Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.210108

$PM_{2.5}$ Forecast System by Using Machine Learning and WRF Model, A Case Study: Ho Chi Minh City, Vietnam

Vo Thi Tam Minh[1,2*], Tran Trung Tin[3,2], To Thi Hien[1,2]

[1] Faculty of Environment, University of Science, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Faculty of Applied Science, University of Technology, Ho Chi Minh City, Vietnam

ABSTRACT

Predicting has necessary implications as part of air pollution alerts and the air quality management system. In recent years, air quality studies and observations in Vietnam have shown that pollution is increasing, especially the concentration of $PM_{2.5}$. There are warnings about excessively high concentrations of $PM_{2.5}$ in the two major cities of Vietnam as Ho Chi Minh City and Hanoi. Projections for $PM_{2.5}$ concentrations in these cities will provide short-term predictive data on air quality. Using the WRF model to forecast $PM_{2.5}$ in Ho Chi Minh City is new research for providing forecast information on air pollution. Experiments with six machine learning algorithms show that the Extra Trees Regression model gives the best forecast with statistical evaluation indicators including RMSE = 7.68 $\mu g\ m^{-3}$, MAE = 5.38 $\mu g\ m^{-3}$, R-squared = 0.68, and the confusion matrix accuracy of 74%. The experimental setting of the Extra Trees Regression algorithm to predict $PM_{2.5}$ for the next two days with WRF's simulated meteorological data compared with the forecast with observed data showing high accuracy of over 80%. The results show that machine learning with the WRF model can predict $PM_{2.5}$ concentration, suitable for early warning of pollution and information provision for air quality management system in large cities as Ho Chi Minh City.

Keywords: Machine learning, Extra Trees Regression, WRF, Predict $PM_{2.5}$, Ho Chi Minh City

OPEN ACCESS

Received: June 6, 2021
Revised: September 7, 2021
Accepted: October 28, 2021

*** Corresponding Author:**
vttminh@hcmus.edu.vn

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Air pollution is a threat to sustainable development for countries and communities worldwide in general and Asia in particular (Marsden *et al.*, 2019; World Bank and Institute for Health Metrics and Evaluation, 2016). Air pollution is defined as the presence of the atmosphere (indoor and outdoor) of one or more contaminants, such as particulate matter, fumes, gas, mist, odor, smoke, or vapor in quantities, of characteristics, and duration, affecting the living creature (Hesketh, 1979). Human exposure to particulate matter, especially $PM_{2.5}$, can cause respiratory diseases, entry into the circulatory system, and even the brain (Croft *et al.*, 2018; Zhang *et al.*, 2018). According to statistics in the study by Shaddick *et al.* (2020), more than 99% of the population was exposed to $PM_{2.5}$ concentrations higher than the current WHO Air Quality Guidelines (annual average of 10 $\mu g\ m^{-3}$) in Central Asia, South Asia, East Asia, and Southeast Asia (Shaddick *et al.*, 2020). The average annual $PM_{2.5}$ concentration estimated for Asia ranges from 16 to 58 $\mu g\ m^{-3}$, surpassing Europe, North America, and Oceania (Crippa *et al.*, 2019). The recent increase in $PM_{2.5}$ concentrations in Asia is due to the region's high urbanization rate and developing economies, especially emerging economies in Southeast Asia (Yang *et al.*, 2018). The source of $PM_{2.5}$ pollution in Southeast Asian countries is mainly from residential activities, industrial production, electricity industry and biomass burning (Amnuaylojaroen *et al.*, 2020). In large cities, especially megacities, these emission sources are most concentrated in industrial and residential activities. The megacities of China, India, and Southeast Asia with high population density are directly proportional to high concentrations of $PM_{2.5}$ (Zhang *et al.*, 2020). Ambient in the large cities seriously degraded is a fact of life. $PM_{2.5}$ pollution in a megacity is a leading risk factor for the population exposed to concentration exceeding the World Health Organization (WHO) air quality guideline (Krzyzanowski *et al.*, 2014; Marlier *et al.*, 2016). Therefore, one issue in these cities is implementing pollution assessment objectives to support legislation to prevent air pollution. Predict $PM_{2.5}$ concentration is also one of the action plans to reduce and limit polluting activities. The predicted results will show future $PM_{2.5}$ concentrations for response planning or measures to prevent emissions increase. At the same time, the first direct benefit is forecasting people about pollution levels at any time, day, week, or month.

