limited by the use of a single sounding derived variable (e.g., the vertical wind profile). Nowotarski and Jones (2018) increased the number of variables used but focused on differences in tornadic production between supercells rather than between supercells and nonsupercells.

A machine learning approach is advantageous for complex tasks like diagnosing storm environments because machine learning models have the ability to represent complex and nonlinear relationships between variables as well as the ability to recognize multiple types of environments that support supercell storms (Nowotarski and Jensen 2013). In this study we hypothesize that machine learning can be used to create a forecast tool to discriminate between environments that are supportive of supercells and those that are not. These machine learning models are evaluated against the relatively simple forecast indices currently used by operational forecasters.

2. Data and methods

a. *Environmental data and input variables*

Thompson et al. (2003) developed a climatology of soundings from convective storm environments using soundings from the Rapid Update Cycle version 2 (RUC-2) model (Benjamin et al. 2004). Thompson et al. (2007) expanded upon this climatology to include additional storm environments. This database of soundings is used to train and test machine learning models. In total, 1079 modeled environmental soundings from a combination of near-storm environments of supercells and discrete, nonsupercellular storms were used. Although multiple storms typically occur on a given day, using many soundings from the same day limits the diversity of environments and could lead to leakage between the training and test dataset if environments are too similar. The database used in this study has an average of two cases for a given calendar day with data collected over the conterminous

United States over six different calendar years. Additionally, any cases from the same day must be separated by at least 3 h and 185 km. As a result, this database includes cases from a wide variety of environments with limited correlation between each case. The manual classification of events as supercell or nonsupercell performed by Thompson et al. (2003, 2007) is used here. They based their classification of supercells on the presence of a visible hook echo or inflow notch and azimuthal shear ≥ 20 m s^{-1} across a distance of less than 10 km at a radar elevation angle of 0.5° or 1.5° and 3°, that persists for at least 30 min. Nonsupercells were defined as discrete storms possessing a max composite reflectivity of 40 dBZ for more than 30 min and failing to meet any of supercell characteristics. Soundings were collected at the RUC-2 analysis time closest to the storm's peak intensity in the area upwind form the storm. In total the database used contains ~77% right moving supercells and ~23% nonsupercells from across the conterminous United States.

This sampling criteria has an important implication for interpreting the output of the machine learning models developed in this study. These machine learning models are not predicting the probability of a supercell, but rather the conditional probability of a supercell given deep convection that persists for at least 30 min. Because the machine learning models are predicting a conditional probability, the findings of this research only apply to scenarios where deep convection initiation and/or sustenance is not in question but the storm mode is uncertain. Other forecast tools should be used to assess issues such as the likelihood of deep convection initiation and/or sustenance. Additionally, because the models are trained only on right moving supercells the models should not be used to assess the potential for left moving supercells.

The proposed machine learning models rely on nine common sounding derived variables (Table 1) and are calculated using SHARPpy (Blumberg et al. 2017). Following the logic of Thompson et al. (2002a), thermodynamic variables are based on the most unstable parcel in the lowest 400 hPa. This approach enables consideration of both surface-based and elevated deep convection. Low-level CAPE is defined using the 3-km layer immediately above the LFC (Rasmussen and Blanchard 1998). Storm-relative variables are calculated using the internal dynamics method (Bunkers et al. 2000) to estimate storm motion of a right moving supercell. Equilibrium-level storm-relative wind (ELSRW) is defined as the storm-relative wind at the equilibrium level of the most unstable parcel. The effective inflow layer used for ESRH and EBWD is the same as used in Eq. (2). Figure 1 shows bivariate relationships between the variables used in the study. Correlation analysis shows that there is limited correlation between variables ($|r| < 0.6$), with the exception of the variable pairs MUCAPE/LLCAPE, ESRH/0–1 BWD, and EBWD/ESRW. While some correlation between variables is expected, the limited correlation suggests that the nine input variables incorporate multiple environmental factors that regulate supercell likelihood without excessive redundancy. This limited variable set reduces model complexity which can improve generalizability and improve human understanding of the machine learning model (Belle and Papantonis 2021).

TABLE 1. Sounding derived variables used models, grouped by variable type.

Thermodynamic	Kinematic
Most unstable parcel CAPE (MUCAPE)	0–1-km bulk wind difference (0–1 BWD)
Most unstable parcel CIN (MUCIN)	Effective bulk wind difference (EBWD)
Most unstable parcel LCL (MULCL)	Effective storm-relative helicity (ESRH)
Most unstable low-level cape (LLCAPE)	Equilibrium-level storm-relative wind (ELSRW)
	Effective inflow-layer storm-relative wind (ESRW)

b. *Machine learning model development*

Machine learning models were developed with the goal of improving assessment of the storm mode supported by an environment, and the role of each variable in determining storm mode. Three types of machine learning models were created: a support vector machine (SVM), an artificial neural network (ANN), and an ensemble of gradient boosted trees (GBT). To facilitate model development, data were first split into training and testing datasets using a stratified split method where the data are shuffled and then split into different sets: 20% of the data are used for testing and 80% are used for training (step 1 of Fig. 2). This process also ensures that the balance between supercell (~77%) and nonsupercell cases (~23%) is maintained in each set. The test data are set aside and only used for evaluating model performance (section 3).

Each machine learning model has some number of hyperparameters that govern the exact model configuration and overall complexity. To determine the approximate level of model complexity necessary for this task, different model configurations were evaluated by a hyperparameter search using 10-fold stratified cross validation on the training dataset where the model with a specific set of hyperparameters is trained on nine folds of the training data and error metrics are calculated on the other fold (Kuhn and Johnson 2013). This is repeated 10 times with each fold used as the validation set once. This results in 10 values of the error metric [Matthews correlation coefficient (MCC), which is described in detail in section 2c], one for each validation fold. The mean and standard deviation of MCC values across the 10 folds is calculated and configurations were ranked relative to other model configurations for both mean MCC and standard deviation of MCC with high mean and low standard deviation values indicative of a good model configuration. The final model configuration was chosen based on the rankings of each of the configurations. The final model configuration was then trained using the entire training dataset and it was used to make predictions on the test dataset (step 3 of Fig. 2). This methodology allows for the full training dataset to be used to train the final models while maintaining an independent test dataset. While the model performance on the test dataset may be slightly different than the estimate from the cross-validation step, the estimate of model performance from the k-fold cross validation was within the confidence interval of the model performance on the

