Atmospheric Humidity Estimation From Wind Profiler Radar Using a Cascaded Machine Learning Approach

Anas Amaireh, Yan Zhang, *Senior Member, IEEE*, and P. W. Chan

***Abstract*—A method for estimating atmospheric relative humidity using wind profiler radar and a "cascaded" machine learning algorithm is introduced. Unlike existing methods in the literature, the proposed approach uses only I/Q or moment data from the profiler radar to generate an intermediate pressure profile, which serves as training data for humidity estimations without requiring temperature as an input feature. The study examines the potential of various machine learning algorithms and evaluates their performance using field data collected by the Hong Kong Observatory between January and June 2021. Importantly, this is the first time a cascading machine-learning solution has been successfully applied to the humidity estimation problem, resulting in a simplified model with reduced complexity and fewer required features.**

***Index Terms*—Decision tree, ensemble tree, machine learning (ML), neural network (NN), profiler radar, relative humidity (RH).**

I. Introduction

ESTIMATING the atmospheric humidity profile is essential for various fields, including weather prediction, climate studies, aviation safety, agriculture, hydrology, and environmental monitoring [1], [2], [3], [4], [5], [6], [7], [8]. Accurate estimates of the humidity profile can help improve decision-making and planning in these fields and ultimately lead to a better understanding of the earth's atmosphere and its role in the global climate system [9], [10]. The challenge of retrieving low-medium (up to 10 km) atmosphere humidity using Doppler wind profiler radar has been discussed in numerous previous studies [11], [12], [13], [14], [15], [16], [17], [18], [19], [20].

Most of the existing methods depend on the physical models, which are based on a set of assumptions about the behavior of the atmosphere, and they rely on precise input data to make accurate predictions. On the other hand, the physical models are limited in their accuracy and adaptation capability due to the interactions of uncertainties in the physical parameters, the limited data availability, and the model structure limitations [21]. For example, these models require input values for physical parameters such as temperature, pressure, and wind speed; but the measurement of these parameters is often uncertain, leading to small input faults and significant model prediction errors [22]. Additionally, the models require large amounts of data and can be biased by data gaps and errors, which may not accurately reflect the behavior of the real-world system. These limitations could result in physical models providing unreliable or inaccurate predictions of atmospheric humidity profiles. In such cases, alternative methods, such as machine learning (ML) algorithms, may be desirable to offer humidity profile estimations based on less dependence on physical modeling.

ML algorithms capable of constructing complex, nonlinear relationships between input variables from vast datasets had diverse applications across energy, environmental engineering, and atmospheric predictions [23], [24], [25], [26], [27], [28], [29], [30]. The unique characteristics of ML algorithms are the flexibility and reducing necessity for accurate physical parameters [24]. Therefore, ML algorithms supplement traditional physical models through their adaptability to unknown and dynamic environments. In the recent developments, critical parameters like interfacial tension and viscosity have been modeled using techniques such as multilayer perceptron (MLP) optimized with Levenberg–Marquardt (LMA) and gradient boosting decision tree (GBDT) [25], [26], [29], [30].

Exploring ML methods for atmospheric parameter prediction is an active field of study. For example, artificial neural networks (ANNs) have been utilized for refractivity prediction in several studies [31], while these studies do not address the specific needs of profiler radar sensing. Other studies have demonstrated the application of ML to weather forecasting using various types of meteorological data. For instance, linear and functional regression models have been employed to forecast maximum and minimum temperatures for seven days using two days' worth of weather data [32]. Similarly, an hourly rainfall forecast model based on a support vector machine (SVM) was developed to predict rainfall with high temporal resolution and accuracy [33]. For satellite remote sensing, ML has been integrated to improve data interpretation. A standalone cloud detection algorithm was designed for the Microwave Humidity Sounder-2 (MWHS-2) satellite sensors, utilizing a GBDT. This algorithm was

Manuscript received 3 April 2023; revised 4 June 2023; accepted 20 June 2023. Date of publication 5 July 2023; date of current version 19 July 2023. The work of P. W. Chan was supported by the Hong Kong Observatory. *(Corresponding author: Anas Amaireh.)*

Anas Amaireh and Yan Zhang are with the School of Electrical and Computer Engineering and Advanced Radar Research Center, University of Oklahoma, Norman, OK 73019 USA (e-mail: anas@ou.edu; rockee@ou.edu).

P. W. Chan is with the Hong Kong Observatory, Kowloon, Hong Kong (e-mail: pwchan@hko.gov.hk).

Digital Object Identifier 10.1109/JSTARS.2023.3292351

trained on observations from China's new-generation weather radar [34]. Looking at the incorporation of Global Navigation Satellite System-Radio Occultation data, ML has been leveraged to forecast wind fields in the Beijing–Tianjin–Hebei region of China [35]. Initial processing established relationships between thermodynamic and kinetic parameters through historical monitoring data. This step was followed by predictions using ML models, such as long short-term memory (LSTM), convolutional neural networks (CNNs), and deep neural networks (DNNs). ML has also been applied to estimate relative humidity (RH) with specific resolutions. A study used random forest and XGBoost to estimate daily near-surface RH at a 1-km resolution over Japan and South Korea, with separate schemes for clear and cloudy sky conditions [36], has been reported.

In this study, we explore a potentially novel approach that primarily uses ML for improved estimation of humidity. First, we separate the clutter return from useful atmosphere echoes through spectrum processing to obtain the basic wind estimation and clutter spectrum properties. Next, we feed the profiler estimation outputs to generate a set of "features." Finally, those higher level features and part of the low-level moment estimations from spectrum processing are combined as training features for ML. The *bagged learning tree* obtained from tuning and comparison with other algorithms is applied as the optimal solution of ML algorithm. It is proven the most effective and computationally efficient algorithm based on the evaluations. The novel contribution includes: 1) The radar profiler moment data is sufficient to obtain reasonably accurate humidity estimations. 2) This is the first time that a "cascading ML" solution, which produces an intermediate training stage with "synthetic" pressure data, is successfully applied to the humidity estimation problem. 3) The RH is estimated with no need to use the temperature as an intermediate input feature.

This article is organized as follows: Section II summarizes the instrumentation, data collection, and existing methods based on physical models. In Section III, the proposed approach is described in detail. Climatology details for the first half of 2021 in Hong Kong are presented in Section IV. Section V offers a comprehensive methodology for the study, including sections on ML algorithms, performance evaluation, data cleaning and preprocessing, and a dataset description. Validation results based on the available datasets are presented in Sections VI, while Section VII provides conclusions and suggestions for future work.

II. Instrumentation and Physical Modeling Methods

HKO's wind profilers operate at 1299 MHz and locate at multiple locations in Hong Kong. Two of the three profilers are installed at the airport, while the third is installed at the city center. A profile has three beams, one vertical and two 50 °C tilted from the vertical. These three beams can support retrieval of the 3-D vector wind (u, v, w). Raw I/Q data can be collected, and it is useful for estimating the vector wind profiles. On the other hand, spectral moment data directly from the profiler outputs can be used to study atmospheric turbulence parameters, which are related to the RH profiles. For the radiosonde truth data, HKO uses Vaisala RS41-SG radiosondes to perform daily routine soundings for both automatic (AUTOSONDE) and manual launches. These radiosondes are used to measure RH, temperature, and pressure with 2% to 4% range of accuracies. The sounding data are collected every 6 min. In addition to information on RH, temperature, and pressure, it also provides information such as wind speed, wind direction, and dew points. The data covered up to an altitude of around 10 km. The HKO used the measurement data collected by the city center's profiler for training and testing. The combination of the wind profiler data and the radiosonde data provide a comprehensive view of the atmospheric conditions, which improves weather parameters forecasting.

The classical method of estimating the humidity profile is based on the following (1) and (2). In these equations, q_0 is the humidity at level z_0. T represents the temperature, P represents the atmospheric pressure, and θ is $\theta = (1000/P)^{\frac{2}{7}}$. M is the vertical (z-direction) gradient of the atmosphere refractivity. C_n^2 is the turbulence structure parameter, which can be estimated from the zeroth radar moment (or total power) of radar return signals. α^2 is a scalar constant dependent on specific regions. $\frac{dV}{dz}$ is the vertical shear of the horizontal wind vector, which can be estimated from the second radar moment (wind velocity estimations). Turbulent kinetic energy ϵ is directly related to the third radar moment or spectrum width.

Although (1) and (2) are used as the initial guidance to RH retrieval algorithms, they reveal the importance of accurate estimations of the spectrum moments. They also inspire the application of ML algorithms in the way that temperature, pressure, or radar spectrum moments might be directly used as feature vectors for humidity retrieval. However, the precise estimation of radar moments is affected by many factors, such as clutter, noise, and equipment quality. Even with preprocessing as a way to enhance the estimation performance, using the physical models still have many challenges. Our approach as follows will then focus on ML solution that is "inspired" by the physical model parameters

$$q(z) = \theta^2 \int_{z_0}^{z} \left(1.67 \times 10^{-6} \frac{MT^2}{P} + \frac{1}{7750} \frac{d\theta}{dz}\right) \theta^{-2} dz + q_0 \tag{1}$$

$$C_n^2 = \alpha^2 \frac{\epsilon^{2/3} M^2}{\left(\frac{dV}{dz}\right)^2}. \tag{2}$$

III. Approach and Methodology

A. Processing Flow

The overall technical solution is described in Fig. 1. Typically, the first step is using classic Doppler radar signal processing to obtain an accurate estimation of the Doppler spectrum and wind estimation for different altitudes (up to 10 km). Spectral data from the three beams of the wind profiler have been used to estimate three moments, power (first moment), velocity (second moment), and spectrum width (third moment). Next, we separate the clutter return from useful atmosphere echoes through spectrum processing and obtain both the basic wind estimation

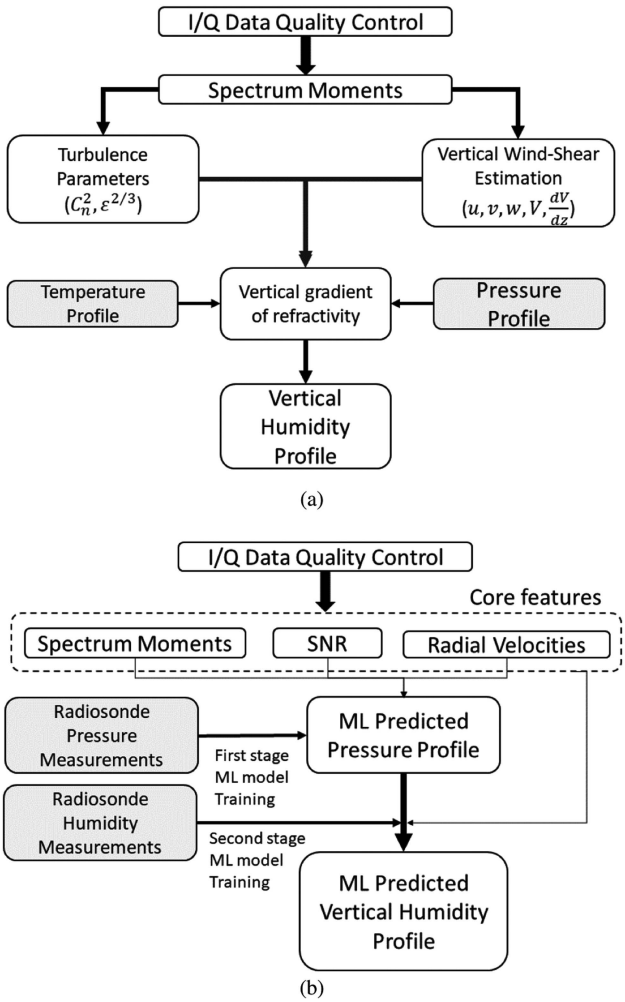


Fig. 1. Comparison of traditional physical model-based method (a) and the cascaded ML method (b) for humidity profile estimation.

and clutter spectrum properties. Next, the profiler estimation outputs and other in-situ probe information (such as pressure) are combined to generate a set of training features for ML.

For a complex physical model in (1) and (2), the normal way of processing, which was adopted in existing literature, is to estimate the wind vectors first and then calculate turbulence parameters, an M, as in Fig. 1. However, the lack of access to vendor-specific and proprietary information on the profiler processing algorithms has been an obstacle to obtain the validated wind vector estimations. Therefore, we explored a new approach that moves the "entry-point" of ML closer to the raw spectrum data. The proposed idea is that instead of performing precise wind estimation, we use the moment data from the raw spectrum directly as ML feature variables. Here the moments are calculated from the three beams of the wind profiler as feature variables for ML models to predict RH directly. So, instead of relying on precise wind estimation and temperature and pressure parameters from radiosonde data, which is not always available, the moment data from the raw spectrum are used to predict the RH in two-cascaded steps. In the first step, the moment data are used as input features to predict the pressure. In the second step, the predicted pressure and moment data are combined to predict the RH without using temperature. In the cascaded ML solution, the temperature, which is typically regarded as a crucial element in predicting humidity according to the literature, is no longer necessary. This simplifies the ML model and reduces the computational loads for the RH estimation.

The level-II radar profiler moments are acquired as the inputs. To ensure the integrity of the processing pipeline, the processing procedure includes a range of data cleaning steps, such as outlier removal and smoothing, which are detailed in the subsequent sections. The feature engineering phase was exploratory and iterative. It involved exploring potential features and assessing their correlation with the target variables of pressure and humidity. A set of features was selected based on their optimal correlation with the target variables and their wide variance. An innovative aspect of the study is the creation and application of the "synthetic" pressure data. Following a rigorous data-cleaning process, the algorithm begins with selecting suitable moment data and other features for pressure prediction. Next, various ML algorithms are trained to achieve an optimal pressure estimation. This optimal pressure estimation is an additional input feature for humidity estimation. The moment data and the predicted pressure are then trimmed again to ensure optimal correlation with RH and the highest variance. These high-quality input features are then used in training different ML algorithms to achieve the most accurate possible humidity prediction.