Ho Chi Minh City (HCM City) is almost a megacity with a 2,095 km^2 area and 9.04 million people in 2019 with the rapid growth in the industry, exports, tourism, and services (General Statistics Office of Vietnam, 2020; Ho Chi Minh City Statistical Office, 2020; Krzyzanowski *et al.*, 2014). This development has led to many environmental problems in which air pollution is one of the severe consequences. The air quality issue in HCM City is getting worse and worse, especially the increase of $PM_{2.5}$ concentration. The $PM_{2.5}$ annual mean concentration in HCM City varied 36.3 ± 13.7 $\mu g\ m^{-3}$ (during March 2017–March 2018), exceeds the limit of the Vietnamese standard (25 $\mu g\ m^{-3}$) and is considerably higher than the WHO's guideline for the $PM_{2.5}$ level of 10 $\mu g\ m^{-3}$ (Hien *et al.*, 2019). In another research, the annual average concentrations of $PM_{2.5}$ modeled with air emission inventory data in 2019 were 23 $\mu g\ m^{-3}$, which is higher about 2.3 times than the WHO guideline (Vu *et al.*, 2020). Exposure to high-level concentrations of $PM_{2.5}$ has a substantial impact on health and increases the risk of admission among young children in HCM City (Luong *et al.*, 2020). The increase in $PM_{2.5}$ concentrations and its human effects presents an urgent problem that needs to solve in HCM City. Therefore, forecasting the level of $PM_{2.5}$ pollution in order to prevent or reduce the exposure risk is necessary to be done.

Several pieces of research work on air quality prediction (such as $PM_{2.5}$, PM_{10}, Ozone) using machine learning models, combining meteorological and emission data, have been published (Du *et al.*, 2019; Zhai and Chen, 2018; Zhang and Ding, 2017). Delavar *et al.* (2019) propose an efficient machine learning model to predict PM_{10} and $PM_{2.5}$ pollution in Tehran City (Iran). Using machine learning regression models to predict $PM_{2.5}$ with emissions data in Taiwan showed the expected result value and the actual value are similar (Doreswamy *et al.*, 2020). In addition to the meteorological and $PM_{2.5}$ emissions data, the Aerosol Optical Depth (AOD) data can also predict $PM_{2.5}$ by the machine learning method (Zamani Joharestani *et al.*, 2019). These predictive results are also significant in health studies due to the impact of $PM_{2.5}$ concentrations (Lary *et al.*, 2015).

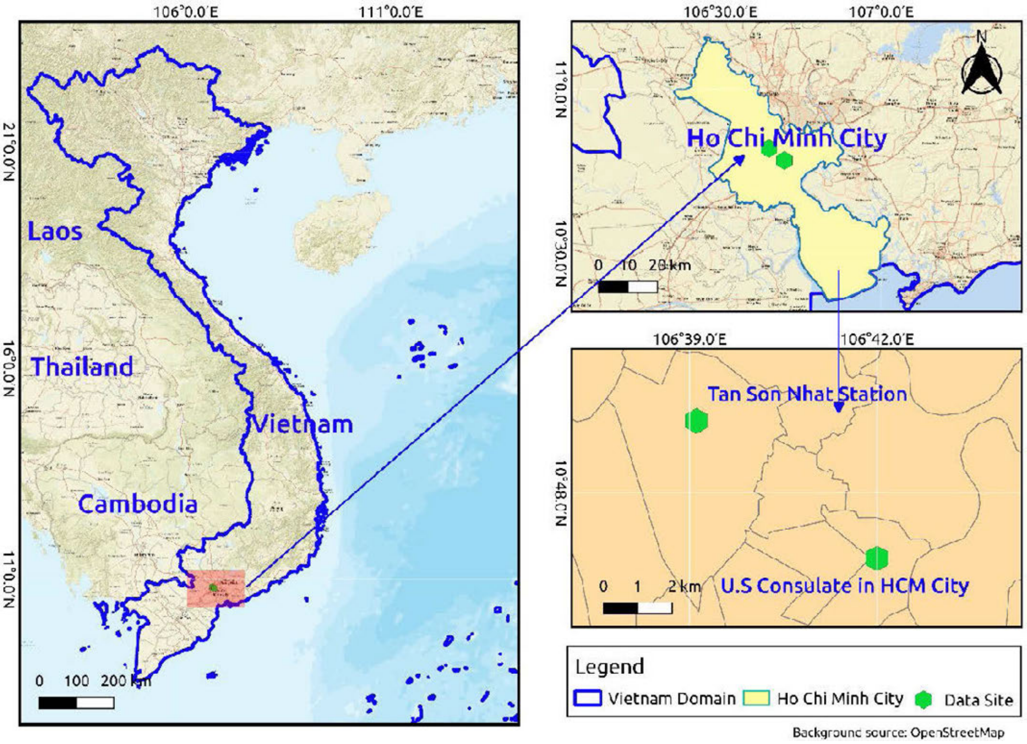
In HCM City, reports on $PM_{2.5}$ are mainly results of actual monitoring; very few studies predict $PM_{2.5}$ concentrations. There are currently very few studies predicting $PM_{2.5}$ concentrations. Information to warn people about $PM_{2.5}$ pollution is also scarce, and there is no official and reliable information channel.