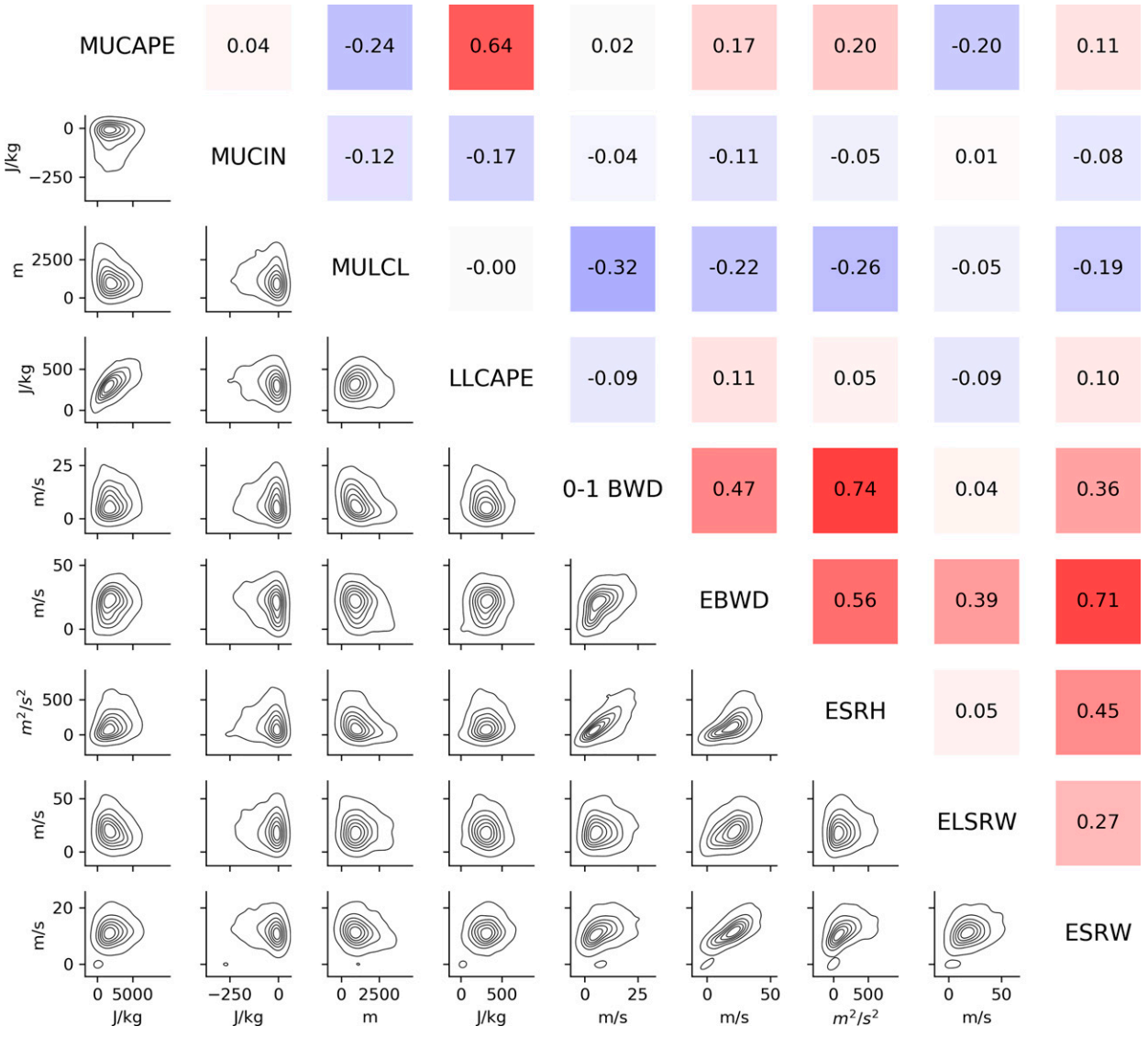

FIG. 1. Predictor variable matrix. The variables are listed on the diagonal while the Pearson correlation coefficient between each pair of variables is shown in the upper-right portion the grid. The bottom left shows the parameter space occupied by the dataset via kernel density estimation.

independent test dataset shown in section 3 for all three models. This suggests the methodology employed both prevents overfitting and is an effective way to test different model configurations when the dataset size makes independent training, validation, and testing sets impractical. For the ANN and SVM, the input data were scaled to have zero mean and a standard deviation of one. The data scaling was performed using the mean and standardized deviation of variables in the training dataset only to prevent information leakage (Kaufman et al. 2012).

An SVM works by finding an optimal hyperplane to separate the two classes (supercell and nonsupercell) in the nine-dimensional parameter space. [See chapter 5 of VanderPlas, (2016) for an entry-level explanation of support vector machines and how they are constructed.] An SVM was chosen because they are effective in modeling complex relationships when relatively small training datasets are available (Géron 2017). Three hyperparameters were varied in the model configuration stage: 1) a penalty parameter, 2) the kernel, and 3) class weights. The penalty parameter C controls the penalty applied for each misclassified case and was varied from 0.25 to 1.25. The kernel controls the type of decision boundary and can either be linear or include transforms of the variables to allow for nonlinear decision boundaries. Each configuration used one of the following kernels: linear, polynomial, radial basis function, and sigmoid. Class weights are used so the penalization of supercell misclassification is different than for a nonsupercell case. Since there are significantly more supercells than nonsupercells in the dataset, equal penalties could result in the model favoring predicting a supercell. The weight

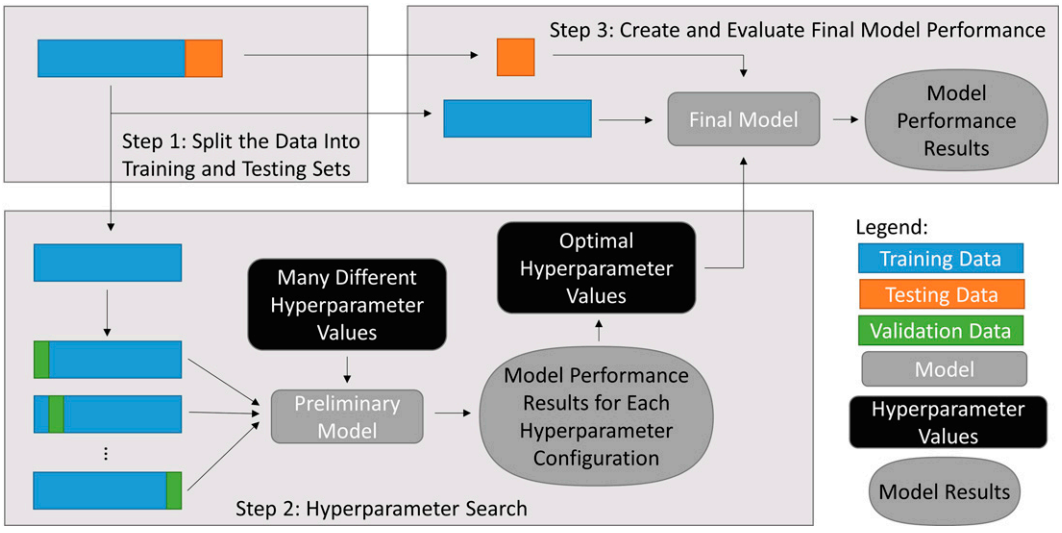

FIG. 2. Workflow for developing machine learning models. Step 1: The data are split into training and testing sets using 80% for training and 20% for testing. Step 2: A hyperparameter search is conducted using *k*-fold cross validation where a portion of the training data is used to validate the performance of each model configuration. Based on the performance of each model configuration a final set of hyperparameters are chosen. Step 3: The optimal hyperparameter values are used to construct the final model that is trained on the entire training dataset and evaluated on test dataset to determine the model's performance.

of nonsupercell cases relative to supercell cases was varied from 2.125 to 5.25. The SVM was implemented with scikit-learn (Pedregosa et al. 2011) and the final model configuration utilized a linear separation boundary, a *C* value of 0. 25, and relative weight of nonsupercells to supercells of 2.57.

An ANN uses a network of hidden layers of fully connected neurons to make predictions. An ANN was chosen because ANNs are theoretically capable of representing any mathematical function. Relatively shallow networks like the one developed in this study are capable of performing well even with limited training data (Aggarwal 2018). Each connection uses a different weight with each neuron representing a unique weighted sum based on the weights and input variable values. To generate predicted probability, variable values are inputted into this network and the prediction is based on the weighted sum of the neurons in the output layer. The result is a model that is able to handle very complex and nonlinear relationships. The weights of each connection are adjusted during the training to maximize the performance of the algorithm. In the model configuration stage thousands of randomly selected model architecture and hyperparameter combinations are evaluated. Each model configuration tested varies 1) the number and/or size of hidden layers, 2) the regularization method/amount, 3) the activation function, and 4) the solver or optimizer. Chapter 13 of Boehmke and Greenwell (2020) provides a brief introduction to artificial neural networks and describes the role of each of the hyperparameters in the model. In the hyperparameter search stage of model development, the number of hidden layers ranged from two to four with between 4 and 64 neurons in each layer. The regularization method/amount is used to decrease the likelihood of the model overfitting. Two methods of regularization were used: L2 regularization which introduces a penalty term based on the model weights to the loss function, and dropout which randomly ignores or "drops" some neurons during the training process. The amount of penalty (L2) or the probability of an individual neuron being "dropped" (dropout) was also varied during the hyperparameter search. The activation function controls how the weighted sum of inputs to a neuron are converted to outputs. Each configuration used one of the following activation functions which converts the weighted sum of inputs to an output value slightly differently: rectified linear unit (ReLU), exponential linear unit (ELU), sigmoid, tanh, scaled exponential linear unit (SELU). The solver, or optimizer, controls the weight adjustments within the network during the model training. Each of the following solvers were evaluated during the hyperparameter search since each solver uses a different method of updating weights within the model: adaptive moment estimation (Adam), root-mean-square prop (RMSProp), Nesterov-accelerated adaptive moment estimation (Nadam), and adaptive gradient (Adagrad). The ANN was implemented with Keras (Chollet et al. 2015) and the final model configuration had 3 hidden layers with 40, 36, and 28 neurons in each layer, respectively, L2 regularization (Aggarwal 2018) was used with the amount ("alpha") parameter set to 0.001, with a SELU (Klambauer et al. 2017) activation function, the solver "Adagrad" (Duchi et al. 2011) with default parameters, and a binary cross-entropy loss function were used to train the model.