B. General Climatology of Hong Kong

Understanding the climatology trend of the region will help us better understand data and algorithm verification. The first half of 2021, from January to June, was unusually warm, primarily due to the four months' worth of temperatures that were much above average. The average maximum, the average minimum, and the entire time average temperature were 26.3, 23.3, and 21.3 °C, respectively. For January 2021, the early half experienced cooler temperatures than the second half, which was comparatively warmer. The mean temperature for the month was 16.2 °C, 0.3 °C cooler than the average of 16.5 °C. It was drier and sunnier in January 2021 than typical. Total sunlight hours for the month were 217.3, which is 49% more than the average of 145.8 h. In the month, there was very little rain. February 2021 was significantly sunnier and warmer in Hong Kong than usual due to the northeast monsoon across southern China being less than typical for the majority of the month. The average maximum and minimum temperatures were 23.5 and 17.5 °C, respectively, which are 4.1 and 2.2 °C higher than the corresponding normal. More than double the average of 101.7 h, the total number of hours of bright sunlight in February was 205.1. As a result, the winter in Hong Kong during December 2020, January 2021, and February 2021 was hotter than typical, mainly due to the unusually sunny and warm weather. March 2021 in Hong Kong continued to be unusually warm despite fewer outbreaks of cold air from the north. The monthly average minimum and maximum temperatures were 22.0 and 24.8 °C, respectively. These three temperatures were 2.5 and 2.9 °C, above their respective normals, and were the

TABLE I
SUMMARY OF THE CLIMATE IN HONG KONG DURING THE FIRST HALF OF 2021 [37]

Meteorological Element	January	February	March	April	May	June
Mean Relative Humidity	62 %	75 %	79 %	79 %	78 %	82 %
Mean Cloud Amount	47 %	41 %	69 %	71 %	75 %	83 %
Total Rainfall	Trace	62.1 mm	3.5 mm	32.5 mm	65.0 mm	628.0 mm
Total Evaporation	87.9 mm	84.2 mm	87.3 mm	95.8 mm	141.0 mm	99.6 mm

maximum monthly average values for March on record. The total rainfall for the month was only 3.5 mm, or around 5% of the average amount of 75.3 mm, making it significantly drier than typical. April 2021 stayed substantially warmer than usual. The average monthly minimum temperature was 22.4 °C, and the average monthly maximum temperature was 27.0 °C. These values were 1.3 and 1.4 °C higher than normal. With a total rainfall of just 32.5 mm, or around 21% of the average of 153.0 mm, April 2021 was also significantly drier than usual due to the dominance of upper-air anticyclones across southern China for the majority of the month. May 2021 was the warmest May in Hong Kong, primarily due to the stronger-than-normal subtropical ridge across southern China. The monthly average low temperature of 27.0 °C was recorded for May and was 2.5 °C beyond its respective normal. The average high temperature, 32.1 °C, was 3.3 °C higher than average. Hong Kong saw the hottest spring in history from March to May 2021, along with unusually warm temperatures in March and April 2021. With rainfall data of only 65.0 mm, May was also significantly drier than typical. The local weather was unpredictable on the first day of June 2021 due to a low-pressure trough with heavy rains and violent thunderstorms. This month had rainfall totals of 628.0 mm or roughly 28% more than the average of 491.5 mm. The heavy rains helped Hong Kong recover from the extreme dryness of the previous several months. With a mean temperature of 28.8 °C, which is 0.5 °C higher than the average of 28.3 °C, June was also hotter than typical. In June 2021, three tropical cyclones passed across the South China Sea, and the western North Pacific [37]. Table I presents the meteorological values for January through June, including the mean RH, mean cloud amount, total rainfall, and total evaporation. The mean RH starts at 62% in January and gradually increases to 82% in June. On the other hand, the mean cloud values vary throughout the months, starting at 47% in January and reaching the highest value of 83% in June. The total rainfall in January is recorded as a trace, while it is 62.1 mm in February. It reached its highest value in June, with a total of 628.0 mm. The total evaporation is similar to the other elements, in which the lowest value was in January at 87.9 mm, and the highest was in May at 141.0 mm. Fig. 2 shows the pressure and humidity values for the study region during the first half of 2021 at different altitudes (from 0 to around 10 km).

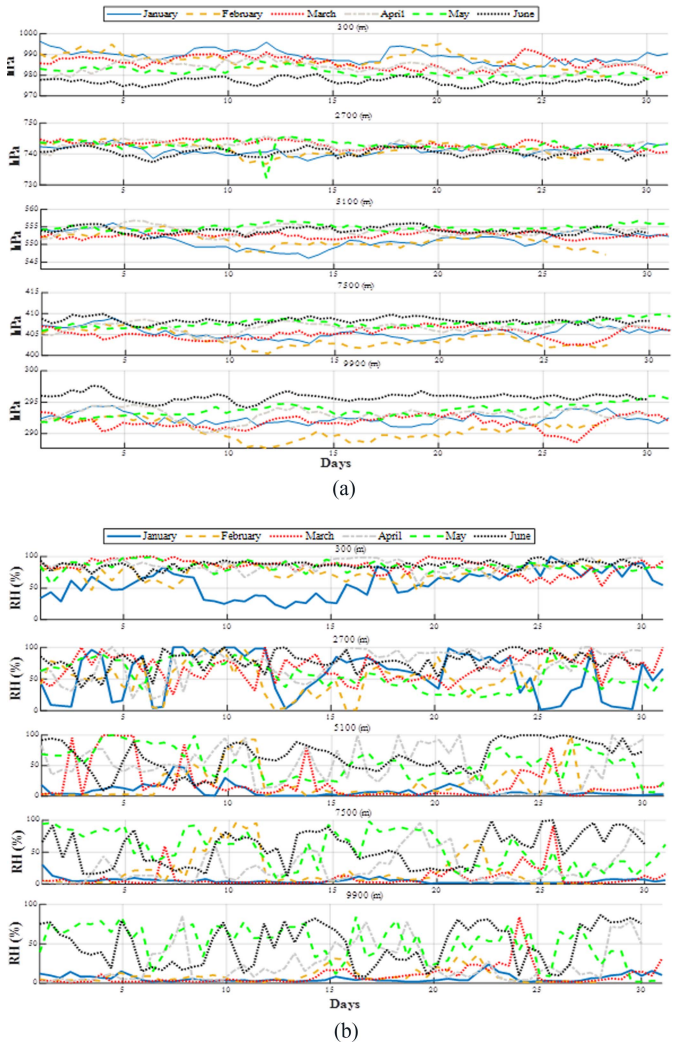


Fig. 2. Comparison (a) pressure values, (b) RH values of the region with different times (month) and different altitudes.

C. ML Algorithms and Methods Used in This Study

1) Regression Decision Tree: A decision tree is a multivariate method that takes into account events that may miss a certain condition. Instead of immediately dismissing such events, the decision tree evaluates whether other conditions could help categorize them correctly. In theory, a decision tree may handle numerous output classes, with every branch breaking into many subbranches [38]. The regression tree algorithm typically consists of several steps: setting the accuracy of the prediction criterion, picking splits, deciding when to finish splitting, and determining the ideal tree [39]. For example, the criterion for accuracy in the first stage might be cross-validation, resubstitution error, or test sample error. As indicated in (3), the resubstitution error is determined as the mean squared error of the same data used to generate the prediction p

$$E(p) = \frac{1}{N}\sum_{i=1}^{N}\left(u_i - p\left(v_i\right)\right)^2 \tag{3}$$

where N is the number of samples, u_i and v_i represent the learning samples. To compute cross-validation error, the samples are divided into k smaller samples of nearly equal sizes. The small sample is utilized to build the predictor p. The cross-validation

error is then calculated based on this small sample, as indicated in the following equation:

$$E^{CV}(p) = \frac{1}{N_k} \sum_{k} \sum_{(u_i, v_i) \in X_k} \left(v_i - p^{(k)}(u_i) \right)^2 \quad (4)$$

$p^{(k)}$ is calculated from the small sample X_k. The test sample error splits the entire number of instances into two subsamples; X_1 with size N_1, and X_2 with size N_2. As indicated in (5), the error of the test sample is determined

$$E^{ts}(p) = \frac{1}{N_2} \sum_{(u_i, v_i) \in X_k}^{N} (v_i - p(u_i))^2 \quad (5)$$

where X_k is the small sample that is not utilized to build the predictor. The regression tree technique then determines splits to estimate continuous variable values. Splits are often quantified using a node impurity measure, which indicates the relative homogeneity in terminal nodes. As illustrated in (6), the least-squared deviation is employed as a metric in a regression tree for node impurity

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i (u_i - \bar{v}(t))^2 \quad (6)$$

where $N_w(t)$ shows instances' weighted number in node t. w_i presents the amount of the weighting variable for i. f_i, u_i, and $\bar{v}(t)$ show the frequency variable, the response variable, and the node t weighted mean, respectively.

The third stage is to decide when to cease splitting, which is determined by the number of nodes required. The next stage is to choose the tree with the right size, often known as the optimum tree, typically attained by tree pruning. The tree pruning approach utilized in this study is the smallest sized tree with the least amount of error.

This work estimates pressure and RH using regression ML techniques. These regression-based algorithms make predictions based on one or more categorical or continuous predictor features [40]. The selection of the DTR method for forecasting pressure and humidity is due to several advantages over other methods. The DTR method provides less training time than the other algorithms, shows a higher accuracy with large datasets, and is more resilient to changes in data patterns and shifts in distribution. Also, it uses if–then conditions to determine optimal value predictions. These factors make the decision tree method more efficient in resource utilization and a reliable and robust choice for predicting values in various environments.

2) Bagged and Boosted Regression Trees: Bagging is a method that minimizes prediction variance and hence enhances prediction accuracy. Its basic concept is that numerous bootstrap samples are taken from the available data, a prediction algorithm is applied to each bootstrap sample, and the results are aggregated. In the case of regression, the findings are averaged to give the overall forecast, with the variance minimized due to this averaging [41]. So, the bagged trees are based on the observation that a portion of the error in a given regression tree is related to the unique choice of the training data. By resampling with replacement and growing regression trees without averaging and pruning them, the output's variance component is decreased [41]. Boosting decreases variance and bias in supervised learning and turns weak learners into powerful ones with a high correlation with the real classification [42]. Most boosting techniques involve iterative learning weak classifiers, which are combined to form a final strong classifier. Once they are included, they are evaluated in a way that would be relevant to the accuracy of the poor learners. The data weighting is reevaluated once a poor learner is provided. Misclassified input data gains weight, whereas correctly classified instances lose weight. As a result, new poor learners focus more on cases misclassified by prior weak learners [43]. Boosting involves fitting models to training data iteratively, gradually increasing focus on observations inadequately described by the current group of trees. The methods through which boosting algorithms measure the lack of fit and pick settings for the following iteration differ [43]. Furthermore, boosting is a method for reducing the loss function that involves adding a new tree at each step that best decreases the loss function. The initial regression tree in BRT is the one that minimizes the loss function the most. The focus of the subsequent steps is on the residuals: variance in the response that is not yet described by the model [44].

3) Artificial Neural Networks (ANNs): In this study, different ANN configurations are tested, such as wide ANN (large layer size), medium ANN (medium layer size), and narrow ANN (small layer size), and ANN with two layers (BiNN) or three layers (TriNN). We use the Rectified Linear Unit (ReLU) function as the activation function for different types of ANNs. This function, which controls neuron activation, has proven to be more effective in backpropagation and gradient descent than other activation functions, and it can prevent potential issues of gradient vanishing and exploding. A systematic manual tuning process was carried out to identify the most suitable hyperparameters for the neural network (NN) models. This process entailed examining various combinations of hyperparameters, including different learning rates, numbers of hidden layers and neurons, activation functions, and regularization techniques. The model's performance was evaluated after implementing these various combinations, and then the mean squared error was used to assess the effectiveness of each combination. The guiding principle for the selection was the set of hyperparameters that delivered the most robust performance on a validation dataset. The manual tuning process employed in this study yielded considerable benefits. First, its simplicity helped deepen the understanding of how various hyperparameters influenced the model's performance. Moreover, it demonstrated its efficiency and practicality in selecting appropriate hyperparameters.

4) Selection of the ML Solutions: The selection of ML solutions included model complexity, ease of interpretation, computational efficiency, and prediction accuracy. The bagged ensemble tree (BET), coarse tree, and wide neural network (WNN) emerged as the primary algorithms for this task. However, several alternative methods were also examined and compared to validate these choices and gain a broader perspective. The bagged ensemble tree emerged as a particularly effective tool in this study. Its strength lies in robustly managing high-dimensional data and its natural ability to resist overfitting.

TABLE II
DETAILS OF THE HYPERPARAMETERS OF THE PROPOSED ML MODELS

Model Type	Hyperparameters
Interactions Linear Regression (ILR)	Terms: Interactions; Robust option: Off
Fine Tree (FT)	Minimum leaf size= 4
Medium Tree (MT)	Minimum leaf size= 12
Coarse Tree (CT)	Minimum leaf size= 36
Boosted Ensemble Tree (BoostedET)	Minimum leaf size= 8; Number of learners: 30; Learning rate: 0.1
Bagged Ensemble Tree (BET)	Minimum leaf size: 8; Number of learners: 30
Narrow Neural Network (NNN)	1 layer; layer size= 10; Activation: ReLU; Iteration limit: 1000
Medium Neural Network (MNN)	1 layer; layer size= 25; Activation: ReLU; Iteration limit: 1000
Wide Neural Network (WNN)	1 layer; layer size= 100; Activation: ReLU; Iteration limit: 1000
Bilayered Neural Network (BiNN)	2 layers; layers' size= 10; Activation: ReLU; Iteration limit: 1000
Trilayered Neural Network (TriNN)	3 layers; layers' size= 10; Activation: ReLU; Iteration limit: 1000

As an ensemble method, the BET combines the predictions of multiple decision trees, thereby reducing variance and enhancing model accuracy. Plus, it yields crucial insights into the factors influencing humidity prediction via feature importance scores. However, it is worth noting that bagged trees can demand substantial computational resources, especially when dealing with huge datasets. In contrast, the coarse tree, known for its simplicity and interpretability, facilitates easier visualization and understanding of data relationships. This straightforwardness reduces the risk of overfitting, which is crucial given the intricate nature of atmospheric data. However, this simplicity might also restrict its ability to understand complex, nonlinear relationships between features, potentially compromising accuracy. The wide NN model was chosen for its proficiency in modeling nonlinear relationships. With a single layer and numerous neurons, the WNN successfully grasps the complicated interplay between atmospheric parameters. However, it might struggle to effectively capture more intricate hierarchical representations that deeper networks handle better. Also, careful tuning of neurons is needed to prevent overfitting.

The models and their chosen hyperparameters are shown in Table II.

D. Method for Performance Evaluation

The root-mean-square-error (RMSE), mean square error (MSE), correlation coefficient (ρ), mean absolute error (MAE), and the coefficient of determination (R-Squared or R^2) metrics are used to examine the estimation errors of pressure and RH, in order to assess the performance of the different algorithms. The RMSE is calculated by the following equation:

$$\mathrm{RMSE} = \sqrt{\left[\frac{\left(\sum_{i=1}^{N}\left(X_{\mathrm{true}}(i) - X_{\mathrm{predicted}}(i)\right)^2\right)/}{N}\right]} \tag{7}$$

where N is the total number of samples, X_{true} and $X_{\mathrm{predicted}}$ are the true and estimated (or predicted) values, respectively.