This study will propose a simple, fast, and accurate machine learning method to predict $PM_{2.5}$ concentration with the case study in HCM City. The research results are new to be able to predict $PM_{2.5}$ pollution in HCM City as well as in Vietnam. This result is the first step for the following projects on pollution forecasting, building an application to warn people, and proposing options to reduce HCM City pollution. According to the evaluation of existing studies in Vietnam, this will be a new study using machine learning and the WRF model to predict pollution in HCM City.

2 DATA AND METHODS

2.1 Study Area and Data Set

HCM City's geographic coordinates are between 10.375833–11.371389 North latitude and 106.017222–107.019444 East longitude. The case study area is the center of HCM City (10.782773°N, 106.700035°E) where the $PM_{2.5}$ monitoring station is placed (Fig. 1). The dataset used in this study includes $PM_{2.5}$ concentrations and meteorological data. Meteorological data comprise temperature, relative humidity, wind direction, and wind speed. Meteorological data is collected from Tan Son Nhat Station—HCM City (representing the city) at the NOAA National Centers for Environmental Information website (NOAA National Centers for Environmental Information, 2001). In addition,



Background source: OpenStreetMap

Fig. 1. Study area and data sites.

hourly averaged $PM_{2.5}$ data were obtained from the monitoring station of the U.S. Consulate in HCM City, and these data are available on https://airnow.gov (AIRNow Program (U.S), 2000). Data used for the machine learning experiment are hourly data for five years, from January 1st, 2016, to December 31st, 2019, for training and from January 1st, 2020, to December 31st, 2020, for the testing model.

2.2 WRF Model Simulations

The Weather Research and Forecasting (WRF) model is developed by many meteorological research and forecasting centers in the United States, such as the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (NOAA), National Center for Environmental Prediction (NCEP), and the participation of a team of scientists working at many universities (Skamarock *et al.*, 2008). WRF is used for research and numerical weather prediction purposes with two versions: Advanced Research (ARW) and Non-hydrostatic Meso Model (NMM) (Janjic, 2003). The WRF-ARW version was developed by the US National Center for Atmospheric Research (NCAR) (Skamarock *et al.*, 2008). With wide application, WRF has become a community model, bringing benefits and the contribution of the number of users in the world.

In this study, the WRF model version 4.1.1 with ARW dynamic core (Skamarock *et al.*, 2019) is used for simulating the meteorological prediction data. The model configuration with three nesting domains are shown in Fig. 2. Domain 3, with a 1.0 km grid resolution covering the HCM City area (163 × 163 grid points), is nested with a 3.0 km domain covering southern Vietnam (domain 2, 193 × 175 grid points) and the 9.0 km resolution largest domain whole Vietnam (domain 1, 100 × 184 grid points).