A GBT uses a large number of decision trees to make a probabilistic prediction. A gradient boosted tree ensemble is a sequential ensemble where each tree added to the ensemble after the initial tree is trained to correct the error of the ensemble up to that point. A gradient boosting ensemble was chosen because they have proven effective in a variety of domains of

Boehmke and Greenwell (2020) and can result in impressive performance by combing a series of weak learning decision trees (Hastie et al. 2009). [See chapter nine of Boehmke and Greenwell (2020) for a good introduction on decision trees, and chapter 12 for an explanation of how boosting can be used to combine multiple weak learners (in this study the individual decision trees) to create a sequential ensemble.] Values for four hyperparameters were tested in the model configuration stage: 1) the number of trees used, 2) the number of variables evaluated for each split in each tree, 3) the size, or maximum depth of each decision tree, and 4) the learning rate. The number of trees represents the size of the ensemble, and the number of trees used ranged from 50 to 800. The optimal number of variables used to construct each tree controls the bias/variance of the model. The number of variables used to construct each tree ranged from 1 to 4. The optimal size or depth of each decision tree depends on the degree of interaction between variables and the tolerance for overfitting with larger depths able to capture variable interactions but also increase the risk of overfitting (Boehmke and Greenwell 2020). The depths tested ranged from 1 to 11. The learning rate controls the impact of each new decision tree added to the ensemble. Learning rates from 0.05 to 0.15 were evaluated. The GBT was implemented with scikit-learn (Pedregosa et al. 2011) and trained using the default loss function "deviance" and the default criterion "Friedman mean squared error" to determine the quality of splits within each tree. The final model configuration utilized 200 trees and 2 variables per tree, a maximum depth of 1 and a learning rate of 0.125.

c. *Model evaluation*

Final model configurations are trained on the entire training dataset and used to make predictions on test data. Due to the small dataset and need to generate confidence intervals, 1000 bootstrap samples (Efron and Tibshirani 1994) of test data are generated and performance metrics calculated for each sample. The 95% confidence interval is estimated by the 2.5th–97.5th percentiles in the performance metric for the 1000 bootstrap samples. MCC is used as a measure of overall model performance due to its ability to evaluate binary classifiers on imbalanced data (Boughorbel et al. 2017):

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}, \quad (3)$$

where TP represents the number of correct predictions of supercell storm mode, TN represents the number of correct predictions of nonsupercell storm mode, FP represents the number of incorrect predictions of supercell storm mode, and FN represents the number of incorrect predictions of nonsupercell storm mode. An MCC value of one indicates a perfect prediction while an MCC value of zero indicates a prediction with no skill.

While the MCC is the preferred metric due to the imbalance between supercells and nonsupercells, other metrics can provide additional insight and a more intuitive understanding of performance. For this reason model performance is plotted on a performance diagram (Roebber 2009) where the x axis is success ratio, or one minus the false alarm ratio (FAR), and the y axis is probability of detection (POD). To facilitate this visualization POD and FAR were calculated. They are defined by Eqs. (4) and (5), respectively, with TP representing the true positive rate, FP representing the false positive rate, and FN representing the false negative rate:

$$\text{probability of detection} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad (4)$$

$$\text{false alarm ratio} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}. \quad (5)$$

The three models were compared to threshold predictors based on EHI and SCP. For each sounding EHI and SCP values were calculated and a prediction of storm mode was made based on if the value of the index for that instance was above or below a specific threshold. The thresholds were optimized using the same cross-validation technique used to determine the optimal machine learning model hyperparameters. For EHI the threshold was varied from 0.05 to 5 in increments of 0.05 and the optimal threshold was determined to be 0.3. For SCP the threshold was varied from 0.1 to 10 in increments of 0.1, and the optimal threshold was determined to be 0.5.

A potential limiting factor to the robustness of these performance evaluations is the fact that supercells made up of a majority (~77%) of cases in the dataset used, while in reality supercells occur much less frequently than nonsupercells. As a result the raw MCC, POD, and FAR values may not accurately represent the performance of models/indices in a more climatologically realistic dataset. While the actual occurrence of supercells is unknown, the work of Duda and Gallus (2010) suggests a reasonable approximation of supercell occurrence for severe convective weather events is around 25%.

To assess performance on a more climatologically realistic ratio of supercells, weighted versions of each metric were created. In the weighted versions both the true positive (TP) and false negatives (FN) are multiplied by 0.1 to reduce their impact. This approximates the performance that would be expected when ~25% of the samples are supercells instead of ~77% of samples as is the case in the original dataset.

For analysis on a more climatologically representative ratio, model predicted probabilities were adjusted using the following formula, which was developed by Saerens et al. (2002):

$$\text{adjusted predicted probability} = \frac{\mathrm{IPP} \times \dfrac{\mathrm{TR}}{\mathrm{CR}}}{\mathrm{IPP} \times \dfrac{\mathrm{CR}}{\mathrm{TR}} + (1 - \mathrm{IPP}) \times \dfrac{(1 - \mathrm{CR})}{(1 - \mathrm{TR})}}, \quad (6)$$

where IPP represents the initial predicted probability, the ratio of supercells to nonsupercells in the training data is represented by TR and the climatological ratio of supercells to nonsupercells is represented by CR. For the ANN and GBT, TR was set at 0.77 while for the SVM TR was set at 0.5 since weights were used in model training to balance supercells and nonsupercells. CR was set at 0.25 for all models. The thresholds used for both EHI and SCP were also adjusted with the cross-validation search technique repeated using the weighted MCC. Optimal thresholds for the more climatologically realistic case were found to be 0.6 for EHI and 3.0 for SCP.

d. Model interpretation methods

Several methods are used to interpret the machine learning models. Multipass permutation variable importance (Lakshmanan et al. 2015; McGovern et al. 2019) is used to rank variable impact in each model. Multipass permutation variable importance works by using the original performance of the model on the test dataset as a baseline. A single input variable from the test set is then permuted and the model performance is evaluated on the test set with the permuted variable. This is repeated for each input variable, and the variable which results in the largest drop in performance when permuted is ranked as the most important variable. This process is then repeated with the exception that any variable which has already been ranked remains permuted, this continues until all variables are ranked. In this study each permutation was done randomly 100 times with the average performance decrease used. This was done to reduce the impact of unimpactful permutations (e.g., a random permutation where by chance a large majority of samples end up with values similar to their initial values). While multipass permutation importance is more computationally intensive than ordinary permutation variable importance measures, theoretically this method results in more accurate results when correlation is present between variables (McGovern et al. 2019).

Examining the accumulated local effect (ALE) (Apley and Zhu 2020) of each variable is another method to examine the individual impact of each variable even when correlation between variables exists (Molnar 2018). Local effects are calculated by isolating a small window of observations (in this study the deciles of the data) and measuring the average change in predictions across that window for a given variable. This is done by taking all observations in a window and setting the value of the variable of interest to the lower bound of the window making predictions, and then repeating for the upper bound of the window. The average difference in predictions is the local effect of that variable. In this study, a moving decile window was used to create ALE curves with the center ranging from the 5th percentile to the 95th percentile of the variable of interest.

3. Results and discussion

a. Model performance

Model performance metrics were calculated from model predictions on the held out test dataset (step 3 of Fig. 2).

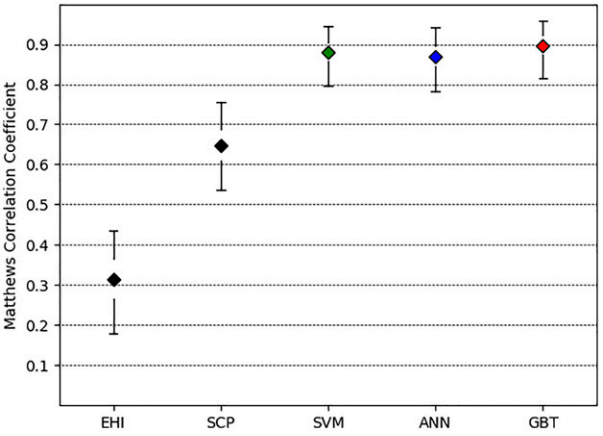

FIG. 3. Model performance for the supercell composite parameter (SCP), the support vector machine (SVM), the artificial neural network (ANN), and the gradient boosted tree (GBT) ensemble. Whiskers represent the 95% confidence interval of the model performance.

