Airborne Radar Quality Control with Machine Learning

ALEXANDER J. DESROSIERS[a] AND MICHAEL M. BELL[a]

[a] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*

(Manuscript received 31 July 2023, in final form 9 December 2023, accepted 14 December 2023)

ABSTRACT: Airborne Doppler radar provides detailed and targeted observations of winds and precipitation in weather systems over remote or difficult-to-access regions that can help to improve scientific understanding and weather forecasts. Quality control (QC) is necessary to remove nonweather echoes from raw radar data for subsequent analysis. The complex decision-making ability of the machine learning random-forest technique is employed to create a generalized QC method for airborne radar data in convective weather systems. A manually QCed dataset was used to train the model containing data from the Electra Doppler Radar (ELDORA) in mature and developing tropical cyclones, a tornadic supercell, and a bow echo. Successful classification of ~96% and ~93% of weather and nonweather radar gates, respectively, in withheld testing data indicate the generalizability of the method. Dual-Doppler analysis from the genesis phase of Hurricane Ophelia (2005) using data not previously seen by the model produced a comparable wind field to that from manual QC. The framework demonstrates a proof of concept that can be applied to newer airborne Doppler radars.

SIGNIFICANCE STATEMENT: Airborne Doppler radar is an invaluable tool for making detailed measurements of wind and precipitation in weather systems over remote or difficult to access regions, such as hurricanes over the ocean. Using the collected radar data depends strongly on quality control (QC) procedures to classify weather and nonweather radar echoes and to then remove the latter before subsequent analysis or assimilation into numerical weather prediction models. Prior QC techniques require interactive editing and subjective classification by trained researchers and can demand considerable time for even small amounts of data. We present a new machine learning algorithm that is trained on past QC efforts from radar experts, resulting in an accurate, fast technique with far less user input required that can greatly reduce the time required for QC. The new technique is based on the random forest, which is a machine learning model composed of decision trees, to classify weather and nonweather radar echoes. Continued efforts to build on this technique could benefit future weather forecasts by quickly and accurately quality-controlling data from other airborne radars for research or operational meteorology.

KEYWORDS: Aircraft observations; Data quality control; Radars/radar observations; Machine learning

1. Introduction

Airborne Doppler radar has a rich history of advancing knowledge in the meteorology community through targeted and close-range data collection of three-dimensional wind and precipitation features in weather systems over land or ocean. Despite the utility provided by airborne Doppler radar, the analysis process comes with challenges unique to the platform. For accurate dual-Doppler wind synthesis to take place, quality control (QC) is required to remove the motion of the aircraft carrying the radar, establish Earth-relative locations of radar echoes, and remove any nonmeteorological echoes from the data (Lee et al. 2003). Although several advances have been made on the platform motion issue to correct errors in the aircraft inertial navigation system (Testud et al. 1995; Bosart et al. 2002; Cai et al. 2018), the removal of nonmeteorological data has received less attention. Previous attempts to automate the process have shown some success (Gamache et al. 2008; Bell et al. 2013) but existing techniques provide a suboptimal classification of weather and nonweather echoes, either removing too much weather data in real-time applications or requiring additional time-consuming manual QC to produce high-quality wind analyses for research. In the current study, we improve this process through the use of complex, automated decision-making available via machine learning techniques.

Although recent progress has been made with QC techniques for ground-based radars (Tang et al. 2020; Ośródka and Szturc 2022), older threshold-reliant QC techniques for airborne Doppler radar (Gamache et al. 2008; Bell et al. 2013) remain the state-of-the-art methods for airborne radar observations in tropical (Fischer et al. 2022) and midlatitude (Stechman et al. 2020) convective systems. A full automation of airborne radar QC was developed for P-3 Hurricane Hunter tail Doppler radar (TDR) data in real time (Gamache et al. 2008) based on a rules-based approach that sets thresholds for data retention. A similar approach was developed by Bell et al. (2013) for research analysis, significantly reducing the effort required in the QC process prior to interactive manual editing in the Solo editing software from the National Center for Atmospheric Research (NCAR) (Oye et al. 1995). Bell et al. (2013) provided three different levels of the QC algorithm (high, medium, and low) that represented a trade-off between how much "good" weather data are retained versus how much "bad" nonweather data are removed. If the thresholds are increased to the "high" level, then 95% of the nonweather data can be removed by the algorithm, but 15% of the valuable weather data are discarded with it. Similarly,

Corresponding author: Alexander J. DesRosiers, adesros@rams.colostate.edu

DOI: 10.1175/AIES-D-23-0064.1 e230064

"low" thresholds allow for retention of 95% of the good data, but also leaves 20% of the bad data that must be removed manually by a trained expert. An improved QC algorithm requires more complex decision-making than the thresholding of the approaches utilized in previous techniques to identify and remove nonmeteorological data.

To develop an automated airborne Doppler radar QC methodology appropriate for a wide variety of weather systems, Bell et al. (2013) used an extensive manually QCed dataset from the NCAR Electra Doppler Radar (ELDORA) (Hildebrand et al. 1996) for development and verification. The ELDORA dataset used in Bell et al. (2013) and herein was collected from several field campaigns investigating different convective phenomena including both mature (Hence and Houze 2008) and developing tropical cyclones (Bell and Montgomery 2010), tornadoes (Wakimoto and Liu 1998), and bow echoes (Wakimoto et al. 2006). The field campaigns that gathered these data were the Hurricane Rainband and Intensity Change Experiment (RAINEX), Tropical Cyclone Structure (T-PARC/TCS08), Verification of the Origins of Rotation in Tornadoes Experiment (VORTEX), and Bow Echo and Mesoscale Convective Vortex Experiment (BAMEX), respectively. The diverse dataset collected by ELDORA and QCed by different experienced radar meteorologists make it a prime candidate for developing an improved QC method for airborne Doppler radar data in convective environments.

In the new approach employed here, we use machine learning (ML) to provide complex decision-making capable of discriminating between sometimes subtle distinctions in weather and nonweather echoes. ML methods have been applied to a range of tasks in many disciplines, with the advantages of ML currently being realized in the field of meteorology (Boukabara et al. 2019) to advance remote sensing retrievals, data assimilation, model physics calculation, forecasting, and data QC. Although scanning geometry and wavelength create considerable differences between airborne and ground-based radar data, the recent successful use of ML for ground-based radar QC (Lakshmanan et al. 2014) further motivated an attempt with airborne data. A relatively straightforward ML technique, the random forest (Louppe 2014), creates an ensemble of decision trees that can be used to classify each radar gate individually. The capabilities of ML were used to improve on the current QC algorithms for the purpose of producing a research quality dual-Doppler wind synthesis from TDR data with minimal effort. Section 2 describes the data used and methodology employed to train and test a random-forest model. Section 3 presents and evaluates results from testing data withheld from the training set and a separate example case. Discussion of the results takes place in section 4, followed by conclusions in section 5.