All three domains have the set up with the same dynamic and physical parameterization configurations. Microphysics parameterization uses a 3-class WRF Single-Moment (WSM) scheme (Kessler, 1969). Other model physics options include the Rapid Radiative Transfer Model (RRTM) for the long-wave radiation parameter (Mlawer *et al.*, 1997), the Dudhia scheme for short-wave radiation (Dudhia, 1989), the Kain-Fritsch (new Eta) scheme (Kain, 2004) for cumulus parameterizations. The schemes for the surface and boundary layers set up in the model are surface layer using Monin-Obukhov Similarity (Monin and Obukhov, 1954; Obukhov, 1946) schemes; land surface set by Noah Land-Surface Model (Chen and Dudhia, 2001), Yonsei University (YSU) for boundary

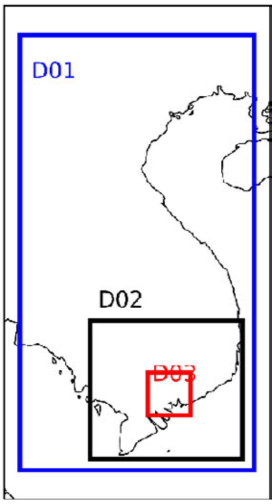


Fig. 2. WRF model domains.

Table 1. WRF model's configurations.

Parameterizations	Schemes
Microphysics	WRF Single-Moment (WSM 3-class)
Longwave Radiation	RRTM
Shortwave Radiation	Dudhia
Surface Layer	Monin-Obukhov Similarity
Land Surface	Noah Land-Surface Model
Boundary Layer	YSU
Cumulus	Kain-Fritsch (new Eta)
Dynamics	PBL scheme (2D)

layer parameterizations (Hong *et al.*, 2006). Table 1 presents a brief of configuration parameter schemes for all three domains in the WRF model.

Global Forecast System (GFS) data is used as a boundary condition and initial data to the WRF model to run meteorological simulations and predictions (NOAA National Centers for Environmental Prediction (NCEP, 2011). GFS's weather forecasting model generates datasets with many atmospheric and land-soil variables, including temperatures, winds, precipitation, soil moisture, and atmospheric ozone concentration. The GFS forecast data has a resolution of 0.5 degrees (about 55 km horizontal resolution), and the forecast steps out to 192 hours (8 days). This dataset is run four times daily at 00z, 06z, 12z, and 18z with a 3-hour temporal resolution (3-hourly).

The distribution of $PM_{2.5}$ concentrations tends to differ according to weather and atmospheric conditions, so there is a seasonal variation. A simulation of the historical period was performed to evaluate the WRF model's applicability to generate meteorological data for the $PM_{2.5}$ predictive machine learning model. Meteorological simulation data for 2 months representing the rainy season (September 2020) and dry season (January 2021) are used to run a machine learning model to predict $PM_{2.5}$ and compare the results run with actual monitoring data at Tan Son Nhat station. The WRF model then runs a forecast of future meteorological data (April and May 2021) to input a machine learning model that predicts $PM_{2.5}$ concentrations. Predictions are made with short-term (24 hours, 48 hours, 72 hours) and long-term (7 days) respectively for analysis and evaluation.

2.3 Machine Learning Algorithms

Machine learning models are run with several different algorithms to evaluate the efficiency and choose the best predictive algorithm. In this study, the machine learning algorithms are run in Python version 3.7 (Van Rossum and Drake, 2009) with the Scikit-Learn library (Pedregosa *et al.*, 2011). The machine learning algorithms are Extremely Randomized Trees Regression (Extra

Trees Regression) (Geurts *et al.*, 2006), Linear Regression (Pires *et al.*, 2008), Extreme Gradient Boosting (XGBoost) (Chen *et al.*, 2019; Zhai and Chen, 2018; Zhang *et al.*, 2018), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Neural Network Regression (Choubin *et al.*, 2016), and Decision Forest Regression (Criminisi, 2011). Furthermore, these six machine learning algorithms are all regression models, which means that the results give a specific predictive value.

In this study, the input data of the $PM_{2.5}$ predictive machine learning model includes meteorological data (temperature, humidity, wind direction, and wind speed) and $PM_{2.5}$ concentration data. The dataset used for the machine learning model is five years, divided into two parts, with one part having four years for the training period and the other part (one year) for the testing period.

The flow diagram in Fig. 3 shows the steps for running a machine learning model with six algorithms. The first step is to separate the preprocessing data into two data sets, including the training and test sets. Second, train the machine learning model with each algorithm by the training dataset. The next step is to place the test set to check the training efficiency of each algorithm. The final step is to evaluate the performance of each model through the evaluation parameters (details in section 2.4).