The MSE is the average of the squared errors between the true and estimated (or predicted) values. The equation of the mean squared error is as follows:

$$\mathrm{MSE} = \frac{1}{2N}\sum_{i=1}^{N}\left(X_{\mathrm{true}}\,(i) - X_{\mathrm{predicted}}\,(i)\right)^2. \tag{8}$$

The correlation coefficient has values range from -1 to 1. The following equation depicts the correlation coefficient (ρ) between the true and estimated values:

$$\rho\left(X_{\mathrm{true}}\,,\ X_{\mathrm{predicted}}\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{X_{\mathrm{true}}\,(i) - \mu_{\mathrm{true}}}{\sigma_{\mathrm{true}}}\right)\left(\frac{X_{\mathrm{predicted}}\,(i) - \mu_{\mathrm{predicted}}}{\sigma_{\mathrm{predicted}}}\right) \tag{9}$$

where μ and σ are the mean and standard deviation, respectively.

The fourth evaluation metric in this study is the MAE, which calculates the errors' average magnitude in a group of predictions without considering their direction. MAE is the mean of the absolute differences between actual and forecast observation over the test samples, with equal weights for all individual differences [45]:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|X_{\mathrm{true}}\,(i) - X_{\mathrm{predicted}}\,(i)\right|. \tag{10}$$

Finally, the coefficient of determination (also known as R-Squared or R^2) is a statistical metric used to evaluate the performance of a regression model. It measures the amount of variation in the dependent variable that can be explained by the independent variable. The R-squared value ranges from 0 to 1. A higher R-squared value indicates that the model explains more of the variability in the dependent variable and that there are fewer discrepancies between the observed and fitted data [46]. R-squared is typically calculated by comparing the total sum of squares to the residual sum of squares, where $\bar{X_{\mathrm{true}}}$ is the mean of the truth values

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(X_{\mathrm{true}}\,(i) - X_{\mathrm{predicted}}\,(i)\right)^2}{\sum_{i=1}^{N}\left(X_{\mathrm{true}}\,(i) - \bar{X_{\mathrm{true}}}\right)^2}. \tag{11}$$

E. Data Quality Control and Prepossessing

The outliers and missing numbers could impact the ML algorithm's reliability and validity, leading to inaccurate results. Detecting and handling missing variables and outliers to maintain the quality and validity of study results can be accomplished using various strategies, such as imputing missing values, eliminating outliers, or changing the data to a more appropriate scale [47]. Missing data can either be discarded

during preprocessing or substituted with values calculated using statistical algorithms. In this study, all missing values have been removed to ensure accurate and reliable results [48]. Outliers must either be modified once their sources have been found or replaced with replacement values. In this work, all the outliers have been removed (similar to [47]). After removing the outliers and missing data from the complete dataset, the data were further smoothed to enhance the correlation between the input features and the targets [49]. The moving median smoothing method [50], also known as the linear Gaussian filter, was used in this study. It is remarkably robust against uncommon events, such as sudden shocks, which can be well handled by the Laplace distribution [51]. Furthermore, this method can enhance the correlation between the input features and targets.

F. Description of the Dataset

A six-month dataset encompasses both Doppler spectral moments and radiosonde data taken at 49 different heights, from 300 to 9900 m, with an interval of 200 m, were used for the verification. In the Doppler spectral data, one sample was recorded every 1 min and 42 s, while in the radiosonde data, samples were collected at 12 PM and 12 AM every day. Each set of radiosonde data consisted of samples taken every 2 s until the 47th minute. The data were adequately assigned to match the corresponding height, date, and time, ensuring accuracy and reliability in the results. The dataset collected over the course of 6 months offers a comprehensive understanding of the pressure and RH patterns. To effectively utilize this information, we employed a cross-validation approach with $K = 5$ to evaluate the performance of our proposed ML algorithms. The cross-validation dataset consisted of 85% of the 1 406 688 unique samples, while the remaining 15% was used as the testing dataset to validate the accuracy of the algorithms. The complete 6 month data were divided into a cross-validation dataset with 1 195 688 samples and a testing dataset with 211 000 samples.

G. Prevention of Overfitting

To prevent overfitting, a strategy was adopted to apply the K-fold cross-validation technique with K set to 5. This technique divides the data into five distinct subsets, running a training cycle and validating the model five times, using a different subset as validation data. This comprehensive approach gives a reliable estimate of how the model might perform on unseen data. Moreover, out of the 1 406 688 unique samples, a majority, 85% or 1 195 688 samples, were used in the cross-validation dataset, with the remaining 15% (211 000 samples) set aside as a testing dataset. This large set of previously unseen data, held throughout the model's training phase, helps correctly measure the model's performance and provides a safety net against overfitting. Furthermore, early stopping was used during the model's training process. This strategy continuously evaluates the model's performance on the validation set during training and stops the procedure if the validation error rises. This proactive technique protects against overfitting caused by intrinsic noise in training data. As a result, these metrics balanced the tradeoff between bias and variance, which leads to limiting models' overfitting.

IV. Results and Discussion

We first evaluate the entire 6 month dataset, then investigate the result data from each individual month. The results of both evaluation are compared and analyzed to provide a comprehensive understanding of the pressure and RH patterns and the performance of the proposed ML algorithms. For either evaluations, we first present a section, which estimates the pressure values using regression techniques with Doppler spectral moments as input features. Then, based on the cascading ML algorithm, the second section estimates RH using the same features plus the predicted pressure values from the first step. The experiments were conducted on a desktop computer equipped with an i7-2600 K CPU at 3.40 GHz and 24 GB of RAM. The detailed data tables related to this section are included in the appendix of this article.

A. Analysis of Results From 6 Months Data

1) Pressure Prediction: Tables IV and V summarize comparisons of the performance of various ML algorithms, evaluated using six different metrics, namely root mean squared error (RMSE), mean squared error (MSE), correlation coefficient, mean absolute error (MAE), R-squared, prediction speed (observations/second), and training time in seconds. The analysis was conducted on both the cross-validation and testing datasets. The best four RMSEs and MSEs values achieved by the ML methods of the predicted pressure for the cross-validation dataset are (87.09, 7585.12), (92.43, 8542.43), (92.55, 8564.87), and (94.54, 8938.46) for the bagged ensemble tree (BET), coarse tree (CT), medium tree (MT), and wide neural network (WNN), respectively. Therefore, these models have the highest accuracy in evaluating the predicting mistakes for the cross-validation dataset and the lowest distances between the predicted mean and actual values. Moreover, the achieved MAE values of the same ML algorithms are 61.42, 64.03, 64.07, and 64.67 for the bagged ensemble tree (BET), coarse tree (CT), fine tree (FT), and medium tree (MT), respectively. Furthermore, based on cross-validation dataset analysis, the correlation coefficients for the best four ML models are 0.91, 0.9, 0.89, and 0.88 for FT, BET, MT, and CT, respectively. The tree models and the BET offer the best outcomes in correlation. The best three R-squared values of the utilized models for the cross-validation dataset are 0.78, 0.75, and 0.75 for the BET, MT, and CT, respectively. This demonstrates that the BET model explains 78% of the variability found in the pressure variable. Linear regression, on the other hand, performs poorly, implying that the selection criteria are not linear.

On the other hand, the best four RMSEs and MSEs values achieved for pressure estimation are (86.64, 7505.88), (92.26, 8511.95), (92.28, 8516.51), and (93.63, 8766.93) for the WNN, BET, narrow neural network (NNN), and boosted ensemble tree (BoostedET), respectively. Further, the MAE values of the same ML algorithms were 60.98, 63.19, 63.56, and 64.36 for the WNN, BoostedET, BET, and NNN, respectively. These results

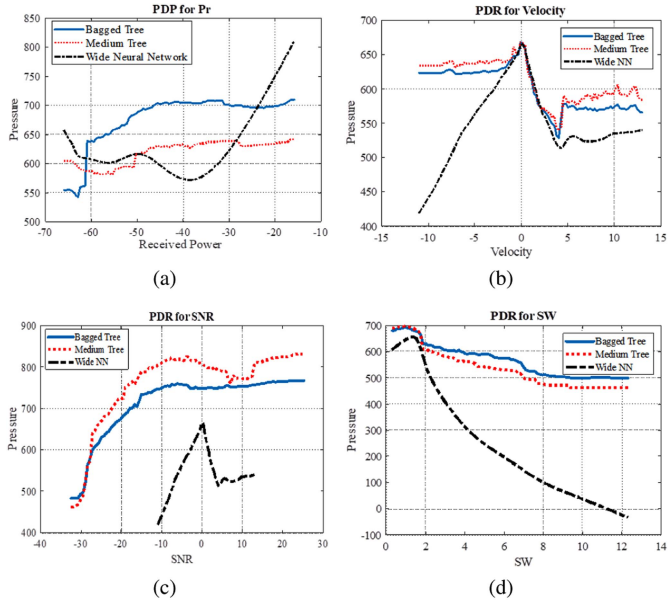


Fig. 3. Partial dependence plots (PDR) between pressure and inputs features of the six-month dataset. (a) PDR for Pr and pressure, (b) PDR for velocity and pressure, (c) PDR for SNR and pressure, (d) PDR for SW and pressure.

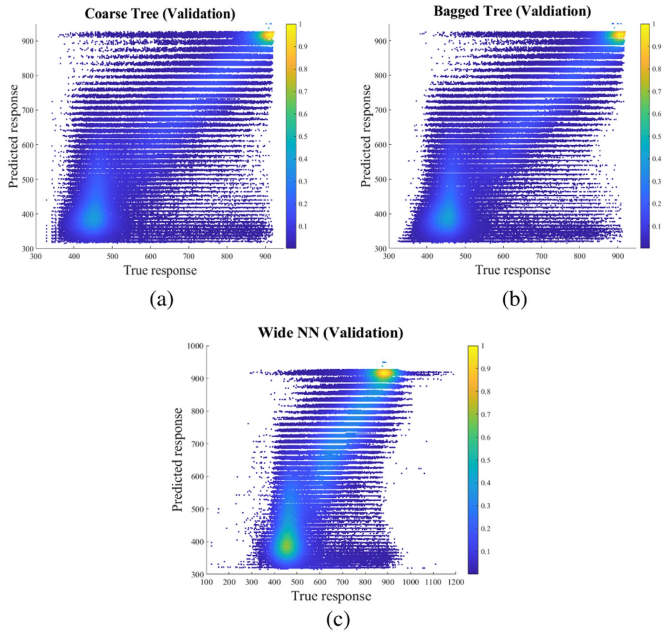


Fig. 4. Scatter density graphs with a color map of predicted versus actual pressure of the (all months) validation dataset using different ML methods. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

prove that the BET and WNN methods have better stability than the others in having low estimation errors for pressure parameters in cross validation and testing datasets. According to the testing dataset evaluation, the correlation coefficients for the top four ML models were 0.91, 0.9, 0.89, and 0.88 for BoostedET, WNN, BET, and NNN, respectively, correspondingly. As can be seen, the ensemble tree and NN models produce the highest performance in terms of correlation. The top three R-squared values for the BET, WNN, and NNN for the testing dataset were 0.78, 0.78, and 0.76, respectively, indicating that the BET and WNN models can account for 78% of the variability in the pressure variable.

In term of computational speed, all of the employed NNs can predict up to 1 445 728 observations per second, the linear regression model can estimate roughly 652 257 observations per second, the decision trees can estimate approximately 603 824 observations per second, and the ensemble trees can handle 128 835 observations per second. However, the NN methods and the FDT have longer training times than the other models, with the WNN, which has 100 neurons, being the slowest. In contrast, the linear regression and ensemble trees were the fastest models during training, taking around 37 s and 560 s, respectively.

In summary, the bagged ensemble tree was found to have the optimal overall performance. So it is used to estimate pressure value, which are then used as an input feature to estimate the RH in the next step.

A partial dependence plot (PDP) is shown in Fig. 3 to better illustrate the relationship between the input features and the target response. This plot demonstrates how the predicted target response varies with changes in the input features. For example, the expected pressure increases as the received power increases above -40 dBm. In the case of the WNN model, there is a linear relationship between the velocity and the pressure when the velocity is less than 0. On the other hand, the BET and MT models exhibit an exponential relationship between the SNR and the pressure, mainly when the SNR is around -25 dB. Conversely, all models demonstrate an inverse relationship between the spectrum width and the pressure values. These observations highlight the importance of considering the impact of individual input features on the predicted target response.

Scatter plots of truth and the estimated pressure values are displayed as density figures with color maps in two ways to address the issue of many sample points in the plots. Fig. 4 shows a typical scatter density plot with a color map of the estimated pressure by the multiple regression models versus the actual pressure data. Fig. 5 presents a different shape of scatter density plot that includes minisquares with a color map inside the plot. All these figures describe using the cross-validation data of 6 months. According to these plots, a slight improvement can be observed for the bagged tree model compared to the other two mentioned models: coarse tree and wide NN, since the scatter points for the bagged tree are concentrated more around the center than the other models. Figs. 6 and 7 show a normal scatter density plot with a color map and a scatter density plot with squares of the predicted pressure by the three ML models versus the truth pressure values. Similar to the cross-validation dataset, the bagged tree model performs better than CT and WNN because the points are clustered closer to the center.