2. Data and method

Training and testing of the random-forest (RF) model utilized the same dataset from Bell et al. (2013), which consists of 6, 11, 22, and 9 min of TDR data collected by ELDORA during the RAINEX, TPARC/TCS08, BAMEX, and VORTEX field experiments, respectively. The combined dataset contains approximately 87.9 million total radar gates, excluding gates

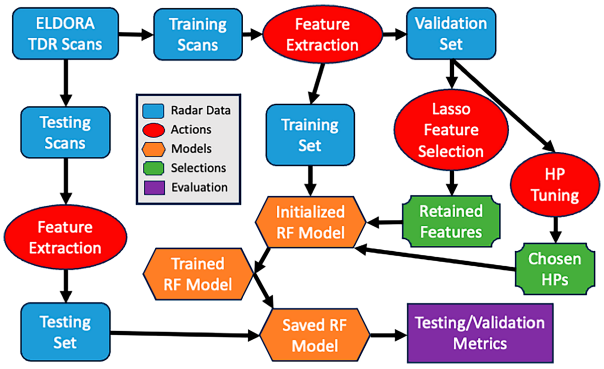

FIG. 1. Flowchart of the experimental setup to create a random-forest model for airborne Doppler radar QC.

devoid of data, containing a variety of weather and nonweather echoes. All data used in training was QCed via a combination of automated and manual techniques by radar experts, providing an extensive training dataset from their dedicated efforts. The variety of airborne data in the training dataset is representative of different convective environments and slightly different approaches to the interactive QC process depending on the person performing QC. The QCed data provided the echo classification required for training a machine learning model to emulate the interactive process with greater speed and minimal user input.

A flowchart of the method, shown in Fig. 1, visually represents the technique. Predictors for the model extracted from the ELDORA dataset contain the radar range, radar reflectivity in reflectivity decibels (dBZ), Doppler velocity in meters per second, normalized coherent power (NCP; unitless), and derived mathematical quantities from the raw radar moments. NCP measures the quality of the Doppler velocity measurement by the coherence of the phase shifts within the signal and is normalized to range from 0 to 1, where 1 is likely high-quality data (similar to the signal quality index produced by some radar signal processors). The derived quantities include the average and standard deviation for the radar moments relative to neighboring radar gates, isolation of a radar gate, probability of a radar gate being affected by the ground, and range normalized through division by aircraft altitude. Range normalized to the aircraft altitude may help identify artifacts due to the reflection of the TDR's antenna-pattern sidelobes off of the ground (Bell et al. 2013). The isolation parameter calculates the ratio of neighboring radar gates with echo to the total neighboring gates in a square area centered on the gate and helps identify spurious "speckles" unlikely to be associated with weather systems. The isolation predictor is less useful in ELDORA scans that are characterized by continuous swaths of data (Fig. 6c) but may be useful when applying this method to newer airborne Doppler radars. The probability of a ground gate is adapted from geometric radar beam considerations in Testud et al. (1995) and aims to identify radar gates that may be affected by surface echo. The original formula produces a binary classification of ground contamination based on half-power beamwidth of the radar antenna and

the scanning geometry. The current probability of ground gate formula improves upon this estimate by producing a probabilistic estimate of the amount of ground contamination in a radar gate assuming a full Gaussian beam volume. In total, 14 predictors (Table 1) are extracted for each radar gate in the dataset and will be considered for model inclusion.

The mined predictors make up the **X** array that is used to classify data as weather or nonweather. The classification made by a human radar expert was stored in the **Y** array as a binary flag for each radar gate. If a radar gate was present in the pre-QC field and post-QC field, it was considered weather data in the **Y** array and assigned a value of 1. If it had been deleted in QC, it was assigned the nonweather class and recorded as 0. Radar gates in ELDORA scans containing no data were excluded from the dataset as they can automatically be left blank during QC. Gates in the dataset that had NCP values of 0.2, indicating poor signal quality, or lower or a value of 1 for probability of ground gates, indicating subterranean echoes, were also excluded. Data meeting either of these criteria are easily identifiable as likely nonweather and excluding them allows the RF model to focus on more difficult classifications. Approximately 45.9 million total radar gates remain in the ELDORA dataset for this study after the above exclusion criteria are met. A successfully trained model classifies radar gates as weather or nonweather data using the predictors in the **X** array to predict the binary classification in the **Y** array.

a. Training, validation, and testing split

The ELDORA dataset was split into training, validation, and testing sets with a percentage split of roughly 72%, 8%, and 20% respectively. Partitioning the ELDORA dataset into the aforementioned sets starts by removing and setting aside the testing set for evaluation of the final tuned RF model. A random split of the data with shuffling is not used on the **X** and **Y** arrays from the dataset. Shuffling the data allows for data leakage due to spatial autocorrelation of neighboring radar gates in the same scan. When radar gates that originally neighbored each other in a scan end up split with one in the training set and the other in testing, the model has a considerable advantage classifying the testing gate given it has seen something with nearly identical predictors in training. There may also be temporal autocorrelation with consecutive radar scans containing similar data, but this cannot be completely avoided without reducing the size of the valuable dataset, so the spatial correlation will be addressed as the main priority in dataset splitting. To separate the testing set and mitigate spatial autocorrelation, scans were set aside for testing before the feature extraction script produced the training and testing sets. The ELDORA dataset contains 1780 TDR scans in total across all four observation cases; 20% of this value is 356. Dividing this figure by four shows that a representative training set across cases should be composed of 89 TDR scans from each case. To make sure the variability with time of each case is represented, 89 was rounded up to 90 scans and 30 consecutive scans were taken from the beginning, middle, and end of each observational case. Selecting scans in groups is a purposeful step to limit temporal autocorrelation that may still occur, but mainly at selection boundaries. Selected scans were set aside and processed for predictor and classification arrays (**X** and **Y**), which were stored in the h5 file data format to make up the testing set.

TABLE 1. All features collected for potential use in the machine learning model and their abbreviations used in this study. Feature names followed by an asterisk are the features retained for the model.