2.4 Evaluate the Machine Learning Model

2.4.1 Statistical indices-based evaluations

The statistical indices used to evaluate models include the coefficient of determination (R^2), root-mean-square-error (RMSE), mean-absolute-percentage-error (MAPE), and mean-absolute-error (MAE). R^2 is the measure of the variance in the observation variable that can be predicted using the predictor variable. R^2 for machine learning models with one independent variable can be calculated as below:

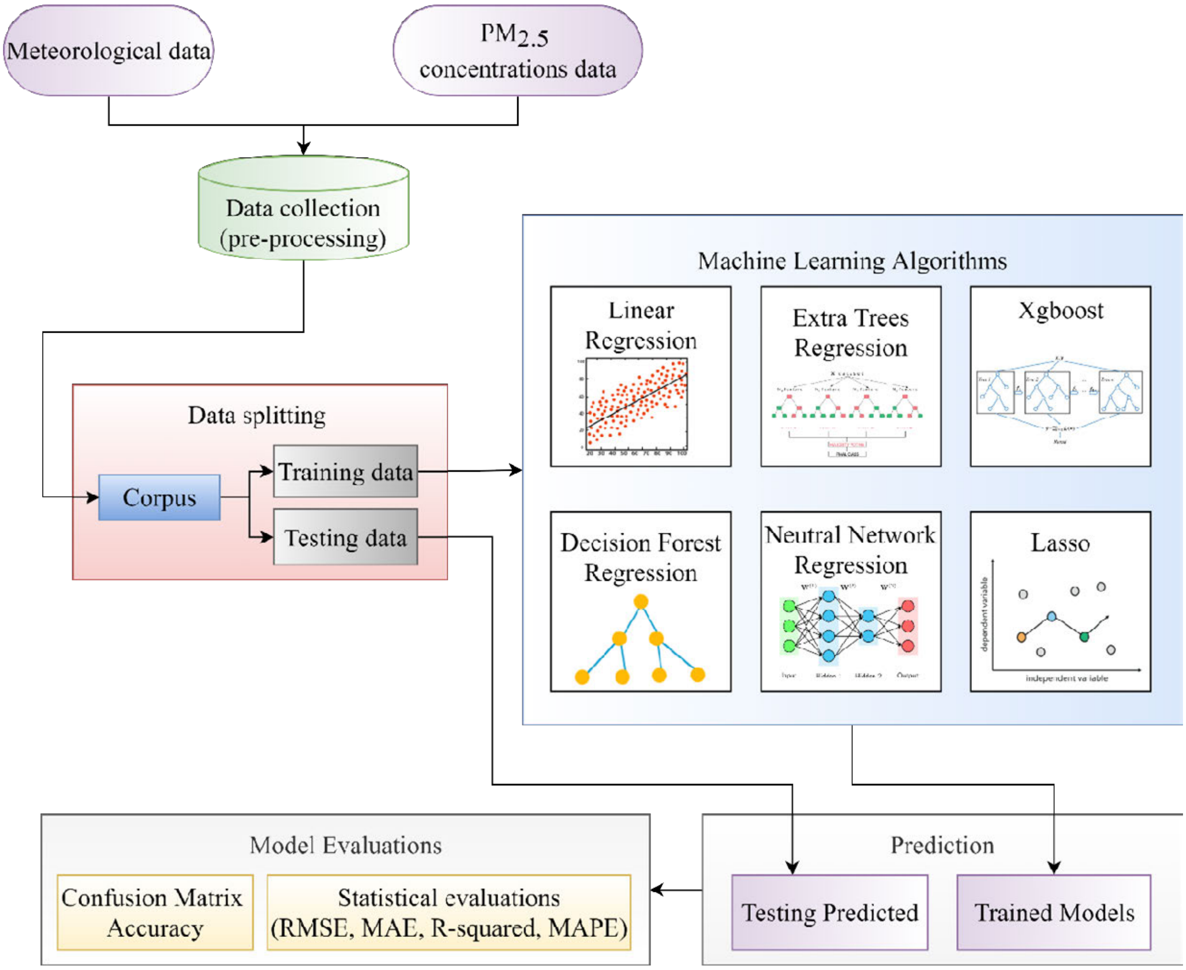

Meteorological data
PM2.5 concentrations data
Data collection (pre-processing)
Data splitting
Corpus
Training data
Testing data
Machine Learning Algorithms
Linear Regression
Extra Trees Regression
Xgboost
Decision Forest Regression
Neutral Network Regression
Lasso
Model Evaluations
Confusion Matrix Accuracy
Statistical evaluations (RMSE, MAE, R-squared, MAPE)
Prediction
Testing Predicted
Trained Models