As illustrated in Fig. 8, three pressure profiles are computed sequentially from the predicted pressure values using the BET, CT, and WNN and then compared to simultaneous radiosonde profiles. This figure shows how well the proposed ML algorithms can predict the pressure at different heights from 0 to 10 km, which is evident that the Bagged tree could estimate the pressure values at various altitudes with high accuracy, significantly

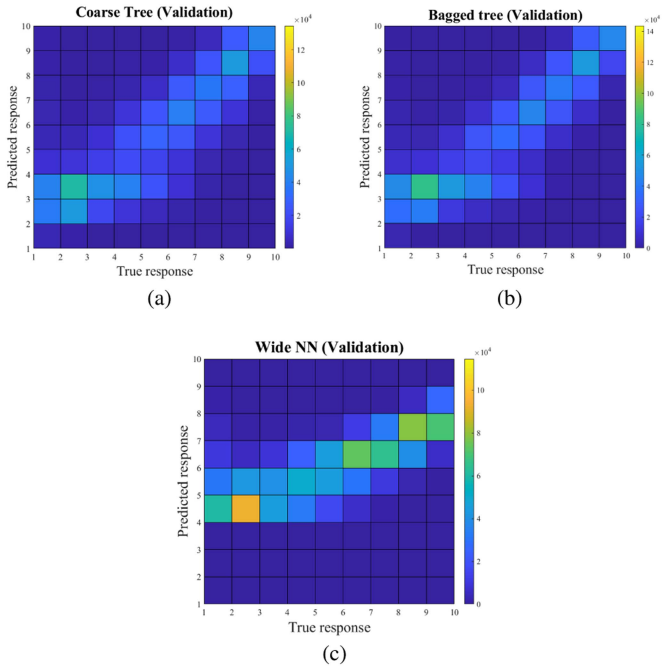


Fig. 5. Different scatter density plots with color maps within of expected versus actual pressure of the (all months) validation dataset using various ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

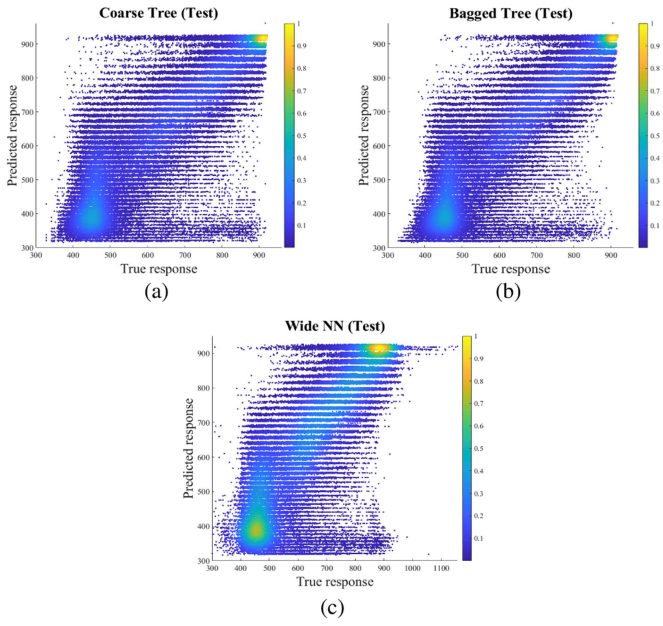


Fig. 6. Scatter density graphs with a color map of expected versus actual pressure of the (all months) test dataset generated by the following ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

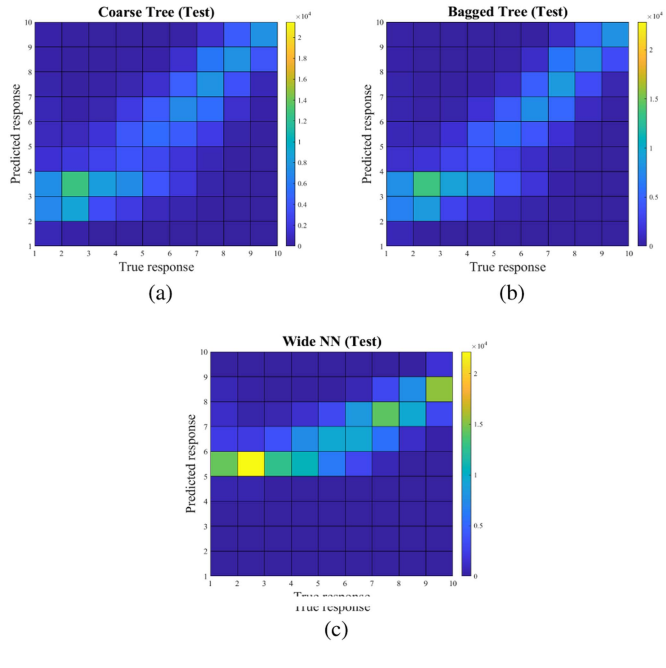


Fig. 7. Scatter density graphs with color maps of estimated versus truth pressure of the (all months) test dataset using several ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN (WNN).

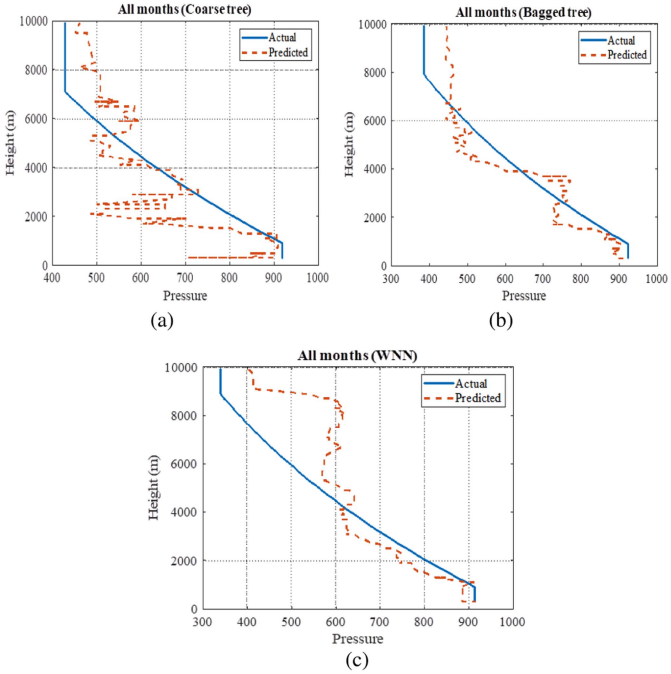


Fig. 8. Vertical profiles of predicted and actual/true pressure values for the three ML algorithms. (a) Bagged tree. (b) Coarse tree. (c) Wide NN.

below 7 km. Similar observations can be seen in both WNN and coarse tree plots. Finally, it is worth mentioning that all the plotted samples are random samples from the 6 month dataset to prove the ability of these ML methods to predict the pressure values during different seasons and times.

2) Humidity Estimation: Next, we uses the same dataset of 6 month, including the four input features, along with the previously predicted pressure to estimate the RH. The data are split into two datasets: cross-validation, with 1 195 688 samples, and testing, with 211 000 samples. The same 11 ML models with the same hyperparameters were used in the pressure prediction stage to train and test the dataset. Tables VI and VII show the performance details of the ML algorithms in terms of the

six performance metrics for the cross-validation and the testing datasets, respectively.

The top three RMSEs and MSEs results of the cross-validation dataset based on the RH (RH) are (22.2, 493.02), (24.17, 584.14), and (24.47, 598.7), respectively, corresponding to the BET, CT, and WNN methods. On the other hand, the ML algorithms with the best MAE performances are BET, MT, and FT, with result values of 17.44, 18.76, and 18.88, respectively. In addition, the FT and BET models had the highest correlation coefficients at 0.86 and 0.83, respectively, with a significant difference from the other ML approaches. Moreover, the best achievable R-squared value for the BET in the cross-validation dataset is 0.48, indicating that it explains 48% of the variability seen in the humidity.

For the testing dataset, the ML methods that achieved the minimal RMSEs and MSEs values of the predicted humidity are (21.9, 479.51), (24.04, 578.03), and (24.5, 600.05) using BET, CT, and MT. Meanwhile, the best achieved MAE values are 17.14, 18.43, and 18.45 for the BET, FT, and MT. The BET, MT, and CT algorithms obtained the best correlation coefficient values, which are 0.70, 0.62, and 0.62, respectively. As can be noticed, the ensemble and decision tree models have the highest correlation coefficient, indicating that these models have linear correlations between truth and estimated humidity values. Furthermore, the BET technique has the highest R-squared value (0.49) for the testing dataset.

Similarly to the pressure value estimating, all of the NNs, decision trees, linear regression, and ensemble trees can predict up to 1 406 412, 836 348, 519 682, and 121 471 observations per second, respectively. The NNs and FDT had the slowest training time compared to the other models. The linear regression model and the BoostedET are the fastest models throughout the training phase, with times of roughly 53 s and 803 s, respectively. As a result, the bagged ensemble tree is the best ML model to predict RH when compared to the other proposed ML techniques.

The partial dependence plot (PDP) for the four input features (velocity, SNR, SW, and projected pressure) versus RH is shown in Fig. 9. The bagged tree displays a nonlinear connection with humidity in the PDR for velocity and SW values. When the SNR is less than −10 dB, it can be shown that there is a linear relationship between the SNR and the humidity for the BET and CT. The link between anticipated pressure and RH is linear in the last plot for all the models presented.

Figs. 10 and 11 illustrate two scatter density plots with colormaps of the three predicted humidity versus the truth humidity, using three different algorithms. All of these results are based on the 6 months of cross-validation data. These results show that the bagged tree model performs better than the other two models. For the testing dataset, Figs. 12 and 13 show a typical scatter density plot with a colormap and a scatter density plot with squares of the predicted RH versus the actual humidity values. Similarly, the bagged tree model outperforms all of the other techniques. As shown in Fig. 14, a humidity profile is generated progressively from predicted humidity values using the best ML model (BET) compared to the corresponding radiosonde profile. The plotted humidity profile was selected randomly from the 6 month dataset to demonstrate the BET method's ability to

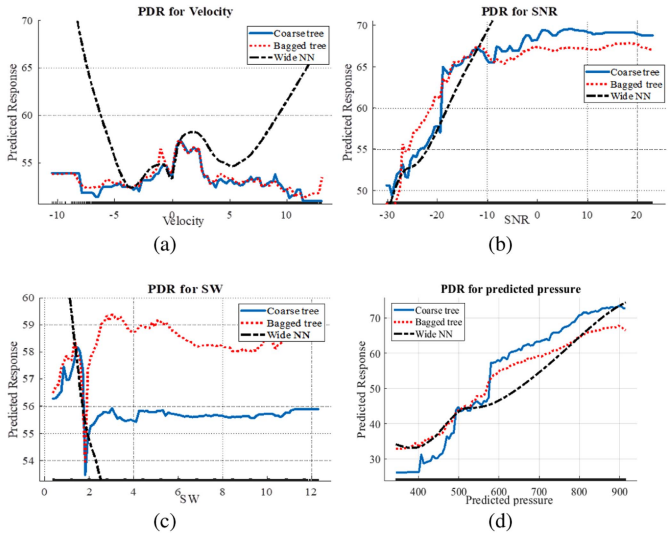


Fig. 9. PDRs between humidity and input variables of the whole 6 month dataset. (a) PDR for velocity and humidity. (b) SNR and humidity. (c) SW and humidity. (d) Predicted pressure and humidity.

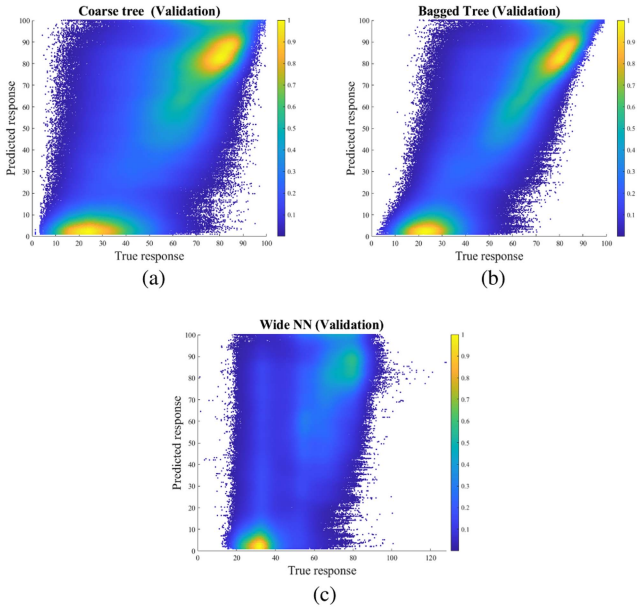


Fig. 10. Scatter density plots with a color map of predicted versus actual humidity of the validation dataset (all months) using several ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

estimate humidity at different periods. The graph depicts the BET technique prediction of RH at various heights ranging from 0 to 10 km. The BET could estimate humidity at various heights with good accuracy, particularly below 5 km.

3) Computational Efficiency: The suggested ML algorithms' computational efficiency is assessed in terms of training time and memory usage. Tables IV and VI show a detailed comparison of several models in predicting atmospheric pressure and humidity with 6 months of cross-validation data. The interactions linear regression (ILR) model was shown to be the most time-efficient in the context of pressure prediction, taking

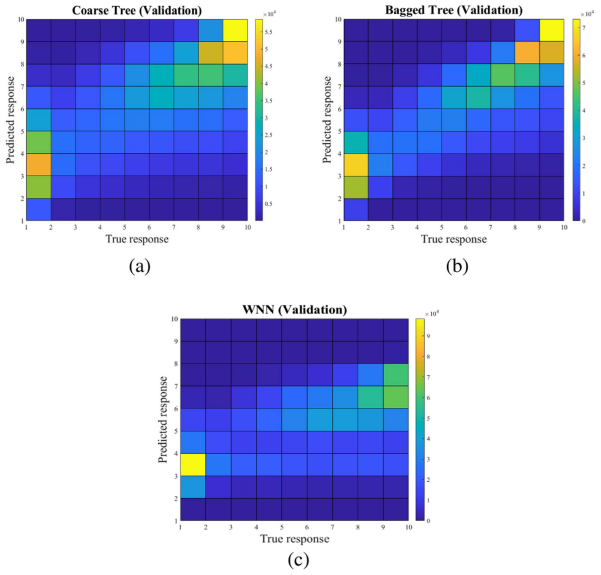


Fig. 11. Different scatter density plots with color maps of predicted versus true humidity of the (all months) validation dataset utilizing the following ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

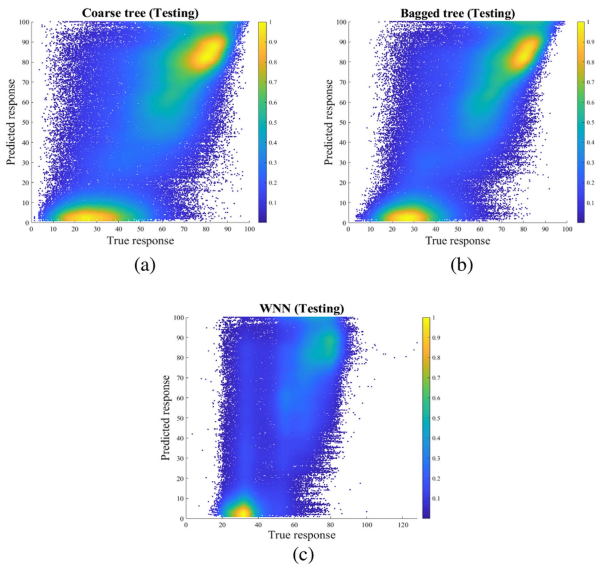


Fig. 12. Scatter density graphs with a color map of predicted versus true humidity of the (all months) test dataset created by the following ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

just 37.51 s for training. The WNN model, on the other hand, required the most training time, clocking in at 18 863.05 s. In addition, ensemble methods such as the boosted ensemble tree and bagged ensemble tree models showcased average training times. Similar patterns were seen during humidity prediction, when the ILR model displayed efficiency in training time, in contrast to the WNN model's long training period. As shown in Table III, the lowest, maximum, and mean memory consumption values are summarized using the whole 6 month dataset, providing an understanding of these methods' computational needs and resource use. The coarse tree approach used the least memory for pressure prediction, with a mean memory

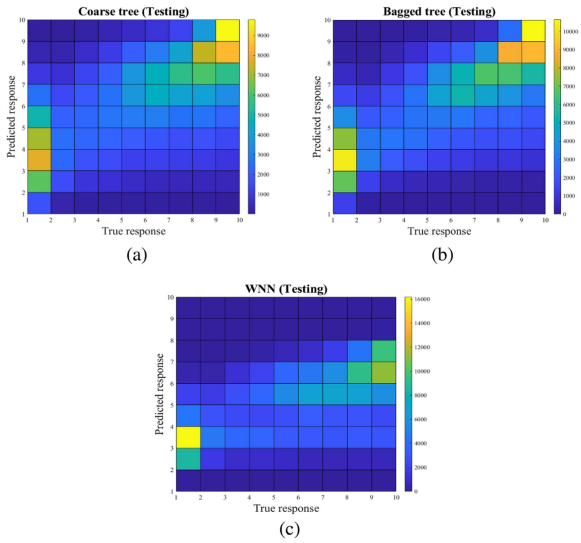


Fig. 13. Different scatter density graphs with color maps of expected versus actual humidity of the (all months) test dataset using the ML algorithms. (a) Coarse tree. (b) Bagged tree. (c) Wide NN.