Feature abbreviation	Feature name
VT	Doppler velocity
ZZ	Raw reflectivity
NCP	Normalized coherent power*
ALT	Altitude of radar gate*
AVT	Avg of velocity
AZZ	Avg of reflectivity
ANCP	Avg of NCP
SVT	Std dev of velocity*
SZZ	Std dev of reflectivity
SNCP	Std dev of NCP
ISO	Isolation parameter
PGG	Probability of ground gate*
RG	Radar range*
NRG	Radar range normalized to aircraft altitude

The remaining scans after the testing set was removed were also run through feature extraction to create an initial training set that was then divided into validation and training sets. Every 10th observation in the training set was removed and pooled into a validation set for model tuning and feature selection. The every-10th-selection method results in the percentage makeups of 72% and 8%, respectively, for training and validation. Although using every 10th observation in the validation set allows for potential spatial autocorrelation of neighboring gates in training to prepare the model unfairly well for the validation set, this choice was purposeful to increase variability in the validation set. An additional testing set beyond the ELDORA dataset used thus far is introduced later. The additional testing set will reverify generalizability while allowing the validation set that is well spread through training to be a representative model tuning tool.

b. Feature selection

Reducing the number of input features that the RF model can use to make a classification decision helps to maintain generalizability and reduce the risk of model overfitting. Feature selection is performed by running 100 different logistic regression classification models using all 14 features, or predictors, while varying a lasso (L1) regularization penalty. In this experiment, the validation set is split randomly into training and testing sets containing 80% and 20% of the samples respectively so that each logistic regression model can be evaluated for performance. The 100 models are initialized with unique lambda L1 penalty values ascending in logarithmic space from 10^{-8} to 1. Lambda sets the inverse of regularization strength. At smaller values, the lasso regularization penalty is large and the logistic regression model is forced to make classifications

with less input features as their coefficients are forced to zero. To select features, a lambda value is objectively selected where there is an acceptable trade-off between model performance and reduced input features required for classification. Each logistic regression model is evaluated with the testing data and scored with two metrics. The first of which is a weighted F_1 score. The F_1 score consists of a combined measure of precision and recall from true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{and}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{1}$$

When evaluating metrics for each data class, a positive is a determination that the data belongs to the class in question while a negative indicates the opposite. True or false denotes correctness of the model classification. Precision is the ratio of correctly predicted positive observations to the total count predictions for positive including those that were false. Recall is the ratio of correctly predicted positives to the total count of positives in the training set. The F_1 score combines precision and recall for each class of data resulting in a score of 0 to 1, with 1 indicating perfect precision and recall. The weighted aspect of the score takes the F_1 score for weather and nonweather data classes and averages them based on the balanced class weights calculated during model training. Weighting is used to get an accurate score despite class imbalance in the ELDORA dataset, which is roughly 3:1 with nonweather being the minority class after easily identifiable data likely to be nonweather were filtered out via NCP and probability of ground gates. The receiver operating characteristic area under the curve (ROC AUC) is also evaluated. The ROC AUC score describes classification performance with a maximum attainable value of 1. Plotting these metrics against lambda in Fig. 2a show how the addition of features impacts performance metrics.

Initially the rise in performance is quite steep as the first predictor (NCP) is added in just after $\lambda = 10^{-6}$, but by around $\lambda = 10^{-4}$ performance has largely leveled out and does not improve much more. A zoomed-in look at the plot in Fig. 2b shows this and also identifies the lambda value selected ($\lambda \sim 2.48 \times 10^{-5}$). Plotting the regression coefficients at this lambda value (Fig. 2c) informs which features to retain for training the RF model. Features, or predictors, and their abbreviations are defined in Table 1. The logistic regression experiment results suggest the retention of NCP, altitude of radar gate (ALT), standard deviation of velocity (SVT), isolation parameter (ISO), probability of ground gate (PGG), and radar range (RG), with NCP being most important to data classification. Given the continuous nature of coverage in raw Eldora data (Fig. 6c) and the low importance of ISO, it is assumed the isolation parameter does not add much in terms of performance for this dataset and will be dropped from this list. The five predictors of NCP, altitude, standard deviation of velocity, probability of ground gate, and radar range are used to tune, and train the RF model. Although reflectivity is useful in visually identifying nonweather data during manual QC, the lasso regression results indicate it may not be necessary for this automated approach.

c. *Hyperparameter tuning*

Tuning, training, and testing of the model were performed with Python's SciKit-learn library (Pedregosa et al. 2011). Hyperparameters, which are set to control the learning process, are very important to maximize model performance. Choice of hyperparameters for a model determines its complexity. A model that is too complex is prone to overfitting to the training set at the expense of performance on unseen data. A model that is too simple fails to capture important characteristics of the data and underperforms on the classification task (Claesen and De Moor 2015). Tuning of the RF model was focused on varying two hyperparameters, the number of trees and the maximum depth of a tree. Adjusting the number of trees simply changes the count of decision trees in the random forest. Maximum depth assigns a limit to how far the trees can branch downward and further sort the data with finer distinctions. Testing the impacts of hyperparameter choices on model performance helps ensure the appropriate level of model complexity is achieved.

There is an unequal distribution of classes between weather and nonweather radar gates in the filtered ELDORA dataset, with a higher frequency of weather echoes. To combat the inequality, class weights used in training were balanced by making the weights of the classes inversely proportional to class frequencies in the data. The adjusted weights assigned greater importance in model training to the less common class of nonweather data. As a consequence of imbalance, accuracy by itself is not an effective metric with which to assess the model's binary classification performance. For example, a model that always classifies radar gates as the most common class would receive a deceivingly good accuracy score in a dataset consisting mainly of that class, regardless of the model not performing the classification task. The weighted F_1 score is used for evaluation again to avoid the pitfalls of simple accuracy scores and class imbalance.

The GridSearchCV functionality from SciKit-learn (Pedregosa et al. 2011) was utilized for iterative training and testing of RF models with different combinations of the two varied hyperparameters. The aim was to find which model exhibits a high weighted F_1 score while not unnecessarily increasing complexity to a point of diminishing return for the added complexity. To further prevent overfitting, a cross validation scheme was employed. Before each model is tested, the validation dataset was split without shuffling into the default count of five stratified groups, or folds, with class distributions similar to the full set. For each combination of hyperparameters, a model was trained on four folds, leaving the remaining one for testing. This step was repeated until each fold had been used for testing and five models were trained and scored. The weighted F_1 score given for each hyperparameter combination is an average of scores on withheld testing data from the five different models created with those specifications. Results of the hyperparameter tuning (Fig. 3), show that past a certain level of model complexity, there is not much performance to be