Dichotomous predictions of storm type for each machine learning model were generated using a 50% probability threshold, while the baseline EHI and SCP predictions were based on the thresholds determined in section 2c. The MCC for each of the three machine learning models as well as the benchmark EHI and SCP threshold predictors, are shown in Fig. 3. Means and confidence intervals rely on the bootstrapped samples. Results show that the GBT has the highest mean MCC of 0.90 followed by the SVM (0.88) and the ANN (0.87). The 95% confidence intervals for all three models overlap and does not indicate a statistically significant difference between the three machine learning models. By comparison the EHI has a mean MCC of 0.31 and the SCP has a mean MCC of 0.65. The 95% confidence intervals indicate that the SVM, ANN, and GBT all have statistically significant better performance than the EHI and SCP threshold predictors.

Further insight into model performance is gained by examining the performance diagram (Fig. 4). The performance diagram indicates that the machine learning models have higher mean POD than EHI or SCP. The difference between EHI and the three machine learning models is statistically significant, while the confidence intervals indicate that the difference between SCP and the machine learning models is only statistically significant for the ANN and GBT. The machine learning models also have lower mean FAR than both SCP and EHI. The difference in FAR between all three machine learning models and EHI is statistically significant, while the confidence intervals do not indicate a statistically significant difference between SCP and any of the machine learning models. From Fig. 4 we can also deduce that the bias of the machine learning models is better than either SCP or EHI with bias represented by the dashed contours and 1.0 being the ideal value. Similarly, the machine learning models have a higher critical success index (CSI) with CSI represented by the continuous contours and 1.0 being an ideal CSI.

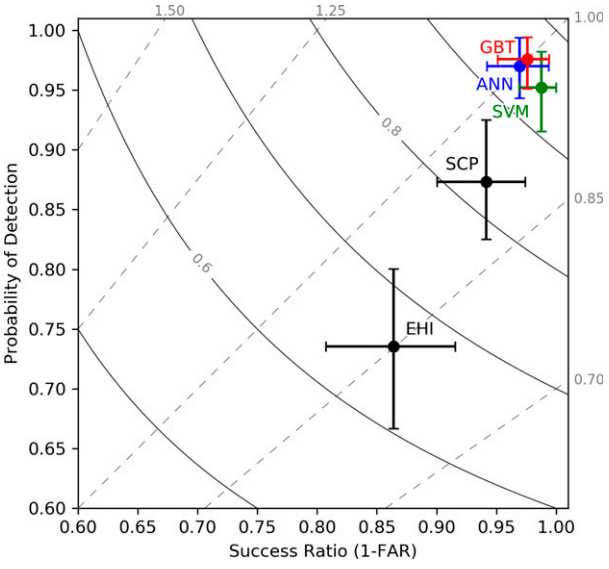

FIG. 4. Model performance diagram: The success ratio is on the *x* axis, and probability of detection is on the *y* axis. The dashed contours labeled on the top and right indicate bias values while the solid contours indicate critical success index values. Performance is plotted for each model as well as SCP and EHI with whiskers indicating the 95% confidence intervals.

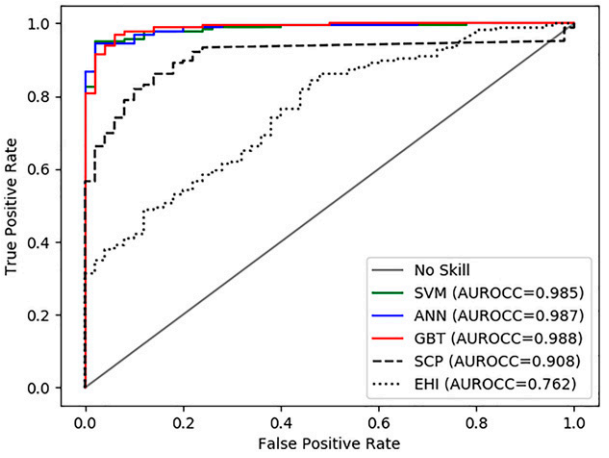

FIG. 5. Receiver operating characteristic (ROC) curves for each of the three machine learning models as well as SCP and EHI. The area under the ROC curve (AUROCC) for each is shown in the legend.

Another useful method for evaluating model performance is the receiving operator characteristic (ROC) curve. The ROC curve shows how the true positive rate and false positive rate change as the decision threshold is changed with each point on the curve representing the true positive and false positive rates for a specific decision threshold. The area under the ROC curve (AUROCC) is a measure of overall skill and represents the average likelihood that two different instances are ranked in the correct order. An AUROCC value of 1.0 indicates a perfect prediction while an AUROCC value of 0.5 indicates a prediction with no skill (i.e., random chance of classification). Both the ROC curves and the AUROCC for each machine learning model and the two benchmark indices is shown in Fig. 5.

Generally, the three machine learning models have higher true positive rates and lower false positive rates than either EHI or SCP across a range of decision thresholds. There are minimal differences between the three machine learning models. This is quantified by the AUROCC values where the three machine learning models have AUROCC values between (0.985 and 0.988) while AUROCC values for EHI and SCP are lower at 0.762 and 0.908. These differences were found to be statistically significant (not shown).

The strong performance of the machine learning models in ROC curve analysis suggests that while their raw probabilities have the potential to be biased toward predicting supercell storm mode as a result of the dataset used, they are still skillful and can be used in situations where supercells are much more rare so long as the decision thresholds or probabilities are adjusted. This is illustrated by testing performance of models using adjusted probabilities (machine learning models) and thresholds (EHI and SCP) on weighted versions of MCC, POD, and FAR.

The weighted MCC for each of the three machine learning models as well as the benchmark EHI and SCP threshold predictors are shown in Fig. 6. The weighted MCC is used to estimate performance for a more climatological realistic ratio of supercells to nonsupercells. Results show that the SVM has the highest mean weighted MCC of 0.92 followed by the ANN (0.91) and the GBT (0.87). The 95% confidence intervals for all three models overlap and does not indicate a statistically significant difference between the three machine learning models. By comparison the EHI has a weighted MCC mean value of 0.32 and the SCP has a weighted MCC mean value of 0.68. The 95% confidence intervals indicate that the SVM, ANN, and GBT all have statistically significant better performance than the EHI and SCP threshold predictors.

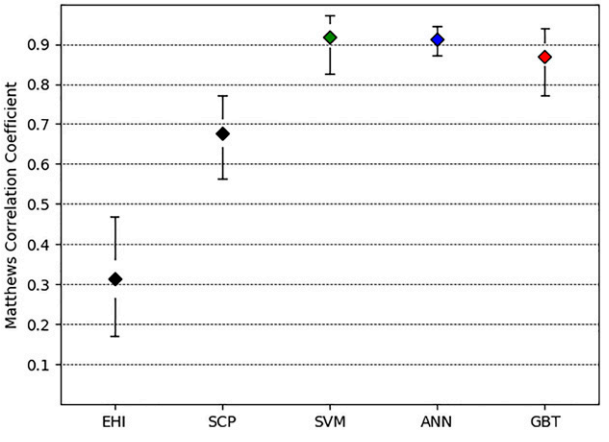

FIG. 6. As in Fig. 3, but for the weighted MCC to replicate more climatologically representative supercell occurrence.

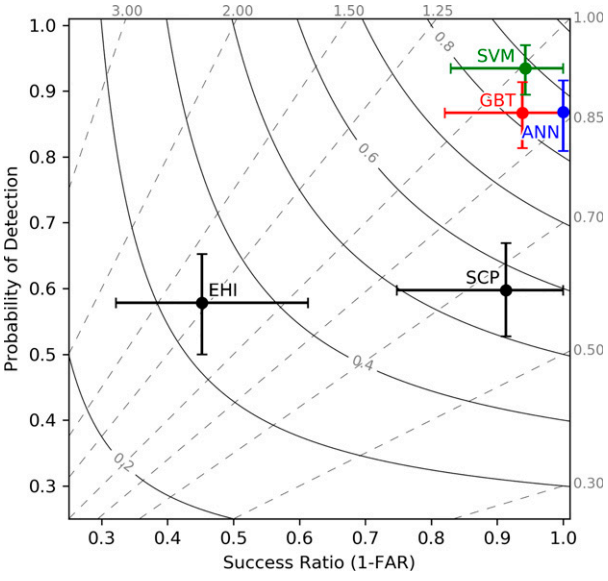

FIG. 7. As in Fig. 4, but for the weighted POD and FAR to replicate more climatologically representative supercell occurrence.