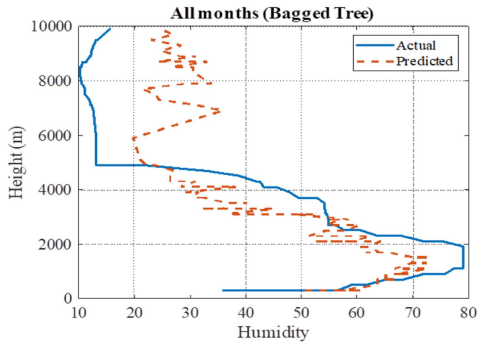


Fig. 14. Bagged tree algorithm's vertical profile of predicted and true RH values.

TABLE III
COMPARISON OF THE MOST SIGNIFICANT ML MODELS' PERFORMANCE IN TERMS OF MEMORY CONSUMPTION

	Min memory consumption (Mbytes)	Max memory consumption (Mbytes)	Mean memory consumption (Mbytes)
Coarse tree (6 months, pressure prediction)	90	595	426.6462
Bagged tree (6 months, pressure prediction)	140	2139	1.0172e+03
WNN (6 months, pressure prediction)	666	3301	2.2607e+03
Coarse tree (6 months, humidity prediction)	469	809	668.9355
Bagged tree (6 months, humidity prediction)	351	2412	1.4913e+03
WNN (6 months, humidity prediction)	416	3086	2.1123e+03

use of 426.6462 Mbytes. In contrast, the WNN algorithm consumed the most memory, with a mean of 2.2607e + 03 Mbytes. Humidity prediction followed a similar trend, with the Coarse Tree approach using the least amount of memory (668.9355 Mbytes) while the WNN algorithm used the most (2.1123e + 03 Mbytes). These memory usage patterns indicate significant differences in processing needs across different ML algorithms

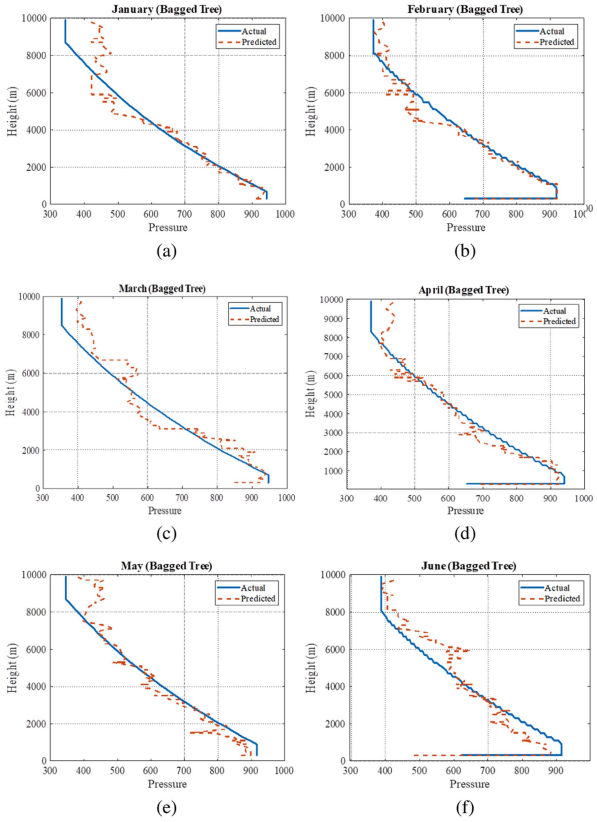


Fig. 15. Vertical pressure profiles of the Bagged tree during the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

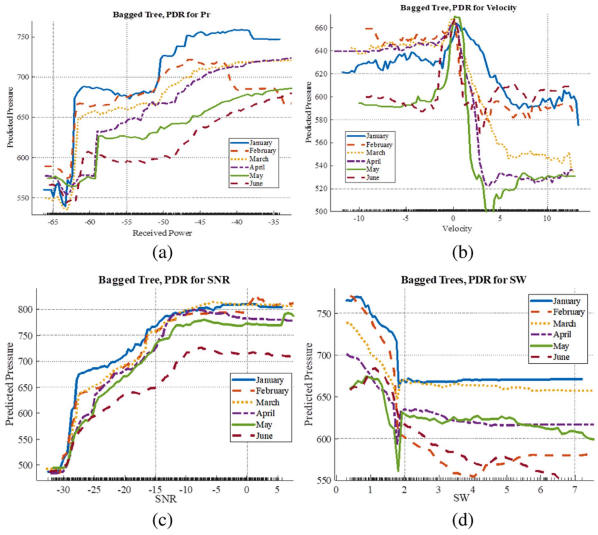


Fig. 16. PDRs between pressure and input features (for each month individually). (a) PDR for P_r and pressure. (b) Velocity and pressure. (c) SNR and pressure. (d) SW and pressure.

and prediction workloads. These findings have practical significance, especially in resource-constrained circumstances where memory restrictions may be challenging.

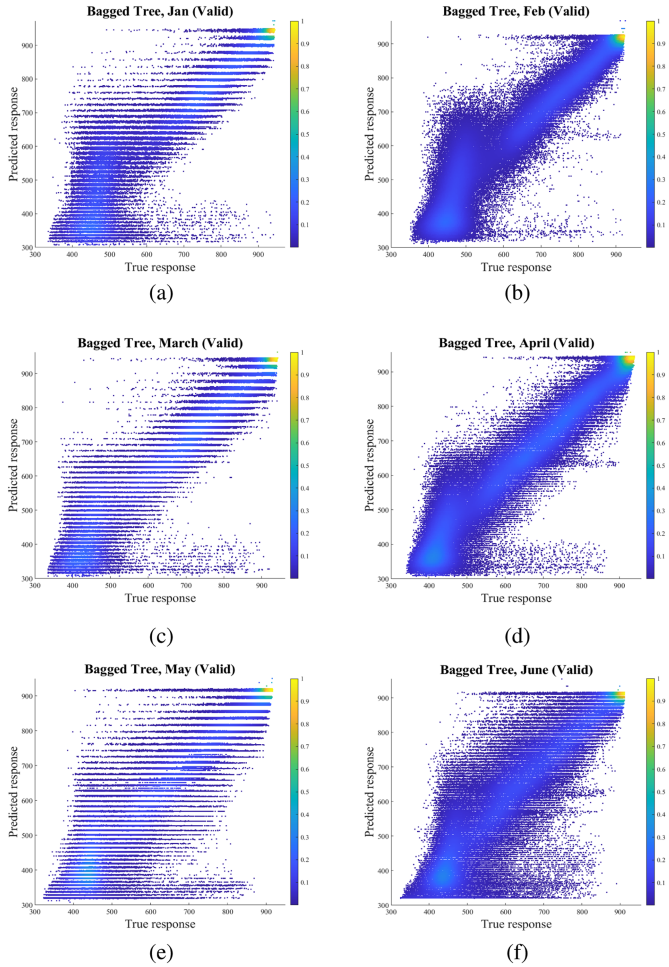


Fig. 17. Scatter density graphs of predicted versus true pressure values of the validation dataset (for each month separately) using bagged tree, for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

B. Month-by-Month Evaluations

1) Pressure Estimation: During various months and different seasons, there can be significant changes in pressure and RH. As a result, estimation performance of each individual month is further examined. To conduct this analysis, the dataset for each month includes all the Doppler spectral moments and radiosonde data collected during that month are used. The total samples collected in January, February, March, April, May, and June are 158 766, 185 791, 194 460, 272 808, 286 727, and 308 140, respectively.

All the datasets for different months were trained using the same ML models. However, for the sake of brevity, we only list the best-performing ML algorithms in the decision tree, ensemble tree, and ANN. Tables VIII and IX include detailed data on each month's effectiveness of the top three ML algorithms. In addition, these tables include information on all performance evaluation methodologies, including root mean squared error (RMSE), mean squared error (MSE), correlation coefficient, mean absolute error (MAE), prediction speed (observations/second), and training time in seconds.

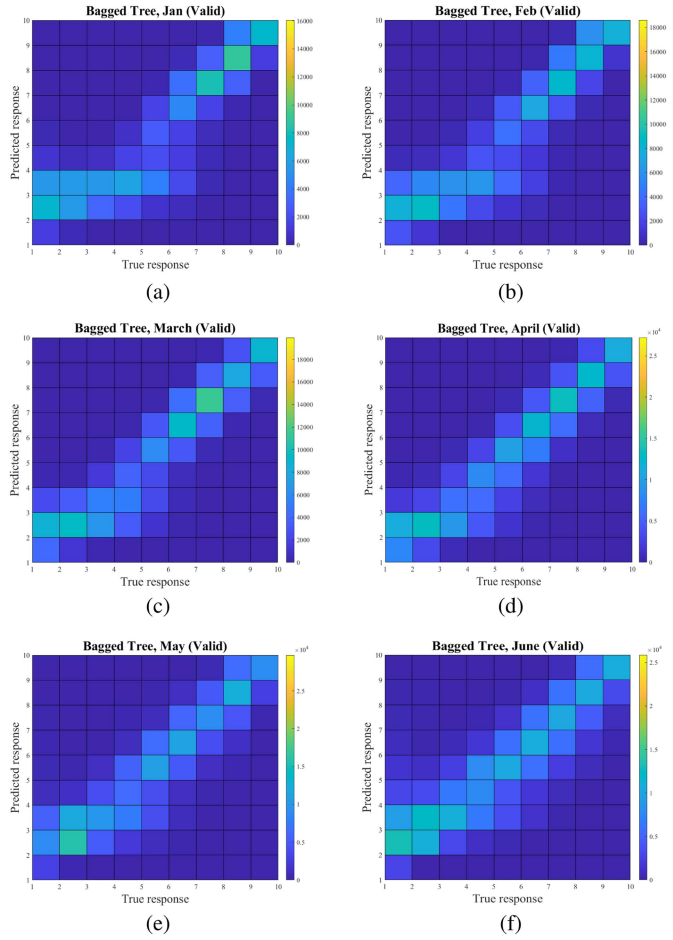


Fig. 18. Different scatter density graphs of predicted versus true pressure values from the validation dataset (for each month individually) using bagged tree, for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

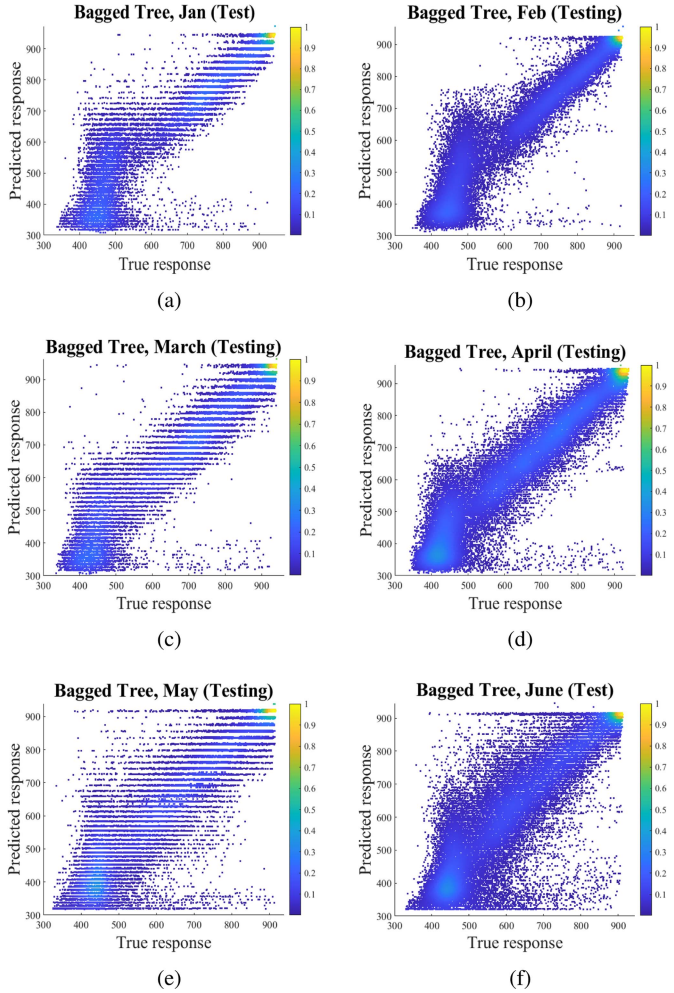


Fig. 19. Scatter density graphs showing the test dataset's predicted versus true pressure values (for each month individually) using bagged tree, for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

According to both tables, the BET technique outperformed the WNN and CT methods in terms of RMSEs, MSEs, and MAE in all months, with a substantial difference. This demonstrates that BET has the best estimation accuracy for pressure. Furthermore, based on cross-validation dataset analysis, the BET technique achieved the following correlation coefficient values: 0.93, 0.94, 0.95, 0.95, 0.93, and 0.9 for January, February, March, April, May, and June, accordingly. Based on testing dataset analysis, the following correlation values were obtained: 0.93, 0.94, 0.95, 0.95, 0.93, and 0.91 for January to June.