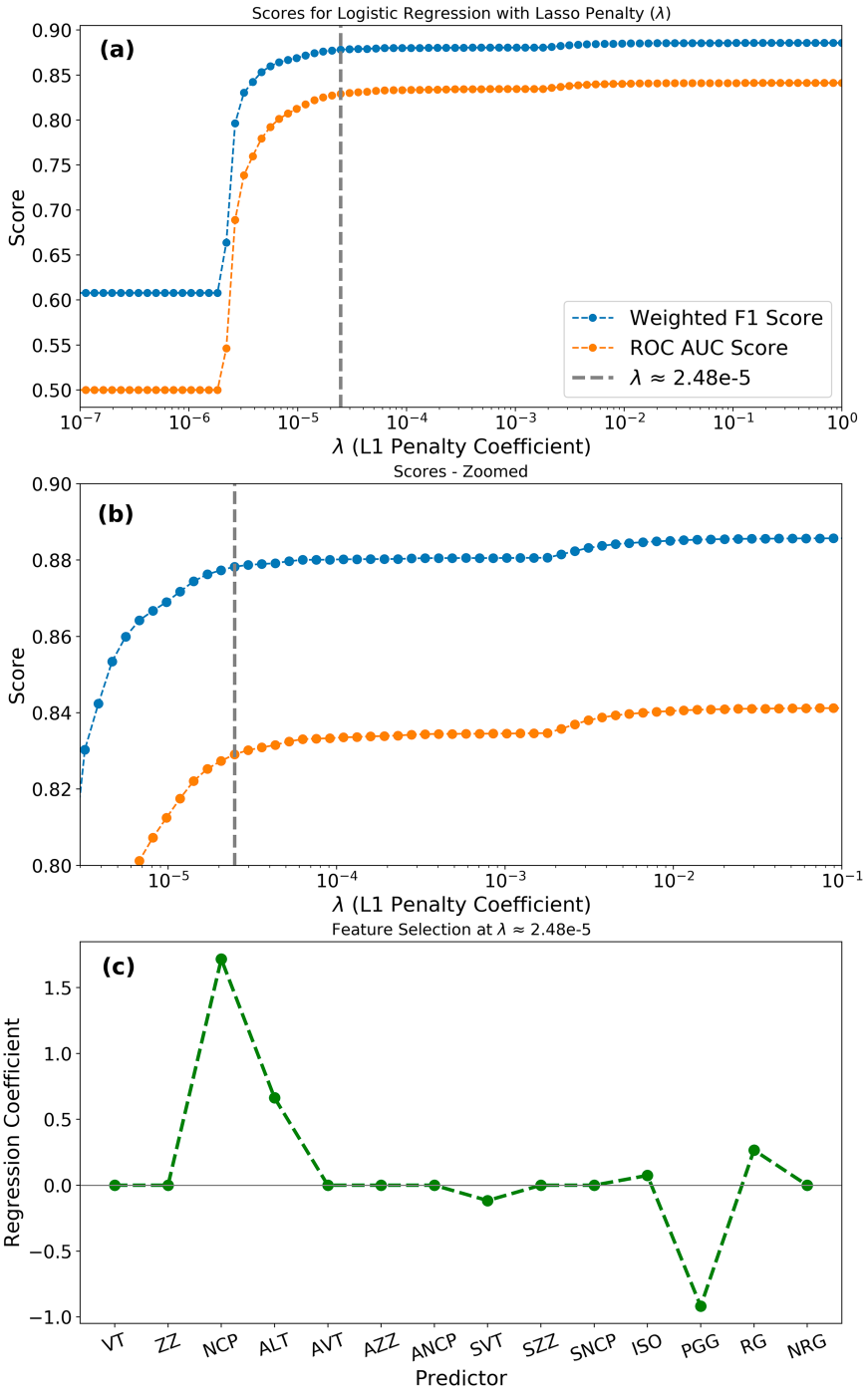

FIG. 2. (a) Logistic regression performance measured by weighted F_1 (blue) and ROC AUC scores (orange) as the L1 lasso penalty coefficient λ values change. (b) As in (a), but zoomed in closer to the chosen λ value. (c) Feature coefficients at the chosen λ value. Feature abbreviations are defined in Table 1.

gained. A larger parameter space extending out to 31 trees and a maximum depth of 20 was tested that confirmed the diminishing returns on performance continue (not shown). After a weighted F_1 score of ~0.945 was achieved further increases in score were very small. An additional test with 300 trees produced similar results indicating this plateau in performance continues with even greater model complexity. The hyperparameter combination of 21 trees with a maximum depth of 14 that achieved the score of ~0.945 was chosen to train and test a more thorough model with the larger training and testing datasets.

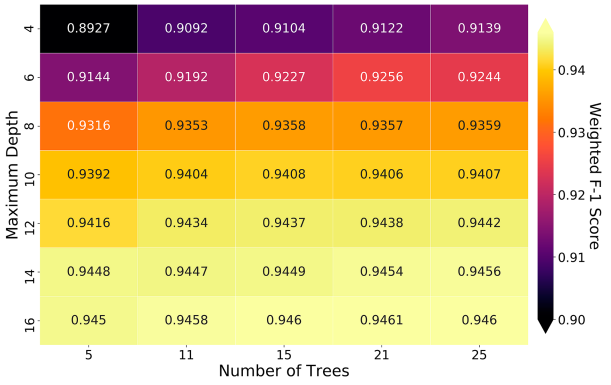

FIG. 3. Heat map depicting weighted F_1 score performance of each hyperparameter combination tested. Colors and values in each square indicate model performance for each combination of hyperparameters. The x axis is the number of trees, and the y axis is the depth to which decisions can be made in each tree.

(a) Training	Predicted Non-weather	Predicted Weather
Non-weather	9,046,253 (96.8%)	298,249 (3.2%)
Weather	1,170,342 (4.8%)	23,271,747 (95.2%)

(b) Validation	Predicted Non-weather	Predicted Weather
Non-weather	1,004,108 (96.7%)	33,900 (3.3%)
Weather	130,966 (4.8%)	2,585,094 (95.2%)

(c) Testing	Predicted Non-weather	Predicted Weather
Non-weather	2,553,320 (96.0%)	106,088 (4.0%)
Weather	401,393 (7.1%)	5,273,855 (92.9%)

(d) Feature Importance

Impurity Based	Rank & Score	Permutation	Rank & Score
SD of Velocity	1) 0.325	Altitude	1) 0.174
Altitude	2) 0.265	NCP	2) 0.159
NCP	3) 0.247	SD of Velocity	3) 0.067
Range	4) 0.092	Range	4) 0.054
Prob. of Ground Gate	5) 0.071	Prob. of Ground Gate	5) 0.009

FIG. 4. Confusion matrices outlining model performance on the (a) training, (b) validation, and (c) testing sets. Percentages are based upon the true classifications and sum to 100 across rows. (d) Predictors used by the model ranked on the basis of impurity- and permutation-based feature importance. SD is short for standard deviation of a quantity calculated with respect to neighboring points. Predictors in red are calculated for the model, and those in black are present in the raw data.