Figure 7 shows the performance diagram with both POD and FAR weighted to represent a more climatologically realistic environment. The performance diagram indicates that the machine learning models have higher mean POD than EHI or SCP. The difference in POD between the three machine learning models and both EHI and SCP is statistically significant. The machine learning models also have lower mean FAR than both SCP and EHI. The difference in FAR between all three machine learning models and EHI is statistically significant while there was not statistically significant difference in FAR between SCP and any of the machine learning models. Similar to the unweighted case, the bias of the machine learning models is better than either SCP or EHI and machine learning models have a higher CSI.

To evaluate the sensitivity of the results to the methodology of estimating performance on a more climatologically reasonable environment, resampling on the test dataset was conducted to construct a separate test dataset made up of 25% supercells and 75% nonsupercells. Results of MCC, POD, and FAR for this dataset were consistent with the weighted versions of these metrics shown in Figs. 6 and 7.

While the machine learning models show a clear improvement in discriminating between storm modes, it is necessary to determine if their predictions are based on physically reasonable relationships between predictor variables and storm mode. This is done by assessing variable importance rankings and individual variable effects in the following sections.

b. *Variable importance rankings*

Variable importance rankings for each machine learning model are shown in Fig. 8. There is agreement between all three models on the top two variables, EBWD and ESRH. EBWD is by far the most influential variable with MCC decreases of 0.58 for the GBT, 0.62 for the ANN and 0.69 for the SVM. This is consistent with previous research which notes the importance of deep-layer shear for supercell organization (Thompson et al. 2003, 2007; Houston et al. 2008). The second highest ranked variable in all models, ESRH, has MCC decreases of 0.14 for the SVM, 0.15 for the GBT and 0.18 for ANN.

While the models show less agreement in the rankings of the other variables, they generally agree on the importance of kinematic variables with the only thermodynamic variable ranked in the top five being MUCIN, ranked fourth in all models. Both MUCAPE and LLCAPE are ranked in the bottom three variables for all three models. A likely explanation for this result is the data sampling strategy used. Because nonsupercells in the dataset are required to maintain radar reflectivity of at least 40 dBZ for a minimum of 30 min, favorable thermodynamic conditions for deep convection are present in both supercells and nonsupercell environments resulting in a large overlap between MUCAPE values for supercells and nonsupercells. Additionally, kinematic effects can serve to increase updraft strength (Peters et al. 2019), meaning that for similar CAPE values, stronger updrafts would be expected in an environment with favorable kinematics.

One variable which has disagreement between the models is ESRW. It is ranked third in the GBT, eighth in the ANN, and ninth in the SVM. While previous research (Peters et al. 2020)

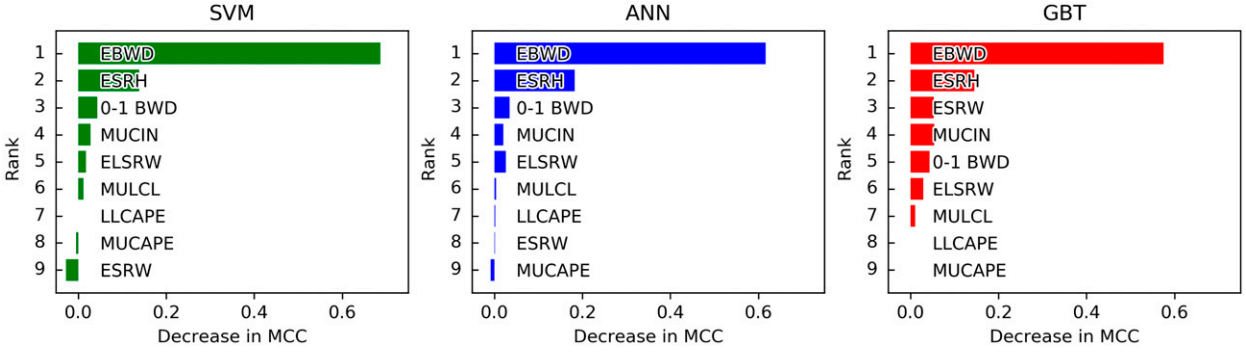

FIG. 8. Variable importance results obtained via multipass permutation variable importance calculation. The rankings indicate the order of the variable ranking, while the decrease in MCC shows the amount the model performance dropped with larger decreases indicating a variable has more influence on model performance.

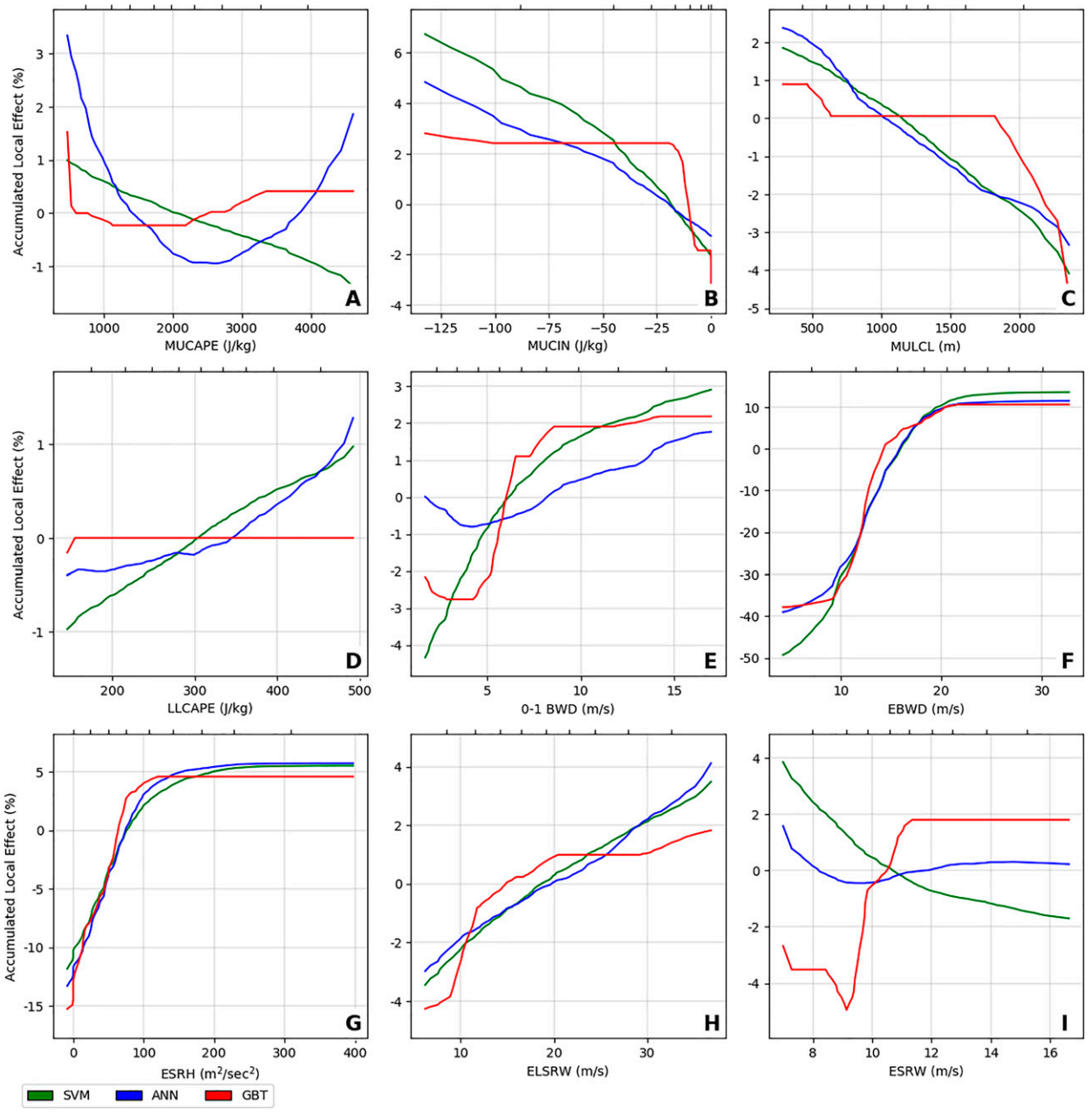

FIG. 9. Accumulated local effect (ALE) for each of the nine predictor variables. The green lines show the ALE for the SVM, the blue lines the ALE for the ANN, and the red lines shows the ALE for the GBT. The *x* axis shows the change in the value of the variable of interest, with tick marks along the top of each panel indicating the deciles of the data. The *y* axis of each plot shows the mean centered change in probability of a supercell according to each model.

has suggested that the ESRW component of ESRH is the primary driver of storm mode, our results show that ESRH is more important in all models, and that in the ANN and SVM ESRW is unimportant. We discuss this finding more in the following section; however, it is likely either related to the correlation between ESRW and EBWD or a limitation of the models.