In addition, the BET model consistently outperformed the other models regarding R-squared values for both cross-validation and testing datasets across all months. For example, the BET in the March dataset obtained the highest R-squared value of 0.88, indicating that the BET model accounts for 88% of the variability in the pressure variable. These findings demonstrate that the BET technique is more reliable and stable than the other methods in terms of having minimal errors in pressure predictions across all datasets. On average, the WNN can predict the highest number of observations per second, while the bagged ensemble trees can observe the least observations per second. However, regarding the training speed, the WNN takes the most time, followed by the BET and finally by the CT. To summarize, it seems clear that the Bagged ensemble tree is the best ML model for effectively predicting pressure.

Fig. 15 displays multiple pressure profiles created from predicted pressure values in all 6 months using the best ML model (BET) compared to the related radiosonde profile. The BET model has high accuracy in predicting pressure at various heights, as seen in the monthly plots. Also, the plots show good agreement between the predicted and truth pressure values, indicating that the BET model effectively captures the underlying patterns and trends in the data. Fig. 16 depicts the partial dependence plot (PDP) for the four input characteristics (Pr, velocity, SNR, and SW) against pressure. Each of these numbers contains the outcomes of all months utilizing BET. The BET behaves differently depending on the pressure readings in other months while having a linear relationship with pressure in the PDR for Pr and SNR values. When the velocity is more significant than zero, it is possible to demonstrate an

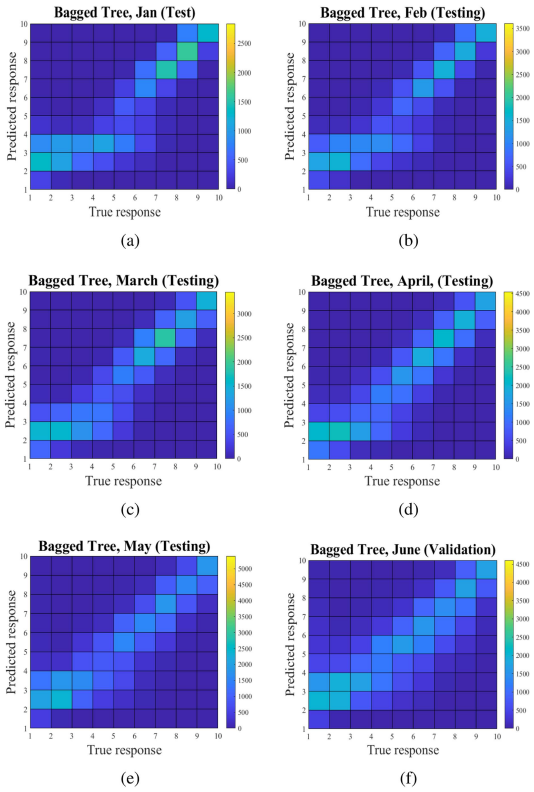


Fig. 20. Different scatter density plots of predicted versus true pressure values from the test dataset (for each month separately) using bagged tree for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

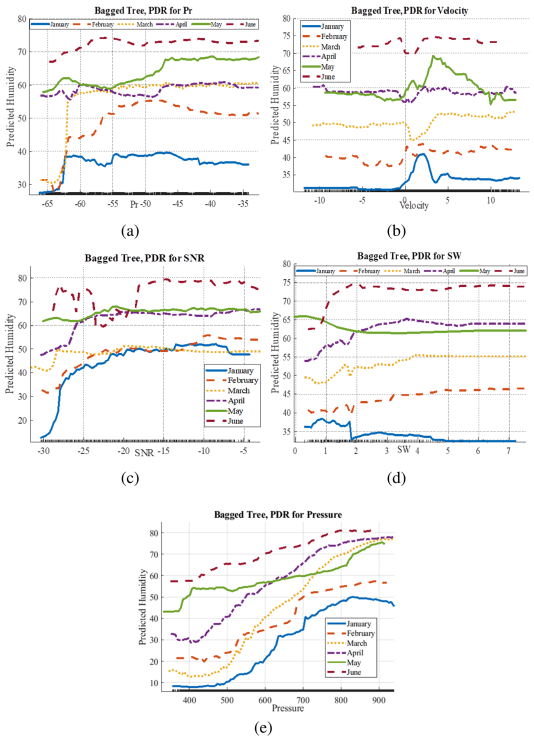


Fig. 21. PDRs for relative humidity and input variables (for each month individually). (a) PDR for P_r and humidity. (b) Velocity and humidity. (c) SNR and humidity. (d) SW and humidity. (e) Predicted pressure and humidity.

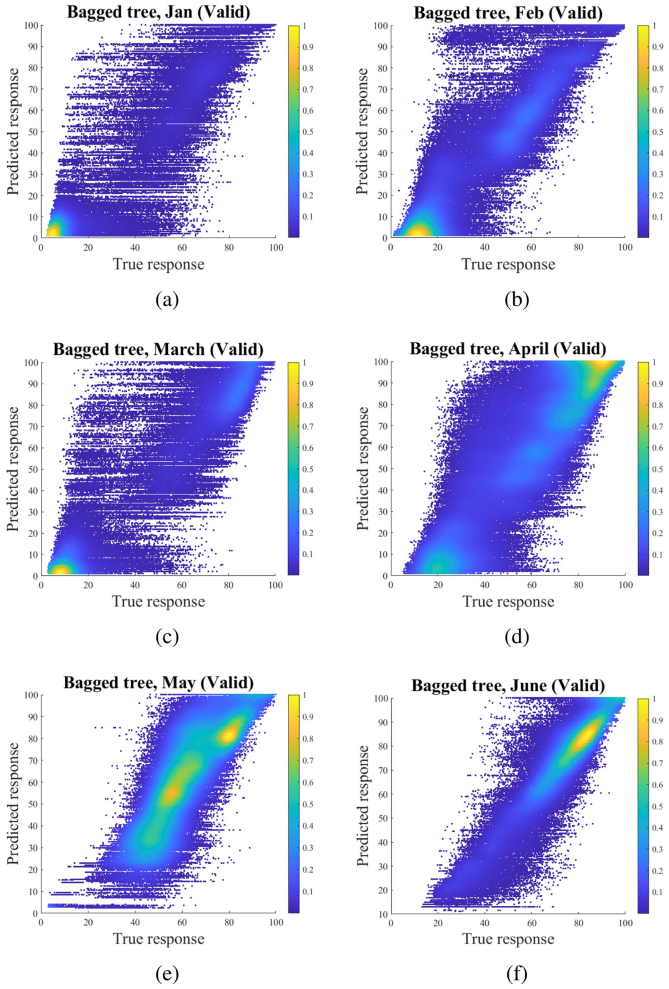


Fig. 22. Scatter density graphs showing the validation dataset's predicted versus true humidity values (for each month individually) using bagged tree, for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

inverse connection between the velocity and the pressure for the BET in all months. The same is true for the SW feature, which exhibits an inverse relationship when SW is smaller than two.

Figs. 17 and 18 show two scatter density plots with color maps of predicted pressure using BET versus the truth pressure data. These results are based on cross-validation data of each individual month. The bagged tree model works well since most scatter points cluster around the center. Figs. 19 and 20 for the testing dataset show a typical scatter density plot with a color map, and a scatter density plot with squares of the BET's predicted pressure versus true pressure data.

2) Humidity Estimation: Each month's dataset, including the Doppler spectral moments and predicted pressure, was trained using the same set of ML algorithms as in the previous section, to predict the RH, and we compared the results of decision tree, ensemble tree, and ANN here. Tables X and XI, which exhibit the cross-validation and testing datasets, include detailed data on each month's results of the three ML algorithms. In addition, these tables include results for all the performance metrics (root

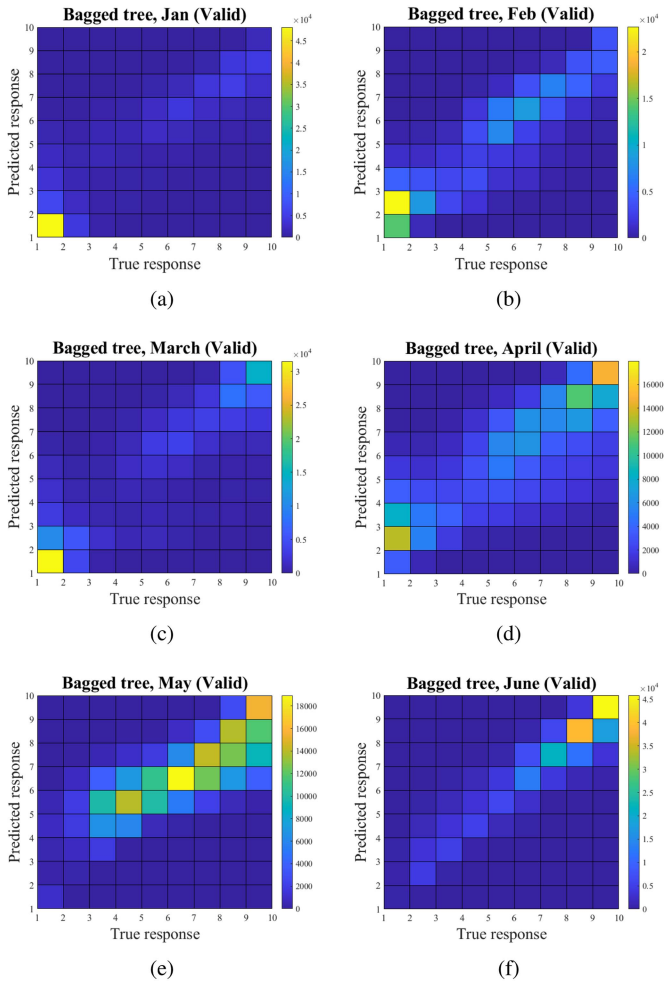


Fig. 23. Different scatter density graphs of predicted versus true humidity values from the validation dataset (for each month individually) using bagged tree, for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

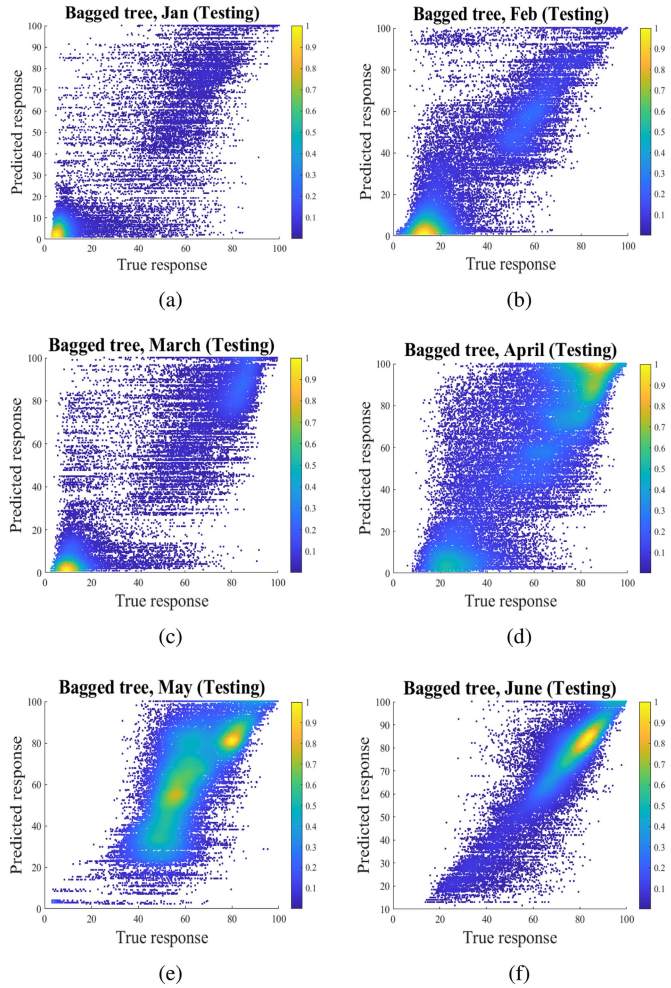


Fig. 24. Scatter density graphs for the following months of the test dataset's predicted versus true humidity results (for each month separately) using bagged tree. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

mean squared error (RMSE), mean squared error (MSE), correlation coefficient, mean absolute error (MAE), prediction speed (observations/second), and training time in seconds). According to both tables, the BET technique surpasses the WNN and CT techniques in all months in terms of RMSEs, MSEs, and MAE. The BET approach produced good correlation coefficient values based on cross-validation dataset analysis, ranging from 0.83 to 0.94 for May and June. While the corresponding values for testing dataset analysis ranged from 0.71 to 0.91 for May and June, respectively. Further, the BET had the good R-squared values for cross-validation and testing datasets across all months. For example, the best R-squared value among all months for the BET is 0.78 in the March dataset. Overall, the bagged ensemble tree is the most accurate ML model for forecasting RH based on the results from each individual month.

The partial dependence plot (PDP) for the five input variables (P_r, velocity, SNR, SW, projected pressure) against RH is shown in Fig. 21. Each feature is seen to have a reasonable correlation with humidity, but the relationship scales differently among different months. Figs. 22 and 23 illustrate scatter density plots of the estimated humidity using BET versus the true humidity values. These results are based on cross-validation data collected for each month separately. The scatter density plot with a colormap of the BET's projected humidity versus actual humidity data is shown in Figs. 24 and 25 for the testing dataset. These figures show that the bagged tree model works well since most of the scatter points cluster around the center line. Fig. 26 shows different humidity profiles generated from predicted humidity values over all 6 month using the BET method compared to the relevant radiosonde profile. The BET method accurately estimates humidity at various heights for all the 6 months.