Fig. 3. Machine learning flow diagram using six algorithms.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{1}$$

Root Mean Square Error (RMSE) is the average magnitude of the error. RMSE tells how concentration the data is around the line of best fit. RMSE is commonly used as a standard statistical metric to measure model performance or predict in meteorology, air quality, and climate research studies (Chai and Draxler, 2014). In this research, RSME is used for the machine learning model to predict the daily $PM_{2.5}$ concentrations. The formula is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{2}$$

Mean Absolute Error (MAE) is another helpful measure widely used in machine learning model evaluations. There are comparisons between RMSE and MAE for reliability in assessing the model through these two types of errors. MAE is shown to be an unbiased estimator, while RMSE is a biased estimator. MAE also has a lower sample variance compared with RMSE (Brassington, 2017). Therefore, dimensioned evaluations and inter-comparisons of average model-performance error should be based on MAE (Willmott and Matsuura, 2005). MAE is calculated according to:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{3}$$

The Mean Absolute Percentage Error (MAPE) expresses forecasting errors as ratios, and they are dimensionless and easy to interpret. MAPE is the mean or average of the absolute percentage errors of forecasts and is defined as actual or observed value minus the forecasted value (Swamidass, 2000). MAPE in the machine learning model is the most common measure used to predict errors and finding the best model (de Myttenaere *et al.*, 2016). The following formula:

$$\text{MAPE} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right|}\right)100 \tag{4}$$

Eqs. (1)–(4) where y, observation value; $\overline{y}$, mean of observation values; $\hat{y}$, predicted values; *i*, i^{th} observation.

2.4.2 Confusion matrix

The statistical indicators evaluate the predictive performance of $PM_{2.5}$ on specific forecast numbers, so the errors may be too large to assess the model's undertaking properly. Besides, the observed data is the hourly average, so the error of the forecasting model can be much higher than the reality. A confusion matrix is an excellent option for reporting results in the performance of a classification model because it is possible to observe the relations between the classifier outputs and the true ones (Diez, 2018). The information in the confusion matrix can be used to determine the accuracy of the predictive model. This study can optimize the model evaluation by a confusion matrix based on the U.S. Environmental Protection Agency (EPA) scale rating for the health impacts of $PM_{2.5}$ (U.S. EPA, 2018). According to EPA's table, the breakpoint scale of $PM_{2.5}$ in $\mu g\ m^{-3}$ is classified into seven categories (Table 2). The forecast results are classified based on the U.S. EPA's $PM_{2.5}$ breakpoint and evaluated with the observed data through a confusion matrix.

In this study, the calculation and presentation of the confusion matrix to evaluate the performance for two cases, including the selected predictive machine learning model and the results of running the model with various types of meteorological data. For the machine learning model: use the confusion matrix to evaluate the predictive results of the testing period. From the results of the confusion matrix analysis, evaluate the model performance. For the results of running the model

Table 2. $PM_{2.5}$ breakpoints for the AQI.

AQI Category	Low Breakpoint $PM_{2.5}$ ($\mu g\ m^{-3}$)	High Breakpoint $PM_{2.5}$ ($\mu g\ m^{-3}$)
Good	0	12
Moderate	12.1	35.4
Unhealthy for sensitive	35.5	55.4
Unhealthy	55.5	150.4
Very unhealthy	150.5	250.4
Hazardous	250.5	350.4
Hazardous	350.5	500.4

with two meteorological data sets: The matrices confuse the observed value with the forecast from the model using two types of meteorological data (observation and simulation by WRF), respectively. Analyze the confusion matrix and conclude the model's effectiveness when using the input meteorological data simulated by the WRF model.