3. Results

Following steps outlined in the flowchart (Fig. 1), a generalized model for airborne Doppler radar QC of nonmeteorological data was trained and evaluated. In this section, model performance and predictor importance were evaluated using common metrics and ELDORA data not included in the ML training and testing datasets were used to test the viability of the model as a generalizable QC prototype.

a. Model performance metrics

Model performance metrics and information are outlined in Fig. 4. The confusion matrix (Figs. 4a–c) reports retention of weather data and removal of nonweather data for the training, validation, and testing sets. Nonweather data are removed in the training set at a 96.8% rate and weather data are retained at a 95.2% rate. Nearly identical performance is recorded with the validation set. Weather data in the withheld testing set are identified at a lesser rate of 92.9% in the testing set, but identification and removal of nonweather data is successful in 96% of instances in the dataset. Although a decrease in performance is expected, model performance on the testing set is still comparable to training and validation. The weighted F_1 score and ROC-AUC scores for the testing dataset are both ~0.94. Impurity-based feature importance for RF predictor variables is ranked in Fig. 4d. The feature importance is a numerical representation of the predictors that affect the most samples and split data most effectively (McGovern et al. 2019). For a predictor to receive a high feature importance relative to others, it should be used higher up in the tree to divide data and be effective at decreasing impurity in the groups into which the predictor splits the data. Permutation importance, also shown in Fig. 4d, evaluates how the model score declines when the chosen feature is randomly shuffled so as to render it useless to prediction. Altitude, NCP, and the standard deviation of velocity rank in the top 3 of both importance metrics. The rankings demonstrate the utility of the derived quantities (2 of the top 3; in red text) that provide context for a radar gate relative to its surroundings and aid in determining its likelihood of being weather data. Probability of ground gates ranks last in both methods. The low ranking is expected given it only contains useful information in a narrow strip of radar data near the ground where the radar beam approaches the surface. How the model uses these predictors is examined in further detail using data from an additional test case described in the following subsection.

b. QC test case

A rigorous test of the model to evaluate its suitability for airborne radar QC of nonmeteorological data is using it with data not included in the training or testing datasets used thus far. During the RAINEX field campaign on 6 September 2005, ELDORA was used to observe intense convective activity on the southern edge of the tropical depression that later became Hurricane Ophelia (Houze et al. 2009). TDR data collected by ELDORA during a closeup 15-min flyby leg of the convective feature were QCed with the RF model for comparison with the original interactive QC used in the published analysis. The MLQC scores showed only a slight decrease in performance from the statistics achieved with the testing dataset. Nonweather data were removed at a 94.5% rate while 91.8% of good weather data were retained. The weighted F_1 score and ROC-AUC scores for the Ophelia dataset are both ~0.93. The slight decrease is expected since the data were not only unseen by the model like the testing set,

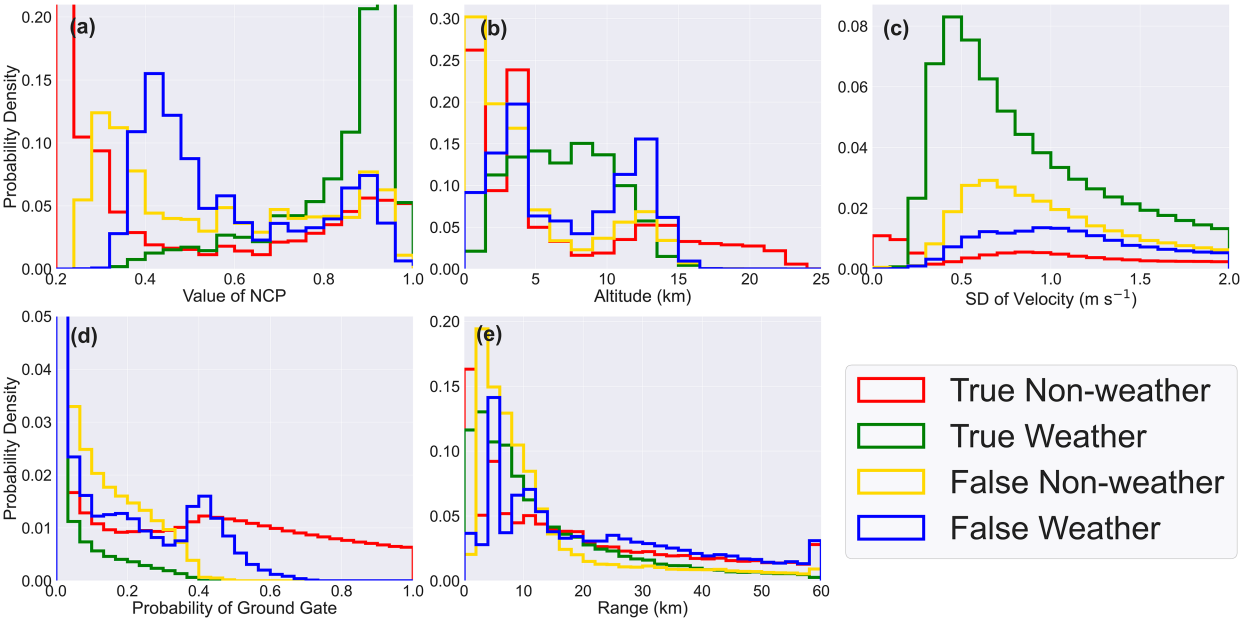

FIG. 5. PDFs of classifications for each predictor for the Ophelia (2005) test case: (a) NCP, (b) altitude, (c) SD of velocity, (d) PGG, and (e) range. The *x* axis indicates the value of the predictor, and the *y* axis displays PDF values that sum to 1. Different color outlines represent the classification of each radar gate, with red bars indicating true nonweather, green bars indicating true weather classification, yellow bars indicating false nonweather, and blue bars indicating false weather. True and false indicate the correctness of model classification.

but entirely separate from the dataset used thus far. Therefore, the algorithm has to generalize its decision-making beyond its training to some extent. Interactive QC on this dataset may also have slight differences in technique specific to the individual performing interactive QC as compared with those responsible for QC in the training dataset. To gain more insight into these results and how the model used the given predictors, histograms were created for the five predictors (Fig. 5) using a probability density function (PDF) that sums to one. Data are grouped by both correct and incorrect predictions of weather and nonweather creating four different histograms for each predictor.

Correct predictions of both good and bad data cluster near high and low bounds for NCP with false weather predictions tending to peak with higher NCP values than those of true nonweather predictions (Fig. 5a). Higher-altitude echoes tended to be nonweather while classification based on this metric alone becomes more complicated as gates approach the surface (Fig. 5b). Low standard deviations of velocity typically indicated good weather data (Fig. 5c). Probability of ground gates is only a useful metric near the surface where values are nonzero, so the *y* axis is capped at 0.05 to show most of the variability in this metric that occurs at lower values of the predictor (Fig. 5d). As the probability approaches higher values, the model exclusively successfully predicts nonweather gates associated with the ground. Radar gates are more likely to be nonweather at greater range (Fig. 5e). The PDFs of all analyzed predictors do not reveal clear and decisive cutoffs providing further evidence of the shortfalls of an approach using only rigidly applied thresholds as in Bell et al. (2013). This analysis makes clear the benefits of allowing a model to classify radar data based on information gathered from multiple predictors.