While the rankings show that each model weights variables slightly differently, overall, the models appear to have logical variable importance rankings, suggesting that they are based on learned relationships that are physically reasonable. Deeper analysis of each variable's impact is contained in the following section.

c. *Individual variable effects*

ALE values for the nine predictor variables appear in Fig. 9. Values can be interpreted as the change in the probability of a supercell for changes in the value of a particular

variable. For example, in Fig. 9a, the ALE for the ANN model increases by about 1% as MUCAPE increases from ~2500 to ~4000 J kg^{-1}, which means that, all else equal, the probability of a supercell would increase by 1% if CAPE were to increase from 2500 to 4000 J kg^{-1}.

Figure 9a shows the ALE for MUCAPE for all three models. The SVM shows a linear decrease in supercell probability; however, the effect is small, around 2%. The GBT has similar small changes in probability, first decreasing then increasing, but the maximum change is small, less than 2%; the small change is consistent with the low importance ranking of MUCAPE in the GBT. The ANN shows a similar decrease then increase in supercell probability as the GBT but with larger magnitudes. Supercell probability decreases by about 4% as MUCAPE increases to ~2500 J kg^{-1} and then increases by around 3% as MUCAPE increases from ~2500 J kg^{-1} to near 5000 J kg^{-1}. While there is little support for the decreases in supercell probability as MUCAPE increases, the increase in probability for larger MUCAPE values is consistent with previous modeling studies (Kirkpatrick et al. 2011) as well as observational climatologies (Rasmussen and Blanchard 1998; Thompson et al. 2002a).

The SVM and ANN show a steady decrease in supercell probability of 6% to 8% as the magnitude of MUCIN deceases from ~−125 to 0 J kg^{-1} (Fig. 9b). The GBT shows little change in supercell probability as MUCIN magnitude decreases from ~−125 to ~−25 J kg^{-1}; however, a decrease in supercell probability of around 6% occurs as MUCIN decreases from −25 to 0 J kg^{-1}. These results agree with Rasmussen and Blanchard (1998), who argue that supercells are more likely with more CIN. Initially, this result appears to contradict the hypothesis of Thompson et al. (2002b), who suggested that supercell formation and sustenance is less likely as CIN increases. However, the data sampling strategy used to train the model came from a collection of storms that were strong, and sustained themselves for at least 30 min. This suggests that while large CIN values are present in the dataset, in these cases the CIN was overcome by other factors such as forcing mechanisms or pressure perturbations resulting from the storm itself. For this reason, it would not be expected that the model would reflect the negative effect of CIN on storm maintenance. If this were not the case and soundings from storms that failed to sustain themselves were included, the ALE plot would be more likely to resemble an inverse U shape which would be expected based on the combined hypotheses of Rasmussen and Blanchard (1998) and Thompson et al. (2002b).

Both the SVM and ANN show linear decreases in supercell probability of around 6% as MULCL height increases from around 500 m to in excess of 2000 m (Fig. 9c). In contrast, the GBT shows little change in supercell probability as LCL increases from near the surface up to 1800 m after which a decrease in supercell probability of around 5% occurs as MULCL height increase above 2000 m. This result is generally consistent with prior work (e.g., Thompson et al. 2002a) and has theoretical backing because as LCL height increases, storms produce colder outflow which can horizontally decouple near-surface vertical vorticity and the low-level lifting (Brooks et al. 1994; Snook and Xue 2008; Markowski and Richardson 2009; Houston 2016; Brown and Nowotarski 2019).

Both the SVM and ANN show around a 2% increase in supercell probability as low-level CAPE increases from near 0 J kg^{-1} to around 500 J kg^{-1} (Fig. 9d). The GBT shows that LLCAPE has no effect which is consistent with the result of the permutation variable importance results (section 3c). The SVM and ANN, on the other hand, show a small increase in supercell probability as LLCAPE increases. Although the relative increase in probability observed is small, these results are consistent with the results of Rasmussen and Blanchard (1998), McCaul and Weisman (2001), and Kirkpatrick et al. (2011), who find that supercells are favored as low-level CAPE increases. Low-level CAPE can be particularly important in cases where overall CAPE was low or LCL heights were high (Davies 2006).

All three models show that as 0–1 BWD increases so does the supercell probability. However, both the ANN and GBT show a small decrease in supercell probability of around 1% as 0–1 BWD increases from 0 to around 3 m s^{-1}. As 0–1 BWD increases beyond that the GBT shows a large increase in supercell probability of 5% as 0–1 BWD increases to around 8 m s^{-1} after which there is little increase in supercell likelihood with increasing 0–1 BWD. The ANN shows a more gradual increase in supercell likelihood of 2% or 3% as 0–1 BWD increases from 3 to over 15 m s^{-1}. The SVM shows constant increase in supercell likelihood with an increase of about 5% as 0–1 BWD increases from 0 to 8 m s^{-1}. Above that the SVM continues to show an increase in supercell likelihood, but the effect is smaller, only an increase of around 2% in supercell probability as 0–1 BWD increases from 8 to over 15 m s^{-1}. The overall trend of increasing supercell likelihood with increasing 0–1 BWD is consistent with the findings of Thompson et al. (2002a), who noted that 0–1-km shear was higher in supercell environments.

Figure 9f shows that all three models have large increases in supercell likelihood as EBWD increases. An increase in probability of ~50% is observed as shear magnitude increases from 10 to 20 m s^{-1}; additional shear above 20 m s^{-1} has little effect on supercell probability. This is consistent with the formulation of SCP for which shear above a 20 m s^{-1} has little impact on supercell probability. In contrast to SCP, supercell likelihood increases in the SVM with increasing EBWD below 10 m s^{-1} and a similar but much smaller change in probability is observed in the ANN and GBT.

As shown in Fig. 9g, all three models show a large increase in supercell likelihood as ESRH increases from 0 to 100 m^2 s^{-2}, after which the curve flattens and additional ESRH above 100 m^2 s^2 has little to no effect on supercell likelihood. The GBT shows no increase in supercell probability as ESRH increases to 300 m^2 s^{-2} and the probability only increases by few percent in the SVM and GBT. These results generally agree with previous research that highlights the importance of ESRH for supercells by serving as an estimate of the streamwise vorticity ingested by a storm which is the primary source of midlevel rotation in supercells (Davies-Jones 1984) and representing the amount of effective inflow-layer storm-relative

flow (Peters et al. 2020). The finding that ESRH above 100 m^2 s^2 has minimal effect on the likelihood of supercell storm mode is something that is not incorporated into EHI or SCP and could result in false alarms from these indices in situations with high storm-relative helicity.

Figure 9h shows the ALE of ELSRW. The ANN and SVM both show a generally linear increase of 8% in supercell probability as ELSRW increases from near 0 m s^{-1} to more than 35 m s^{-1}. The GBT shows the same overall relationship between ELSRW and supercell likelihood, but the effect of additional ELSRW decreases as ELSRW increases, with the largest change in supercell likelihood occurring as ELSRW increases from 8 to 11 m s^{-1}. These results agree with Rasmussen and Blanchard's (1998) as well as Houston et al.'s (2008) observations of higher upper-tropospheric flow in supercell environments, and the theoretical argument that faster upper-tropospheric wind increases the evacuation of hydrometeors form the updraft.

Figure 9i shows that the three models have different impacts of ESRW. The ANN shows a slight decrease in supercell likelihood as ESRW increases from 7 to 9 m s^{-1} but any changes in ESRW above 9 m s^{-1} have negligible impact on supercell likelihood. The SVM shows a consistent decrease in supercell likelihood as ESRW increases. GBT shows a different relationship, with a slight decrease in supercell likelihood as ESRW increases to from 7 to 9 m s^{-1}. As ESRW increases from 9 and 11 m s^{-1} supercell likelihood increases by 7%. Any increase in ESRW above 11 m s^{-1} has no effect on supercell likelihood in the GBT model. The GBT relationship between ESRW and supercell likelihood seems to be the most realistic, agreeing with Peters et al. (2020) on changes in ESRW impacting the transition to supercellular storm mode.

d. Model interpretation summary

Overall, the relationships between input variables and predicted supercell likelihood match the understanding of their influence on storm mode based on previous research with a few exceptions. The deviations from previous research noted in some of the thermodynamic variables (MUCAPE, LLCAPE, and MUCIN) can be explained by the sampling strategy and are likely not limitations of the machine learning models.