C. Advantages and Limitations of the Study

This study has shown numerous significant strengths and improvements in atmospheric humidity estimation. First, the proposed innovative cascaded ML approach outperforms previous approaches significantly. Second, this approach uses raw moment data from a wind profiler radar as ML feature variables,

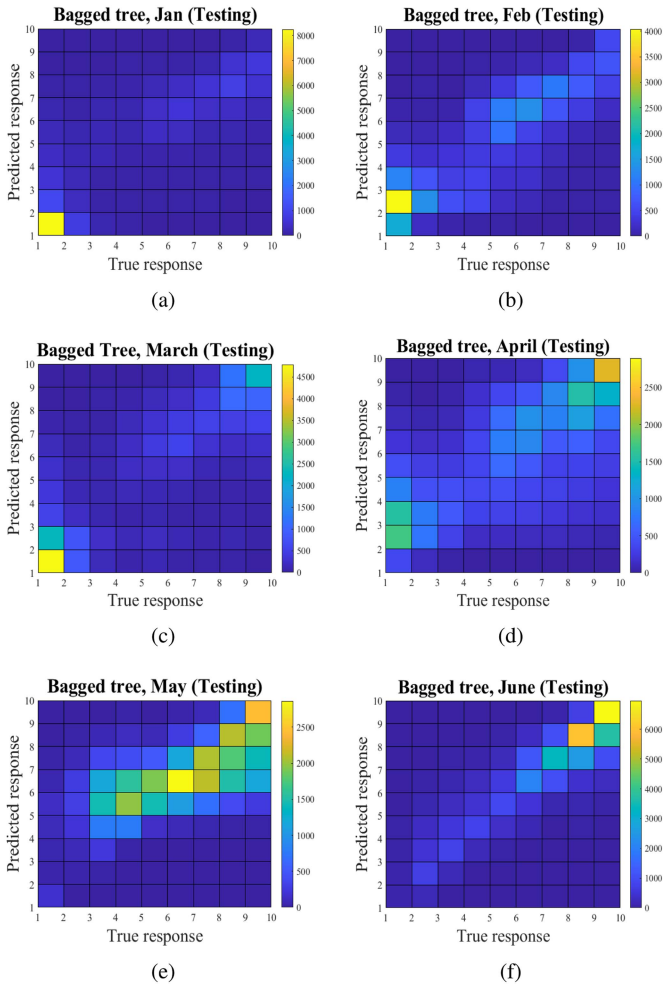


Fig. 25. Different scatter density plots of estimated versus true RH from the test dataset (for each month independently) using bagged tree for the following months. (a) January. (b) February. (c) March. (d) April. (e) May. (f) June.

simplifying and decreasing the model's complexity. Third, the lack of dependence on accurate wind estimates, temperature, and pressure data, which are sometimes unavailable, emphasizes the resilience and usefulness of the suggested technique. As part of the implementations, we applied an intermediate training stage using synthetic pressure data initially introduced in this study. This cascade ML system has demonstrated promising application potential in calculating various atmospheric parameters. Finally, the bagging learning tree algorithm's superior computational efficiency and prediction efficacy distinguish this study and improve the reliability and performance compared to existing physical models. For the limitations, the reliance on high-quality I/Q or moment data from the profiler radar is a significant restriction. Noise, data incompletion, or inconsistencies in this data might impact the model's performance. Furthermore, the model's ability to be generalized may be constrained by its training on data from a specific site over a defined period. Environmental and climatic variability may challenge the model's prediction accuracy across multiple geographical areas and time intervals. Also, although measures were taken to minimize overfitting, the model may find challenges

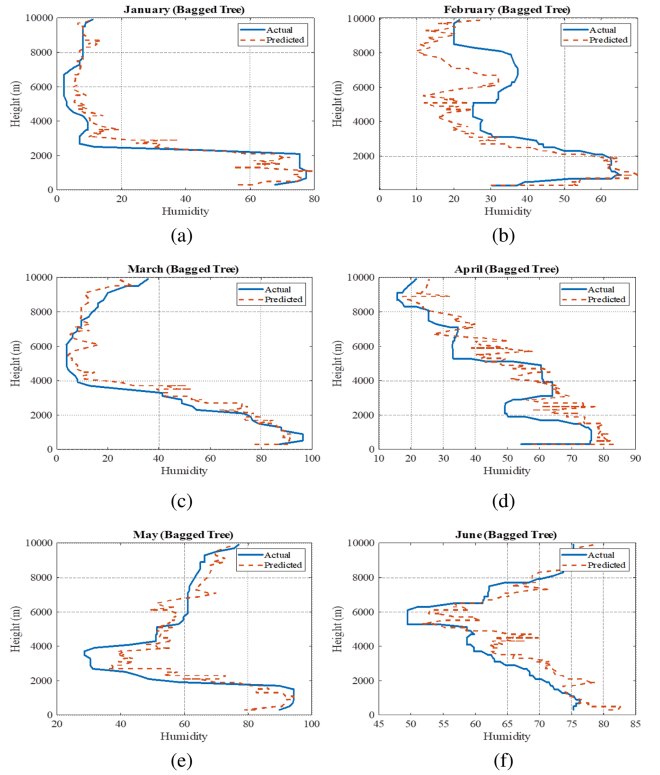


Fig. 26. Vertical RH profiles of the Bagged tree in (a) January, (b) February, (c) March, (d) April, (e) May, and (f) June.

when dealing with scenarios that diverge greatly from those seen in the training data.

V. Conclusion and Future Work

This article proposes a novel cascaded ML approach for estimating atmospheric humidity from wind profiler radar data. Unlike existing methods, this approach uses moment data from the raw spectrum as ML feature variables and predicts RH in two cascaded steps. The bagging learning tree algorithm is identified as the most effective and computationally efficient method for predicting pressure and RH. The proposed approach improves reliability and performance compared to physical models, and does not require precise wind estimates, temperature, and pressure parameters from radiosonde data that are not always available. The results show that this approach can provide reasonably accurate humidity estimates. In addition, the cascaded ML solution produces an intermediate training stage with "synthetic" pressure data, which is applied to the humidity estimation problem for the first time. This method has the potential to be extended to estimate other atmospheric parameters, and future studies could explore the integration of different sensor data, such as radiometers, as well as more sophisticated ML techniques.

Appendix
Summary of the Result Tables

TABLE IV
COMPARISON OF PERFORMANCE OF THE PROPOSED ML MODELS IN PREDICTING PRESSURE FOR THE 6 MONTH OF CROSS-VALIDATION DATA

Model Type	RMSE (Validation)	MSE (Valid)	ρ (Valid)	R^2 (Valid)	MAE (Valid)	Prediction Speed (obs/sec)	Training Time (sec)
ILR	108.52	11776.85	0.81	0.66	80.03	652257.94	37.51
FT	94.57	8942.63	0.91	0.74	64.03	417087.25	12216.47
MT	92.55	8564.87	0.89	0.75	64.07	419819.87	4188.86
CT	92.43	8542.43	0.88	0.75	64.67	603824.96	561.64
BoostedET	103.59	10730.38	0.85	0.69	79.77	128835.49	551.78
BET	87.09	7585.12	0.9	0.78	61.42	23874.35	2615.18
NNN	102.05	10413.95	0.85	0.7	75.11	860477.63	5651.68
MNN	97.41	9487.97	0.85	0.73	70.59	1018433.15	10905.88
WNN	94.54	8938.46	0.86	0.74	67.86	1187889.67	18863.05
BiNN	98.15	9633.41	0.85	0.72	71.53	951444.23	9758
TriNN	97.39	9483.95	0.85	0.73	70.57	1445728.81	13432.38

TABLE V
COMPARISON OF PERFORMANCE OF THE PROPOSED ML MODELS IN PREDICTING THE PRESSURE FOR THE 6 MONTHS OF TEST DATA

Model Type	RMSE (Test)	MSE (Test)	MAE (Test)	ρ (Test)	R^2 (Test)
ILR	118.06	13937.82	87.64	0.78	0.6
FT	109.2	11923.67	80.42	0.81	0.66
MT	120.34	14481.19	85.79	0.77	0.58
CT	109.2	11923.67	80.42	0.81	0.66
BoostedET	93.63	8766.93	63.19	0.91	0.75
BET	92.26	8511.95	63.56	0.89	0.78
NNN	92.28	8516.51	64.36	0.88	0.76
MNN	104.01	10819.05	79.96	0.84	0.69
WNN	86.64	7505.88	60.98	0.9	0.78
BiNN	98.97	9794.77	71.87	0.85	0.72
TriNN	97.44	9493.95	70.71	0.85	0.73

TABLE VI
COMPARISON OF THE ML MODELS' PERFORMANCE IN FORECASTING HUMIDITY ACROSS 6 MONTHS OF CROSS-VALIDATION DATA

Model Type	RMSE (Valid)	MSE (Valid)	ρ (Valid)	R^2 (Valid)	MAE (Valid)	Prediction Speed (obs/sec)	Training Time (sec)
ILR	25.13	631.34	0.57	0.33	20.73	519682.08	53.05
FT	26.26	689.62	0.86	0.27	18.88	836348.75	25183.75
MT	24.75	612.49	0.76	0.35	18.76	463455.07	9708.12
CT	24.17	584.14	0.69	0.38	18.99	511306.23	878.89
BoostedET	24.92	621.14	0.59	0.34	20.69	121471.29	803.27
BET	22.2	493.02	0.83	0.48	17.44	18553.43	3137.75
NNN	24.93	621.59	0.58	0.34	20.46	934521.39	5927.35
MNN	24.68	609.08	0.59	0.35	20.16	749117.18	12012.23
WNN	24.47	598.7	0.6	0.36	19.92	887005.17	22554.27
BiNN	24.71	610.64	0.59	0.35	20.2	986172.83	12031.56
TriNN	24.7	609.85	0.6	0.35	20.18	1406412.82	15824.96

TABLE VII
COMPARISON OF THE PERFORMANCE OF ML MODELS IN PREDICTING HUMIDITY ACROSS A 6 MONTH TEST DATASET

Model Type	RMSE (Test)	MSE (Test)	MAE (Test)	ρ (Test)	R^2 (Test)
ILR	25.14	632.02	20.74	0.57	0.33
FT	25.88	670.02	18.43	0.6	0.29
MT	24.5	600.05	18.45	0.62	0.36
CT	24.04	578.03	18.81	0.62	0.39
BoostedET	24.93	621.68	20.7	0.59	0.34
BET	21.9	479.51	17.14	0.70	0.49
NNN	24.94	622	20.43	0.58	0.34
MNN	24.84	616.95	20.31	0.59	0.34
WNN	24.5	600.15	19.97	0.6	0.36
BiNN	24.77	613.76	20.27	0.59	0.35
TriNN	24.65	607.73	20.13	0.6	0.35

TABLE VIII
COMPARISON OF THE PERFORMANCE OF ML MODELS IN PREDICTING PRESSURE (THROUGHOUT EACH MONTH SEPARATELY) OF CROSS-VALIDATION DATA

Model Type	RMSE (Valid)	MSE (Valid)	ρ (Valid)	R^2 (Valid)	MAE (Valid)	Prediction Speed (obs/sec)	Training Time (sec)
CT (Jan)	84.26	7099.94	0.91	0.82	58.49	535609.04	11.28
BET (Jan)	80.33	6452.44	0.93	0.84	55.53	22469.62	161.9
WNN (Jan)	85.54	7317.74	0.9	0.81	60.8	1006409.75	1930.91
CT (Feb)	78.89	6222.91	0.92	0.82	54.84	850895.2	22.05
BET (Feb)	73.44	5393.84	0.94	0.85	51.05	32731.09	175.29
WNN (Feb)	79.86	6377.2	0.91	0.82	57.79	1010450.46	1965.07
CT (March)	70.98	5037.66	0.94	0.86	49.95	1291171.51	18.34
BET (March)	66.4	4409.26	0.95	0.88	46.77	43060.84	157.34
WNN (March)	71.38	5095	0.93	0.86	51.38	1081371.59	2161.29
CT (April)	71.91	5171.15	0.93	0.85	51.25	414760.43	25.77
BET (April)	67.18	4513.79	0.95	0.87	47.7	18572.45	332.55
WNN (April)	72.47	5252.1	0.92	0.85	52.43	1051270.32	3352.45
CT (May)	83.71	7007.81	0.9	0.8	57.49	1403378.45	32.53
BET (May)	76.77	5894.36	0.93	0.83	52.82	45653.43	261.5
WNN (May)	85.83	7367.31	0.89	0.79	61.09	1069922.17	3299.32
CT (June)	104.02	10821.15	0.85	0.68	72.15	435601.58	34.81
BET (June)	94.54	8938.68	0.9	0.74	65.77	20260.96	378.9
WNN (June)	107.69	11597.88	0.82	0.66	78.6	1063646.44	3864.2

TABLE IX
COMPARISON OF THE RESULTS OF ML ALGORITHMS IN PREDICTING PRESSURE IN THE TEST DATASET (FOR EACH MONTH SEPARATELY)

Model Type	RMSE (Test)	MSE (Test)	MAE (Test)	ρ (Test)	R^2 (Test)
CT (Jan)	83.41	6956.67	58.08	0.91	0.82
BET (Jan)	78.97	6236.99	54.8	0.93	0.84
WNN (Jan)	84.56	7150.8	60.13	0.9	0.82
CT (Feb)	77.65	6030.15	54.03	0.92	0.83
BET (Feb)	71.85	5162.92	49.92	0.94	0.85
WNN (Feb)	79.38	6301.04	57.35	0.91	0.82
CT (March)	70.89	5025.62	49.74	0.94	0.86
BET (March)	65.82	4332.83	46.28	0.95	0.88
WNN (March)	72.09	5197.21	51.51	0.93	0.86
CT (April)	73.03	5333.27	51.48	0.93	0.85
BET (April)	67.38	4540.44	47.38	0.95	0.87
WNN (April)	73.91	5463.33	52.88	0.92	0.85
CT (May)	82.88	6869.09	56.8	0.9	0.8
BET (May)	75.31	5671.3	51.69	0.93	0.84
WNN (May)	86.11	7414.59	61.26	0.88	0.79
CT (June)	102.65	10536.88	71.14	0.85	0.69
BET (June)	92.8	8611.9	64.4	0.91	0.75
WNN (June)	106.89	11425.78	78.09	0.82	0.66

TABLE X
COMPARISON OF THE RESULTS OF ML ALGORITHMS FOR PREDICTING RELATIVE HUMIDITY USING CROSS-VALIDATION DATASET (FOR EACH MONTH INDIVIDUALLY)

Model Type	RMSE (Valid)	MSE (Valid)	ρ (Valid)	R^2 (Valid)	MAE (Valid)	Prediction Speed (obs/sec)	Training Time (sec)
CT (Jan)	19.56	382.81	0.83	0.6483	13.14	1027135.92	23.38
BET (Jan)	17.52	306.97	0.91	0.7179	11.81	43860.84	146.46
WNN (Jan)	19.58	383.73	0.80	0.6475	13.66	1005029.16	1708.79
CT (Feb)	19.29	372.43	0.80	0.5713	14.06	1146322.46	19.66
BET (Feb)	17.33	300.55	0.89	0.6540	12.59	31802.08	212.85
WNN (Feb)	19.53	381.61	0.74	0.5607	14.83	829839.38	2221.16
CT (March)	19.58	383.51	0.86	0.6998	13.89	1172163.53	21.25
BET (March)	17.86	319.23	0.92	0.7501	12.70	41404.25	200.97
WNN (March)	19.56	382.92	0.83	0.7003	14.2.	1049074.34	2155.52
CT (April)	22.69	514.93	0.74	0.4769	17.60	1382439.82	28.19
BET (April)	20.77	431.45	0.86	0.5617	16.16	39426.27	325.61
WNN (April)	22.67	514.33	0.69	0.4775	18.16	1112739.98	3041.01
CT (May)	18.70	349.71	0.66	0.3290	14.41	1112555.65	30.83
BET (May)	16.53	273.32	0.83	0.4756	12.68	36747.93	393.39
WNN (May)	19.19	368.52	0.54	0.2929	15.42	1143452.58	3183.13
CT (June)	12.49	156.00	0.83	0.6310	8.07	1321420.39	35.26
BET (June)	9.68	93.83	0.940	0.7780	6.31	41432.77	389.53
WNN (June)	15.74	248.00	0.64	0.4133	11.74	1169097.01	3393.19