The end goal of airborne radar QC is to create a dual-Doppler wind synthesis from intersecting fore and aft scans through the weather of interest. One approach to creating this synthesis is through the use of a three-dimensional variational technique called "spline analysis at mesoscale utilizing radar and aircraft instrumentation" (SAMURAI) that yields a maximum likelihood estimate of the atmospheric state for the given radar observations (Bell et al. 2012; Foerster et al. 2014). SAMURAI was used to create analyses of the observed convective feature using data processed by both the original interactive QC and novel ML method. An example scan of Doppler velocity is shown in Figs. 6a–c, illustrating that the two methods produce a similar end product of a QCed velocity field in this test case. Only slight differences are detectable in the interactive QC scan when compared with machine learning QC (Figs. 6a,b). Figure 6 shows the SAMURAI wind analyses for each technique performed with a $4\Delta x$ Gaussian filter applied in the horizontal and a $2\Delta x$ in the vertical. Horizontal resolution is 1 km and vertical resolution is 0.5 km. Slight discrepancies in the winds at lower reflectivity values are found when comparing the horizontal cross sections at 2-km altitude in Figs. 6d and 6f. A weak inflow channel in the bottom left corner of Fig. 6f is present in the MLQC but not in the interactive editing. The horizontal flow fields are otherwise largely identical. Vertical cross sections (traced in cyan; Fig. 6d) passing through the highest reflectivity region of the convective feature show comparison of

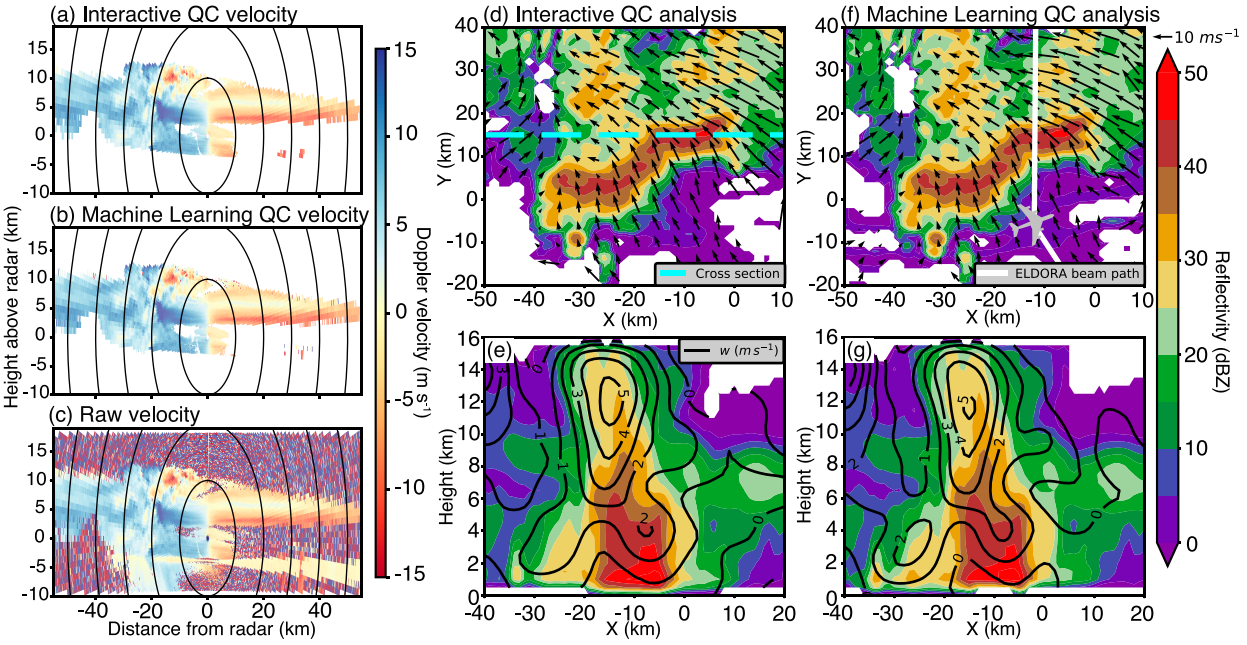

FIG. 6. A range–height indicator scan from the ELDORA radar plotted with Py-ART (Helmus and Collis 2016) showing the (a) interactive, (b) MLQC, and (c) raw velocity fields at 2127 UTC 6 Sep 2005. (d) A horizontal cross section at 2-km altitude of reflectivity (colors; dBZ) and analyzed winds (vectors) using interactively QCed data centered at 2123 UTC. (e) A vertical cross section taken from west to east through the most intense convection, showing the reflectivity (color shading) and vertical wind component (black contours, in 1 m s^{-1} intervals). (f),(g) The same analysis as in (d) and (e), but using the ML method for QC. The path of the radar beam from the scan in (a)–(c) is shown as a white line in (f), and location of the cross section is shown as a dashed cyan line in (d).

the calculated vertical motion (Figs. 6e,g). Both analyses display a similar shape and maximum value of the upward component of motion. There are slight differences in magnitude at lower levels to the left of the convective core. In general, consistencies in the low level planar flow field (Figs. 6d,f) and vertical motion (Figs. 6e,g) indicate the method is capable of producing an analysis that would be scientifically interpreted in a similar way to one produced by more time consuming interactive QC efforts.

4. Discussion

The proof-of-concept model demonstrates the ability of the random-forest ML technique to automate the time consuming interactive manual QC effort required to produce a research quality dual-Doppler analysis. Although the technique has not yet been optimized for speed, its present state offers a faster alternative for researchers to investigate observed convective weather phenomena. QC of a scan from the Ophelia test case proceeded as follows. The radar scan file is read in and the required predictors are calculated. Pointwise classification of weather and nonweather radar gates are made by the RF model. All radar gates that do not contain data or meet filtering thresholds in NCP and probability of ground gates are set to the nonweather class so that the full data arrays remain intact. All radar gates marked as nonweather via the generated binary prediction array are set to fill values in copies of the original fields that effectively removes the data from the scan. The new QCed fields are written into the original radar file with the nonweather data removed. The process in its entirety takes a few minutes per scan. An optimized rewrite of the predictor mining script and parallel processing of scans are steps that could greatly improve the processing speed of the MLQC technique in the future. Regardless of the current time requirements of this method, little user input is required. Automated QC can be performed in the background while a researcher completes other tasks rather than visually inspecting scans for nonweather data.