The only other relationship that does not agree with previous research is between ESRW and storm mode. The effect of ESRW on storm morphology is not yet fully understood. Peters et al. (2020) theorized that as ESRW increases, convergence near the cloud base results in changes to updraft properties that favor supercell formation. The GBT appears to model a similar relationship, whereas the SVM and ANN do not, and variable importance rankings for the SVM and ANN models suggest ESRW is not an important variable in determining the likelihood of supercells. This could be because of the strong correlation between ESRW and EBWD (Fig. 1). It is possible that the primary impact on storm mode is contained in EBWD and once that is accounted for by the ANN and SVM, the ESRW has little effect. Alternatively, ESRW could play an important role in storm morphology, and GBT captures this more accurately than the ANN or SVM due to limitations in those models. For example, the SVM implemented here uses a linear decision boundary which can reduce the negative effects of higher dimensional data; however, it also reduces the model's ability to model complex relationships (Hastie et al. 2009).

4. Conclusions

This study developed machine learning models to predict if discrete supercell or discrete nonsupercell storm mode was favored based on sounding derived environmental parameters. These models were trained and evaluated on a dataset consisting of ~1000 RUC-2 model proximity soundings from near-storm environments. Results indicate that all three of the machine learning models developed could better discriminate storm mode than EHI and SCP indices by a statistically significant margin.

The dataset used to train and test the models results in several potential limitations to the study. First, every possible supercell environment cannot be represented in the dataset used to train and test the models. As such, caution should be used when applying these models, particularly when environments are outside the range of variable values used in this study (Fig. 1). A second potential limitation resulting from the dataset is related to the ratio of supercells to nonsupercell storms. The dataset used in this study contains a much higher ratio of supercells to nonsupercells than climatology would suggest. This could limit the ability of the machine learning models by biasing them to predict supercellular storm mode too frequently. However, when machine learning model probabilities are adjusted and performance results weighted to replicate a more climatologically reasonable ratio show that the machine learning models still outperform SCP and EHI by a statistically significant margin, although this process effectively shrinks the size of the already small test dataset, and has the potential to limit the robustness of the conclusions.

The dataset used also means performance results (e.g., FAR, POD) could be misleading compared to performance in an operational forecasting environment. Even when using performance metrics that perform well for imbalanced data, the skew can influence the metric values (Lampert and Gançarski 2014; Boyd et al. 2012) and care should be taken when interpreting performance metrics. Specifically, the values of the metrics used in this study (MCC, POD, FAR, and AUROCC) are dependent on the dataset to which they are applied. Additionally, the dataset used to train and evaluate the machine learning models is a collection of idealized situations, (e.g., the database only contains discrete storms while in reality other types of storms occur and supercells can be embedded within or grow into larger convective systems rather than simply staying either a discrete supercell or discrete nonsupercell). These factors mean that the POD or FAR listed in this study are likely unrealistic of what could be expected in a true operational forecast environment.

Examination of variable importance rankings showed general agreement with previous work on the importance of EBWD and ESRH. The importance of EBWD and ESRH is reflected both in permutation-based variable importance

rankings and in the magnitude of ALE, which can also be used as a measure of variable importance (Greenwell et al. 2018). EBWD ranked highest the variable importance rankings and showed the largest magnitude of ALE for all models. ESRH was ranked second in variable importance and had the second largest ALE magnitudes for all models. The low ranking of both MUCAPE and LLCAPE is not particularly surprising considering the fact that the machine learning models, by necessity due to the sampling strategy used to generate the training data, generate a supercell probability conditioned on the occurrence of a sustained thunderstorm. Expanding the dataset to incorporate instances of failed convection initiation or sustenance could be done in future research and would allow for the development of machine learning models that more accurately represent the impact of thermodynamics variables and predict an absolute probability rather than a conditional probability of supercell storm mode.

Individual variable effects showed that the machine learning models generated variable relationships that were generally consistent with previous research but there were several exceptions. The models, particularly the SVM and ANN, indicated relatively low impact of ESRW. It is uncertain if the importance of ESRW noted by Peters et al. (2020) is the result of the correlation of ESRW with EBWD or if the SVM and ANN failed to accurately capture the influence of ESRW on storm mode. As a result, future research into the impact of ESRW is warranted. Additionally, for the SVM, an inverse relationship between model performance and both MUCAPE and ESRW is seen. This could be the result of the simplicity of the linear SVM model used, but since the variable importance of these variables were lowest in the SVM the impact on the predictions appears to be minimal. For the most important variables in all models, EBWD and ESRH, the individual variable effects were physically reasonable. EBWD shows a relationship similar to SCP where the greatest change in supercell likelihood occurs as EBWD increases from 10 to 20 m s^{-1}. ESRH shows a similar relationship where the greatest increase in supercell probability occurs as ESRH increases to 100 m^2 s^{-2}. Additional ESRH above this threshold has little effect, which is not reflected in the calculation of EHI or SCP. While not the focus of this study, these findings could be used to modify the weighting of EBWD and ESRH in parameters like SCP to improve parameter performance. In summary, model interpretability results suggest the relationships learned by the machine learning models were generally physically reasonable.

Overall, this study demonstrates the ability of machine learning models to strongly discriminate between storm modes based on physically reasonable relationships between environmental variables and storm mode. Machine learning models like those developed in this study have potential to be used directly as a forecast tool similar to how SCP or EHI are used, or the knowledge gained (e.g., the most important variables and how changes in those variables influence storm mode) can be utilized by forecasters when evaluating environmental conditions. Due to their potential impact on operational forecasting, future work should continue the development of machine learning–based forecasting tools and evaluate their performance in a more realistic operational environment through cases studies.

Acknowledgments. The authors thank the National Science Foundation for providing partial funding for this project under Award AGS-1824649. The comments of three anonymous reviewers greatly improved the manuscript. The authors would also like to thank Matthew B. Wilson for his help in processing the RUC data used in this study, as well as University of Nebraska–Lincoln's Holland Computing Center (HCC) for providing the computing resources for this study.

Data availability statement. The dataset analyzed in this study were a reanalysis of an existing dataset detailed in Thompson et al. (2007). A web-based implementation of the machine learning models created in this study and a code repository can be found at https://eas2.unl.edu/~sshield/ml_supercell.

REFERENCES

Aggarwal, C. C., 2018: *Neural Networks and Deep Learning*. Springer, 497 pp.

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/rssb.12377.

Belle, V., and I. Papantonis, 2021: Principles and practice of explainable machine learning. *Front. Big Data*, **4**, 688969, https://doi.org/10.3389/fdata.2021.688969.

Benjamin, S. G., and Coauthors, 2004: An hourly assimilation-forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495–518, https://doi.org/10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2.

Blumberg, W. G., K. T. Halbert, T. A. Supinie, P. T. Marsh, R. L. Thompson, and J. A. Hart, 2017: SHARPpy: An open-source sounding analysis toolkit for the atmospheric sciences. *Bull. Amer. Meteor. Soc.*, **98**, 1625–1636, https://doi.org/10.1175/BAMS-D-15-00309.1.

Boehmke, B., and B. Greenwell, 2020: *Hands-On Machine Learning with R*. Taylor and Francis, 488 pp.

Boughorbel, S., F. Jarray, and M. El-Anbari, 2017: Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLOS ONE*, **12**, e0177678, https://doi.org/10.1371/journal.pone.0177678.

Boyd, K., V. S. Costa, J. Davis, and C. D. Page, 2012: Unachievable region in precision-recall space and its effect on empirical evaluation. *Proc. 29th Int. Conf. on Machine Learning*, Edinburgh, Scotland, ICML.

Brooks, H. E., C. A. Doswell III, and R. B. Wilhelmson, 1994: The role of midtropospheric winds in the evolution and maintenance of low-level mesocyclones. *Mon. Wea. Rev.*, **122**, 126–136, https://doi.org/10.1175/1520-0493(1994)122<0126:TROMWI>2.0.CO;2.

Brown, M., and C. J. Nowotarski, 2019: The influence of lifting condensation level on low-level outflow and rotation in simulated supercell thunderstorms. *J. Atmos. Sci.*, **76**, 1349–1372, https://doi.org/10.1175/JAS-D-18-0216.1.