TABLE XI
COMPARISON OF ML ALGORITHMS' PREDICTIONS OF RELATIVE HUMIDITY IN THE TEST DATASET (FOR EACH MONTH INDIVIDUALLY)

Model Type	MAE (Test)	MSE (Test)	RMSE (Test)	Correlation coefficient (Test)	RSquared (Test)
CT (Jan)	13.04	375.62	19.38	0.81	0.65
BET (Jan)	11.52	293.05	17.12	0.86	0.73
WNN (Jan)	13.79	384.59	19.61	0.8	0.65
CT (Feb)	13.8	365.06	19.11	0.76	0.57
BET (Feb)	12.2	286.98	16.94	0.82	0.67
WNN (Feb)	14.83	385.61	19.64	0.74	0.55
CT (March)	13.95	386.69	19.66	0.84	0.7
BET (March)	12.51	314.26	17.73	0.87	0.75
WNN (March)	14.37	388.55	19.71	0.83	0.7
CT (April)	17.54	511.72	22.62	0.69	0.47
BET (April)	15.86	419.43	20.48	0.76	0.57
WNN (April)	18.2	515.54	22.71	0.69	0.47
CT (May)	14.17	341.44	18.48	0.6	0.35
BET (May)	12.34	262.1	16.19	0.71	0.5
WNN (May)	15.52	372.39	19.3	0.54	0.29
CT (June)	7.72	146.06	12.09	0.81	0.66
BET (June)	5.91	84.07	9.17	0.9	0.8
WNN (June)	11.77	247.82	15.74	0.64	0.42

REFERENCES

[1] D. Jacob, "The role of water vapour in the atmosphere. A short overview from a climate modeller's point of view," *Phys. Chem. Earth, Part A, Solid Earth Geodesy*, vol. 26, no. 6–8, pp. 523–527, 2001.

[2] D. L. Hartmann, *Global Physical Climatology*, vol. 103. Oxford, U.K.: Newnes, 2015.

[3] T. A. Guinn and R. J. Barry, "Quantifying the effects of humidity on density altitude calculations for professional aviation education," *Int. J. Aviation, Aeronaut., Aerosp.*, vol. 3, no. 3, 2016, Art. no. 2.

[4] J. C. O'Connor, M. J. Santos, S. C. Dekker, K. T. Rebel, and O. A. Tuinenburg, "Atmospheric moisture contribution to the growing season in the Amazon arc of deforestation," *Environ. Res. Lett.*, vol. 16, no. 8, 2021, Art. no. 084026.

[5] M. Chowdhury et al., "Effects of temperature, relative humidity, and carbon dioxide concentration on growth and glucosinolate content of kale grown in a plant factory," in *Foods*, vol. 10, no. 7, 2021, Art. no. 1524.

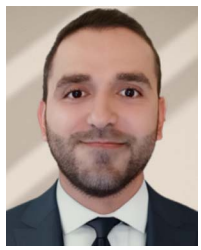
[6] D. Harel, H. Fadida, A. Slepoy, S. Gantz, and K. Shilo, "The effect of mean daily temperature and relative humidity on pollen, fruit set and yield of tomato grown in commercial protected cultivation," *Agronomy*, vol. 4, no. 1, pp. 167–177, 2014.

[7] K. E. Trenberth and C. J. Guillemot, "Evaluation of the atmospheric moisture and hydrological cycle in the NCEP/NCAR reanalyses," *Climate Dyn.*, vol. 14, pp. 213–231, 1998.

[8] S. Sherwood, "Direct versus indirect effects of tropospheric humidity changes on the hydrologic cycle," *Environ. Res. Lett.*, vol. 5, no. 2, 2010, Art. no. 025206.

[9] A. I. Barreca, "Climate change, humidity, and mortality in the United States," *J. Environ. Econ. Manage.*, vol. 63, no. 1, pp. 19–34, 2012.

[10] R. E. Davis, G. R. McGregor, and K. B. Enfield, "Humidity: A review and primer on atmospheric moisture and human health," *Environ. Res.*, vol. 144, pp. 106–116, 2016.
[11] F. Saíd, B. Campistron, and P. Di Girolamo, "High-resolution humidity profiles retrieved from wind profiler radar measurements," *Atmospheric Meas. Techn.*, vol. 11, no. 3, pp. 1669–1688, 2018.
[12] T. Tsuda, M. Miyamoto, and J.-I. Furumoto, "Estimation of a humidity profile using turbulence echo characteristics," *J. Atmospheric Ocean. Technol.*, vol. 18, no. 7, pp. 1214–1222, 2001.
[13] B. B. Stankov et al., "Humidity gradient profiles from wind profiling radars using the NOAA/ETL advanced signal processing system (SPS)," *J. Atmospheric Ocean. Technol.*, vol. 20, no. 1, pp. 3–22, 2003.
[14] S. Imura, J.-i. Furumoto, T. Tsuda, and T. Nakamura, "Estimation of humidity profiles by combining co-located VHF and UHF wind-profiling radar observation," *J. Meteorol. Soc. Japan. Ser. II*, vol. 85, no. 3, pp. 301–319, 2007.
[15] H. Adab, K. D. Kanniah, K. Solaimani, and K. P. Tan, "Estimating atmospheric humidity using MODIS cloud-free data in a temperate humid region," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 1827–1830.
[16] C. C. Kuo, D. H. Staelin, and P. W. Rosenkranz, "Statistical iterative scheme for estimating atmospheric relative humidity profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 2, pp. 254–260, Mar. 1994.
[17] L.-C. Chen and A. A. Bradley, "Adequacy of using surface humidity to estimate atmospheric moisture availability for probable maximum precipitation," *Water Resour. Res.*, vol. 42, no. 9, 2006, Art. no. W09410.
[18] V. Klaus, L. Bianco, C. Gaffard, M. Matabuena, and T. J. Hewison, "Combining UHF radar wind profiler and microwave radiometer for the estimation of atmospheric humidity profiles," *Meteorologische Zeitschrift*, vol. 15, no. 1, pp. 87–98, 2006.
[19] E. Eccel, "Estimating air humidity from temperature and precipitation measures for modelling applications," *Meteorol. Appl.*, vol. 19, no. 1, pp. 118–128, 2012.
[20] L. Bianco, D. Cimini, F. S. Marzano, and R. Ware, "Combining microwave radiometer and wind profiler radar measurements for high-resolution atmospheric humidity profiling," *J. Atmospheric Ocean. Technol.*, vol. 22, no. 7, pp. 949–965, 2005.
[21] J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to infer unresolved scale parametrization," *Philos. Trans. Roy. Soc. A*, vol. 379, no. 2194, 2021, Art. no. 20200086.
[22] H. Olafsson and J.-W. Bao, *Uncertainties in Numerical Weather Prediction*. Amsterdam, The Netherlands, Elsevier, 2020.
[23] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: Advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016.
[24] A. E. Hassanien, M. Tolba, and A. T. Azar, "Advanced machine learning technologies and applications," in *Proc. 2nd Int. Conf.*, 2014, pp. 248–257.
[25] M. N. Amar, H. Ouaer, and M. A. Ghriga, "Robust smart schemes for modeling carbon dioxide uptake in metal-organic frameworks," *Fuel*, vol. 311, 2022, Art. no. 122545.
[26] C. S. W. Ng, H. Djema, M. N. Amar, and A. J. Ghahfarokhi, "Modeling interfacial tension of the hydrogen-brine system using robust machine learning techniques: Implication for underground hydrogen storage," *Int. J. Hydrogen Energy*, vol. 47, no. 93, pp. 39595–39 605, 2022.
[27] M. N. Amar, M. Shateri, A. Hemmati-Sarapardeh, and A. Alamatsaz, "Modeling oil-brine interfacial tension at high pressure and high salinity conditions," *J. Petroleum Sci. Eng.*, vol. 183, 2019, Art. no. 106413.
[28] M. Talebkeikhah et al., "Experimental measurement and compositional modeling of crude oil viscosity at reservoir conditions," *J. Taiwan Inst. Chem. Engineers*, vol. 109, pp. 35–50, 2020.
[29] M. N. Amar, A. J. Ghahfarokhi, C. S. W. Ng, and N. Zeraibi, "Optimization of wag in real geological field using rigorous soft computing techniques and nature-inspired algorithms," *J. Petroleum Sci. Eng.*, vol. 206, 2021, Art. no. 109038.
[30] M. N. Amar, A. J. Ghahfarokhi, and C. S. W. Ng, "Predicting wax deposition using robust machine learning techniques," *Petroleum*, vol. 8, no. 2, pp. 167–173, 2022.
[31] S. Javeed, K. S. Alimgeer, W. Javed, M. Atif, and M. Uddin, "A modified artificial neural network based prediction technique for tropospheric radio refractivity," *Plos one*, vol. 13, no. 3, 2018, Art. no. e0192069.
[32] M. Holmstrom, D. Liu, and C. Vo, "Machine learning applied to weather forecasting," *Meteorol. Appl.*, vol. 10, pp. 1–5, 2016.
[33] Q. Zhao, Y. Liu, W. Yao, and Y. Yao, "Hourly rainfall forecast model using supervised learning algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4100509.
[34] S. Liu, Y. Yin, Z. Chu, and S. An, "CDL: A cloud detection algorithm over land for MWHS-2 based on the gradient boosting decision tree," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4542–4549, 2020.
[35] X. Chu, W. Bai, Y. Sun, W. Li, C. Liu, and H. Song, "A machine learning-based method for wind fields forecasting utilizing GNSS radio occultation data," *IEEE Access*, vol. 10, pp. 30258–30273, 2022.
[36] H. Park, J. Lee, C. Yoo, S. Sim, and J. Im, "Estimation of spatially continuous near-surface relative humidity over Japan and South Korea," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8614–8626, 2021.
[37] "Highlight of Hong Kong climate," 2023. [Online]. Available. https://www.hko.gov.hk/en/cis/climat.htm
[38] L. Breiman, *Classification and Regression Trees.* London, U.K.: Routledge, 2017.
[39] A. Swetapadma and A. Yadav, "A novel decision tree regression-based fault distance estimation scheme for transmission lines," *IEEE Trans. Power Del.*, vol. 32, no. 1, pp. 234–245, Feb. 2017.
[40] W. I. D. Mining, "Data mining: Concepts and techniques," *Morgan Kaufinann*, vol. 10, pp. 559–569, 2006.
[41] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.
[42] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.
[43] Y. Freund et al., "*Experiments With a New Boosting Algorithm*," in *Proc. Int. Conf. Mach. Learn.*, 1996, pp. 148–156.
[44] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Animal Ecol.*, vol. 77, no. 4, pp. 802–813, 2008.
[45] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.
[46] A. C. Cameron and F. A. Windmeijer, "An r-squared measure of goodness of fit for some common nonlinear regression models," *J. Econometrics*, vol. 77, no. 2, pp. 329–342, 1997.
[47] S. K. Kwak and J. H. Kim, "Statistical data preparation: Management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, pp. 407–411, 2017.
[48] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, "Missing data in clinical research: A tutorial on multiple imputation," *Can. J. Cardiol.*, vol. 37, no. 9, pp. 1322–1331, 2021.
[49] C. Loader, "Smoothing: Local regression techniques," in *Handbook of Computational Statistics: Concepts and Methods*. Berlin, Germany: Springer, 2012, pp. 571–596.
[50] G. R. Arce, *Nonlinear Signal Processing: A Statistical Approach*. Hoboken, NJ, USA: Wiley, 2005.
[51] J. S. Simonoff, *Smoothing Methods in Statistics*. New York, NY, USA: Springer Science & Business Media, 2012.



Anas Amaireh received the bachelor's and master's degrees in communication engineering from Yarmouk University, Jordan, in 2015 and 2018, respectively. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Oklahoma Norman, Norman, OK, USA.

He is currently a Research and Teaching Assistant with the Advanced Radar Research Center, University of Oklahoma. He specializes in radar, machine learning in radar, radar signal processing, 5G communication, antenna arrays, and metaheuristic algorithms. His research focuses on leveraging machine learning to enhance the functionality and performance of modern radar and communication systems.

Yan (Rockee) Zhang (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Nebraska, Lincoln, NE, USA, in 2004.

He is a Presidential Professor with the School of Electrical and Computer Engineering, University of Oklahoma (OU). He was one of the founding faculty members of OU's Advanced Radar Research Center (ARRC), the Technical Lead of the ARRC's multi-functional phased array radar development from 2008 to 2011, and is currently the Faculty Leader of the Intelligent Aerospace Radar Team. He is the Principal Investigator for the OU's Polarimetric Airborne Radar Operating at X-Band developments and leading the deployment missions related to radar and radios supporting Advanced Air Mobility, autonomous vehicle systems, and air-surveillance services. He is one of the faculty fellow members of the Cooperative Institute for Severe and High Impact Weather Research and Operations at OU and a representative of OU at multiple RTCA special committees supporting industry standard developments for avionics.

P. W. Chan received the Ph.D. degree from City University of Hong Kong, Hong Kong, with the study on wind engineering and low level wind shear for aviation applications in 2023.

He is the Director of the Hong Kong Observatory. He had been working at the Hong Kong International Airport for over 20 years, with research and operational efforts in airport meteorological instrumentation, low level windshear and turbulence alerting, and high resolution numerical weather prediction. He is a Visiting Professor of a number of universities in mainland China and an Adjunct Associate Professor with the University of Hong Kong. He has published more than 350 papers in SCIE journals, with a significant portion of the papers focusing on the applications at the airport. He is the Vice Chair of a steering group on meteorological measurements at the World Meteorological Organization and the Chair of meteorology subgroup in Asia Pacific region of the International Civil Aviation Organization.

Dr. Chan is a fellow of the Royal Meteorological Society and a Chartered Meteorologist.