A key step missing from this QC process is the unfolding of velocity data. When observed Doppler velocities exceed the Nyquist velocity in either the positive or negative direction, the velocity folds over to the oppositely signed maximum value and reports the value incorrectly. Folding errors are corrected by adding or subtracting the correct number of Nyquist intervals ($2V_{max}$), where V_{max} is the maximum unambiguously detectable velocity (MUDV) for the radar (Houze et al. 1989). The dual-pulse repetition frequency capability of ELDORA allowed for the measurement of velocities of 60–80 m s^{-1} (Hildebrand et al. 1996; Bell et al. 2013), which are rarely exceeded in weather observations. Thanks to this capability, the ELDORA dataset does not usually contain velocity folding errors, but not all TDRs have MUDVs as high as ELDORA. Unfolding of velocity data is an automated process that works well in the absence of nonmeteorological clutter. An earlier version of this technique was employed to QC data collected by the TDR aboard the NOAA P-3 as it

observed Hurricane Michael (2018) during its rapid intensification (DesRosiers et al. 2022). Wind speeds in the storm were well in excess of the 25.6 m s^{-1} (Beven et al. 2019) of the TDR in the scanning mode used in 2018 (Gamache et al. 1997). Pairing the MLQC method, trained on interactive QC of this radar, with automated velocity unfolding allowed for the production of research quality dual-Doppler analyses (DesRosiers 2020). Success of this method despite the addition of velocity folding issues demonstrated the adaptability of this method to current radars. Future efforts to apply this method more broadly to the current generation of NOAA TDR radars and the future Airborne Phased Array Radar (APAR) that is in development (Vivekanandan et al. 2014) will require a large sample of interactively QCed radar data from multiple experts in a wider variety of convective phenomena to provide the confidence level of the ELDORA MLQC experiment described in this study.

The calculated model performance metrics (Fig. 4) should be evaluated with the caveat that the model is tasked with classifying more difficult weather versus nonweather radar gates. Subterranean data and data of lower quality was removed in the preprocessing step when NCP and probability of ground gate thresholds were applied. Nearly half of the data were removed from the ELDORA dataset, which were most likely exclusively nonweather data. Given the simplicity with which these data are removed, the model should be capable of doing so if trained with all data. However, limiting the radar data allows the model to focus on making the more difficult and important classifications that must be made in manual interactive QC.

5. Conclusions

A random-forest machine learning model has been developed to aid in the quality control of airborne radar through identification of nonmeteorological data. The method was trained on data collected in four separate field projects sampling different convective weather systems interactively QCed by different researchers. The ML model performed well on the withheld testing set data with 96% and 92.9% correct classification of nonweather and weather echoes, respectively, after easily rejected data were removed. The testing set results indicate good discrimination ability and a promising step toward reduction of effort required to perform dual-Doppler analyses. Tests on previously unseen data from Hurricane Ophelia (2005) collected during the RAINEX field campaign produced slightly lower classification accuracy but still correctly classified 94.5% of nonweather and 91.8% of weather echoes. Wind fields calculated from the ML and interactively QCed data are very similar, suggesting the method has practical applicability to produce the desired scientific end product with reduced time and effort for researchers. Simple hyperparameter tuning provided an effective number and depth of decision trees that allowed for good performance while limiting model overfitting. An earlier version of the technique was used with more recent P-3 TDR data with velocity folding issues present. The technique is capable of successful identification of weather and nonweather data even when velocity folding errors are present. A more exhaustive dataset is required to evaluate the generalizability of the QC technique in varied convective phenomena with folded velocity data present. The complex decision-making ability of the random forest to effectively combine predictors provides an advantage over previous automated approaches that rely primarily on thresholds of independent predictors. Replacing interactive QC methods with an automated ML technique should reduce the time and effort burdens on researchers who analyze airborne Doppler radar.

This study provides a proof of concept for the ability of an automated ML technique to recreate the interactive QC of ELDORA airborne radar data, paving the way toward a generalized method for other airborne radars with additional training and tuning. The method described herein offers adaptability to other radars due to its pointwise yet contextual classification that uses predictors available on current scanning tail Doppler radars. The algorithm infrastructure also easily allows for additions such as polarimetric radar variables as predictors when they become available for future radars. Continued effort should focus on increased performance and generalization ability as well as decreased computational time with the goal of meeting real-time data assimilation time windows for numerical model guidance. Assimilation of airborne radar observations has been shown to improve tropical cyclone guidance with a trade-off between higher-quality data improving intensity forecasts and greater data coverage decreasing track errors (Zhang et al. 2012). Continued improvements to the method capable of retaining high-quality data with good coverage offers an opportunity to improve weather forecasts by more effectively assimilating airborne radar data in the future.

Acknowledgments. This research was supported by National Science Foundation Awards OAC-1661663 and AGS-2103776, Office of Naval Research Award N000142012069, and NOAA APAR Risk Reduction Award NA19OAR4590245. We thank Bruno Melli for efforts that provided initial direction to the project and Wen-Chau Lee, Huaqing Cai, and Hannah Murphey for the original interactive QC of the dataset. We thank NCAR's Earth Observing Laboratory for the ELDORA data collection.

Data availability statement. The dataset utilized in this study is available on Figshare (https://doi.org/10.6084/m9.figshare.23689194).

REFERENCES

Bell, M. M., and M. T. Montgomery, 2010: Sheared deep vortical convection in pre-depression Hagupit during TCS08. *Geophys. Res. Lett.*, **37**, L06802, https://doi.org/10.1029/2009GL042313.

——, ——, and K. A. Emanuel, 2012: Air–sea enthalpy and momentum exchange at major hurricane wind speeds observed during CBLAST. *J. Atmos. Sci.*, **69**, 3197–3222, https://doi.org/10.1175/JAS-D-11-0276.1.

——, W.-C. Lee, C. A. Wolff, and H. Cai, 2013: A solo-based automated quality control algorithm for airborne tail Doppler radar data. *J. Appl. Meteor. Climatol.*, **52**, 2509–2528, https://doi.org/10.1175/JAMC-D-12-0283.1.

Beven, J. L. II, R. Berg, and A. Hagen, 2019: Tropical cyclone report: Hurricane Michael (AL142018), 7–11 October 2018. NHC Tech. Rep., 86 pp., https://www.nhc.noaa.gov/data/tcr/AL142018_Michael.pdf.

Bosart, B. L., W.-C. Lee, and R. M. Wakimoto, 2002: Procedures to improve the accuracy of airborne Doppler radar data. *J. Atmos. Oceanic Technol.*, **19**, 322–339, https://doi.org/10.1175/1520-0426-19.3.322.

Boukabara, S.-A., V. Krasnopolsky, J. Q. Steward, E. S. Maddy, N. Shahroudi, and R. N. Hoffman, 2019: Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges. *Bull. Amer. Meteor. Soc.*, **100**, ES473–ES491, https://doi.org/10.1175/BAMS-D-18-0324.1.

Cai, H., W.-C. Lee, M. M. Bell, C. A. Wolff, X. Tang, and F. Roux, 2018: A generalized navigation correction method for airborne Doppler radar data. *J. Atmos. Oceanic Technol.*, **35**, 1999–2017, https://doi.org/10.1175/JTECH-D-18-0028.1.

Claesen, M., and B. De Moor, 2015: Hyperparameter search in machine learning. arXiv, 1502.02127v2, https://doi.org/10.48550/arXiv.1502.02127.