Bunkers, M. J., B. A. Klimowski, J. W. Zeitler, R. L. Thompson, and M. L. Weisman, 2000: Predicting supercell motion using a new hodograph technique. *Wea. Forecasting*, **15**, 61–79, https://doi.org/10.1175/1520-0434(2000)015<0061:PSMUAN>2.0.CO;2.

Chollet, F., and Coauthors, 2015: Keras. GitHub, accessed January 2021, https://github.com/fchollet/keras.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29**, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1.

——, and Coauthors, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

——, M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Davies, J., 2006: Tornadoes in environments with small helicity and/or high LCL heights. *Wea. Forecasting*, **21**, 579–594, https://doi.org/10.1175/WAF928.1.

Davies-Jones, R. P., 1984: Streamwise vorticity: The origin of updraft rotation in supercell storms. *J. Atmos. Sci.*, **41**, 2991–3006, https://doi.org/10.1175/1520-0469(1984)041<2991:SVTOOU>2.0.CO;2.

Duchi, J., E. Hazan, and Y. Singer, 2011: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, **12**, 2121–2159.

Duda, J. D., and W. A. Gallus, 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206, https://doi.org/10.1175/2009WAF2222338.1.

Efron, B., and R. J. Tibshirani, 1994: *An Introduction to the Bootstrap*. CRC Press, 456 pp.

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. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Gagne, D. J., S. Haupt, and D. Nychka, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Géron, A., 2017: *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 545 pp.

Greenwell, B. M., B. C. Boehmke, and A. J. McCarthy, 2018: A simple and effective model-based variable importance measure. arXiv, 1805.04755, https://arxiv.org/abs/1805.04755.

Gropp, M. E., and C. E. Davenport, 2018: The impact of the nocturnal transition on the lifetime and evolution of supercell thunderstorms in the Great Plains. *Wea. Forecasting*, **33**, 1045–1061, https://doi.org/10.1175/WAF-D-17-0150.1.

Haberlie, A. M., and W. S. Ashley, 2018: A method for identifying midlatitude mesoscale convective systems in radar mosaics. Part I: Segmentation and classification. *J. Appl. Meteor. Climatol.*, **57**, 1575–1598, https://doi.org/10.1175/JAMC-D-17-0293.1.

Hart, J. A., and W. Korotky, 1991: The SHARP workstation v1.50 users guide. NOAA National Weather Service Doc., 30 pp.

Hastie, T., R. Tibshirani, and J. H. Friedman, 2009: *Elements of Statistical Learning*. Springer, 745 pp.

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.

Houston, A. L., 2016: The sensitivity of deep ascent of cold-pool air to vertical shear and cold-pool buoyancy. *Electron. J. Severe Storms Meteor.*, **11** (3), https://ejssm.org/archives/2016/vol-11-3-2016/.

——, R. L. Thompson, and R. Edwards, 2008: The optimal bulk wind differential depth and the utility of the upper-tropospheric storm-relative flow for forecasting supercells. *Wea. Forecasting*, **23**, 825–837, https://doi.org/10.1175/2008WAF2007007.1.

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2020: Classifying convective storms using machine learning. *Wea. Forecasting*, **35**, 537–559, https://doi.org/10.1175/WAF-D-19-0170.1.

Kaufman, S., S. Rosset, C. Perlich, and O. Stitelman, 2012: Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discovery Data*, **6**, 1–21, https://doi.org/10.1145/2382577.2382579.

Kirkpatrick, C., E. W. McCaul Jr., and C. Cohen, 2011: Sensitivities of simulated convective storms to environmental CAPE. *Mon. Wea. Rev.*, **139**, 3514–3532, https://doi.org/10.1175/2011MWR3631.1.

Klambauer, G., T. Unterthiner, A. Mayr, and S. Hochreiter, 2017: Self-normalizing neural networks. arXiv, 1706.02515, https://arxiv.org/abs/1706.02515.

Kuhn, M., and K. Johnson, 2013: *Applied Predictive Modeling*. Springer, 600 pp.

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

——, ——, C. R. Homeyer, D. J. Gange, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

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

Lampert, T. A., and P. Gançarski, 2014: The bane of skew. *Mach. Learn.*, **97**, 5–32, https://doi.org/10.1007/s10994-013-5432-x.

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.

Markowski, P. M., and Y. P. Richardson, 2009: Tornadogenesis: Our current understanding, forecasting considerations, and questions to guide future research. *Atmos. Res.*, **93**, 3–10, https://doi.org/10.1016/j.atmosres.2008.09.015.

McCaul, E. W., Jr., and M. L. Weisman, 2001: The sensitivity of simulated supercell structure and intensity to variations in the shapes of environmental buoyancy and shear profiles. *Mon. Wea. Rev.*, **129**, 664–687, https://doi.org/10.1175/1520-0493(2001)129<0664:TSOSSS>2.0.CO;2.

McGovern, A., R. Lagerquist, D. J. Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Molnar, C., 2018: Interpretable machine learning: A guide for making black box models explainable. GitHub, https://christophm.github.io/interpretable-ml-book/.

Nowotarski, C. J., and A. A. Jensen, 2013: Classifying proximity soundings with self-organizing maps toward improving supercell and tornado forecasting. *Wea. Forecasting*, **28**, 783–801, https://doi.org/10.1175/WAF-D-12-00125.1.

——, and E. Jones, 2018: Multivariate self-organizing map approach to classifying supercell tornado environments using near-storm, low-level wind and thermodynamic profiles. *Wea. Forecasting*, **33**, 661–670, https://doi.org/10.1175/WAF-D-17-0189.1.

NTSB, 2010: Weather-related aviation accident study 2003–2007. NTSB Rep., 71 pp., https://www.asias.faa.gov/i/studies/2003-2007weatherrelatedaviationaccidentstudy.pdf.

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

Peters, J. M., C. J. Nowotarski, and H. Morrison, 2019: The role of vertical wind shear in modulating maximum supercell updraft velocities. *J. Atmos. Sci.*, **76**, 3169–3189, https://doi.org/10.1175/JAS-D-19-0096.1.

——, ——, J. P. Mulholland, and R. L. Thompson, 2020: The influences of effective inflow layer streamwise vorticity and storm-relative flow on supercell updraft properties. *J. Atmos. Sci.*, **77**, 3033–3057, https://doi.org/10.1175/JAS-D-19-0355.1.

Rasmussen, E. N., 2003: Refined supercell and tornado forecast parameters. *Wea. Forecasting*, **18**, 530–535, https://doi.org/10.1175/1520-0434(2003)18<530:RSATFP>2.0.CO;2.

——, and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164, https://doi.org/10.1175/1520-0434(1998)013<1148:ABCOSD>2.0.CO;2.

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

Saerens, M., P. Latinne, and C. Decaestecker, 2002: Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.*, **14**, 21–41, https://doi.org/10.1162/089976602753284446.

Sander, J., J. Eichner, E. Faust, and M. Steuer, 2013: Rising variability in thunderstorm-related U.S. losses as a reflection of changes in large-scale thunderstorm forcing. *Wea. Climate Soc.*, **5**, 317–331, https://doi.org/10.1175/WCAS-D-12-00023.1.

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

Shield, S. A., S. M. Quiring, J. V. Pino, and K. Buckstaff, 2021: Major impacts of weather events on the electrical power delivery system in the United States. *Energy*, **218**, 119434, https://doi.org/10.1016/j.energy.2020.119434.

Snook, N., and M. Xue, 2008: Effects of microphysical drop size distribution on tornadogenesis in super-cell thunderstorms. *Geophys. Res. Lett.*, **35**, L24803, https://doi.org/10.1029/2008GL035866.

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.

Steinkruger, D., P. Markowski, and G. Young, 2020: An artificially intelligent system for the automated issuance of tornado warnings in simulated convective storms. *Wea. Forecasting*, **35**, 1939–1965, https://doi.org/10.1175/WAF-D-19-0249.1.

Thompson, R. L., R. Edwards, and J. A. Hart, 2002a: An assessment of supercell and tornado forecast parameters with RUC-2 model close proximity soundings. *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., P12.3, https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_46931.htm.

——, ——, and ——, 2002b: Evaluation and interpretation of the supercell composite and significant tornado parameters at the Storm Prediction Center. *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., J3.2, https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_46942.htm.

——, ——, ——, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, ——, and C. M. Mead, 2004: An update to the supercell composite and significant tornado parameters. *22nd Conf. on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., P8.1, https://ams.confex.com/ams/11aram22sls/techprogram/paper_82100.htm.

——, C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115, https://doi.org/10.1175/WAF969.1.

VanderPlas, J., 2016: *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, 548 pp.