3 RESULTS AND DISCUSSION

3.1 Machine Learning Model Evaluation

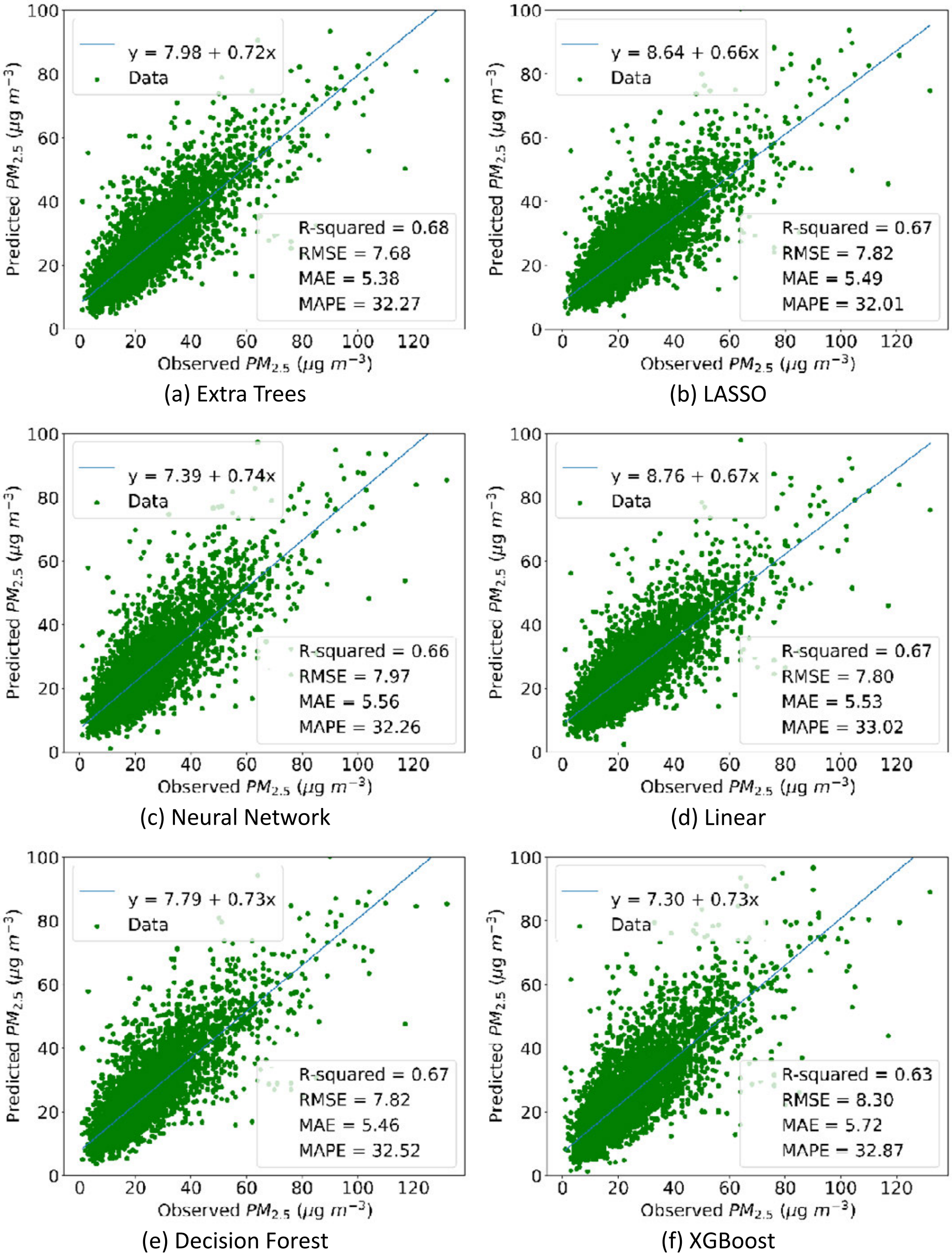
The hourly meteorological and $PM_{2.5}$ data for four years (from January 1st, 2016 to December 31st, 2019) covering 1460 days (24047 standardized samples) are taken as a training set and leaving 6303 samples for 365 days (from January 1st, 2020 to December 31st, 2020) is used as testing data. Compare the prediction results of the six models with the test data set with observations to evaluate the model's performance. The results of the two evaluation methods, including statistical errors and confusion matrix, are presented in the following.