DesRosiers, A. J., 2020: Airborne radar quality control and analysis of the rapid intensification of Hurricane Michael (2018). M.S. thesis, Dept. of Atmospheric Science, Colorado State University, 61 pp., https://mountainscholar.org/bitstreams/b3653b35-3197-495f-980a-da86bb196470/download.

——, M. M. Bell, and T.-Y. Cha, 2022: Vertical vortex development in Hurricane Michael (2018) during rapid intensification. *Mon. Wea. Rev.*, **150**, 99–114, https://doi.org/10.1175/MWR-D-21-0098.1.

Fischer, M. S., P. D. Reasor, R. F. Rogers, and J. F. Gamache, 2022: An analysis of tropical cyclone vortex and convective characteristics in relation to storm intensity using a novel airborne Doppler radar database. *Mon. Wea. Rev.*, **150**, 2255–2278, https://doi.org/10.1175/MWR-D-21-0223.1.

Foerster, A. M., M. M. Bell, P. A. Harr, and S. C. Jones, 2014: Observations of the eyewall structure of Typhoon Sinlaku (2008) during the transformation stage of extratropical transition. *Mon. Wea. Rev.*, **142**, 3372–3392, https://doi.org/10.1175/MWR-D-13-00313.1.

Gamache, J. F., J. S. Griffin, P. P. Dodge, and N. F. Griffin, 1997: Evaluation of a fully three-dimensional variational Doppler analysis technique. Preprints, *28th Conf. on Radar Meteorology*, Austin, TX, Amer. Meter. Soc., 422–423.

——, P. P. Dodge, and N. F. Griffin, 2008: Automatic quality control and analysis of airborne Doppler data: Realtime applications, and automatically post-processed analyses for research. Preprints, *28th Conf. on Hurricanes and Tropical Meteorology*, Orlando, FL, Amer. Meteor. Soc., P2B, https://ams.confex.com/ams/pdfpapers/137969.pdf.

Helmus, J. J., and S. M., Collis, 2016: The Python ARM radar toolkit (Py-ART), a library for working with weather radar data in the Python programming language. *J. Open Res. Software*, **4**, e25, https://doi.org/10.5334/jors.119.

Hence, D. A., and R. A. Houze Jr., 2008: Kinematic structure of convective-scale elements in the rainbands of Hurricanes Katrina and Rita (2005). *J. Geophys. Res.*, **113**, D15108, https://doi.org/10.1029/2007JD009429.

Hildebrand, P. H., and Coauthors, 1996: The ELDORA/ASTRAIA airborne Doppler weather radar: High-resolution observations from TOGA COARE. *Bull. Amer. Meteor. Soc.*, **77**, 213–232, https://doi.org/10.1175/1520-0477(1996)077<0213:TEADWR>2.0.CO;2.

Houze, R. A., Jr., S. A. Rutledge, M. I. Biggerstaff, and B. F. Smull, 1989: Interpretation of Doppler weather radar displays of midlatitude mesoscale convective systems. *Bull. Amer. Meteor. Soc.*, **70**, 608–619, https://doi.org/10.1175/1520-0477(1989)070<0608:IODWRD>2.0.CO;2.

——, W.-C. Lee, and M. M. Bell, 2009: Convective contribution to the genesis of Hurricane Ophelia (2005). *Mon. Wea. Rev.*, **137**, 2778–2800, https://doi.org/10.1175/2009MWR2727.1.

Lakshmanan, V., C. Karstens, J. Krause, and L. Tang, 2014: Quality control of weather radar data using polarimetric variables. *J. Atmos. Oceanic Technol.*, **31**, 1234–1249, https://doi.org/10.1175/JTECH-D-13-00073.1.

Lee, W.-C., F. D. Marks, and C. Walther, 2003: Airborne Doppler radar data analysis workshop. *Bull. Amer. Meteor. Soc.*, **84**, 1063–1075, https://doi.org/10.1175/BAMS-84-8-1063.

Louppe, G., 2014: Understanding random forests: From theory to practice. arXiv, 1407.7502v3, https://doi.org/10.48550/arXiv.1407.7502.

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.

Ośródka, K., and J. Szturc, 2022: Improvement in algorithms for quality control of weather radar data (RADVOL-QC system). *Atmos. Meas. Tech.*, **15**, 261–277, https://doi.org/10.5194/amt-15-261-2022.

Oye, R., C. Mueller, and S. Smith, 1995: Software for radar translation, visualization, editing, and interpolation. Preprints, *27th Conf. on Radar Meteorology*, Vail, CO, Amer. Meteor. Soc., 359–361.

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

Stechman, D. M., G. M. McFarquhar, R. M. Rauber, M. M. Bell, B. F. Jewett, and J. Martinez, 2020: Spatiotemporal evolution of the microphysical and thermodynamic characteristics of the 20 June 2015 PECAN MCS. *Mon. Wea. Rev.*, **148**, 1363–1388, https://doi.org/10.1175/MWR-D-19-0293.1.

Tang, L., J. Zhang, M. Simpson, A. Arthur, H. Grams, Y. Wang, and C. Langston, 2020: Updates on the radar data quality control in the MRMS quantitative precipitation estimation system. *J. Atmos. Oceanic Technol.*, **37**, 1521–1537, https://doi.org/10.1175/JTECH-D-19-0165.1.

Testud, J., P. H. Hildebrand, and W.-C. Lee, 1995: A procedure to correct airborne Doppler radar data for navigation errors using the echo returned from the earth's surface. *J. Atmos. Oceanic Technol.*, **12**, 800–820, https://doi.org/10.1175/1520-0426(1995)012<0800:APTCAD>2.0.CO;2.

Vivekanandan, J., W.-C. Lee, E. Loew, J. L. Salazar, V. Grubišić, J. Moore, and P. Tsai, 2014: The next generation airborne polarimetric Doppler weather radar. *Geosci. Instrum. Methods Data Syst.*, **3**, 111–126, https://doi.org/10.5194/gi-3-111-2014.

Wakimoto, R. M., and C. Liu, 1998: The Garden City, Kansas, storm during VORTEX 95. Part II: The wall cloud and tornado. *Mon. Wea. Rev.*, **126**, 393–408, https://doi.org/10.1175/1520-0493(1998)126<0393:TGCKSD>2.0.CO;2.

——, H. V. Murthey, A. Nester, D. P. Jorgensen, and N. T. Atkins, 2006: High winds generated by bow echoes. Part I: Overview of the Omaha bow echo 5 July 2003 storm during BAMEX. *Mon. Wea. Rev.*, **134**, 2793–2812, https://doi.org/10.1175/MWR3215.1.

Zhang, L., Z. Pu, W.-C. Lee, and Q. Zhao, 2012: The influence of airborne Doppler radar data quality on numerical simulations of a tropical cyclone. *Wea. Forecasting*, **27**, 231–239, https://doi.org/10.1175/WAF-D-11-00028.1.