The first is to present the results of statistical evaluation of the performance of machine learning models. Following are the statistical evaluation results for the performance of the machine learning models. The performance of the different models is listed in Table 3, and a scatter chart of predicted and observed results for each model is shown in Fig. 4. The blue line indicates the fitted simple regression line on scattering points for models (Figs. 4(a)–4(f)). The slope of the regression equation in all models is positive, less than 1, and the residual interval is positive. First, with the positive slope value less than 1, the model will give lower forecast results than the actual observed, especially at the points where the higher concentration is suddenly. The study of Gupta and Christopher (2009) also concluded that the regression equation of the machine learning model has a positive slope value less than 1, the results of the predictive model are lower than those observed at the station (Gupta and Christopher, 2009). Next, positive model residuals also indicate that the model's prediction tends to be lower than the observed value (Kiernan, 2014). The scatter becomes sporadic when the observed or predicted values are overestimated or underestimated. These results explain the trend of models' prediction: lower for high observed values and higher for low observed values. This explanation is very consistent with the future forecast results when the concentration of $PM_{2.5}$ increases suddenly at a particular time (see section 3.2).

R^2 always range between 0 and 1 and is the direct indicator in term of model performance. In Table 3, the results R^2 of the six models reach the value from 0.63 to 0.68, which means that these models all have just fine enough performance. The highest R^2 value earned 0.68 is Extra Trees Regression which means that the $PM_{2.5}$ forecast efficiency of this algorithm is 68%. Models are then considered for RMSE, which is better with a lower RMSE. Extra Trees Regression is also

Table 3. The performance of each model.

	Linear Regression	Neural Network Regression	LASSO	Extra Trees Regression	Decision Forest Regression	XGBoost
R^2	0.67	0.66	0.67	0.68	0.67	0.63
RMSE	7.80	7.97	7.82	7.68	7.82	8.30
MAE	5.53	5.56	5.49	5.38	5.46	5.72
MAPE	33.02	32.26	32.01	32.27	32.52	32.87



(a) Extra Trees

(b) LASSO

(c) Neural Network

(d) Linear

(e) Decision Forest

(f) XGBoost

Fig. 4. Scatter and fitted plots of predicted and observed values of different models.

the model with the lowest RMSE (RMSE = 7.68 µg m^{-3}), which means it gives better performance than others. In addition to the RMSE, the MAE is also a popular metric for evaluating the accuracy of continuous variables. Both errors are negatively-oriented scores, meaning the lower values are, the better. In MAE results, the lower MAE value of Extra Trees Regression (MAE = 5.38 µg m^{-3}) means that this is better than the rest of the models.

This study aims to be able to predict $PM_{2.5}$ concentrations with a minimum of RMSE error. The expected value of the Extra Trees Regression for the slightest error of RMSE is 7.68 µg m^{-3}. Compared with some results in other studies, such as Karimian *et al.* (2019) using LTSM (long short-term memory) model for predictive model results $PM_{2.5}$ with RMSE error of 8.91 µg m^{-3}, R^2 = 0.8. Another study by Brent Lagesse *et al.* (2020) using some algorithms, including the LASSO model predicting indoor $PM_{2.5}$, gives better results with a small RMSE of 2.65 µg m^{-3} and R^2 of 0.65.

However, with the prediction of $PM_{2.5}$ indoors, a small error is achieved because indoor ambient air conditions are more accessible to control than the experiment in our study with continuously changing meteorological parameters. In the study of Karimian *et al.* (2019) higher R^2 performance is possible due to the LTSM model used. The results of R^2 and errors of RMSE, MAE obtained by the Extra Trees Regression model in this study are relative and acceptable to predict $PM_{2.5}$. In summary, the Extra Trees Regression model with the best statistical evaluation results is the model chosen to perform future predictions.

The results of the evaluation of six machine learning models with regression equations and statistical errors show only quite small differences. There was no significant difference in the forecast results between the models. Notice that all six models give predictions that tend to be close to reality. Therefore Fig. 5 shows the variation of the actual and predicted $PM_{2.5}$ concentrations during the test of a best model—Extra Tree Regression. There is a similarity in concentration variation between actual and forecast over the entire time series of the hourly averaged experimental data set from January 1st, 2020 to January 1st, 2021. That shows the close relationship and accuracy between reality and observation using machine learning model. This predictive trend shows that the predictive ability of machine learning models is relatively good. However, to more accurately evaluate the forecast performance, these results are not enough. It is necessary to further consider the method of model evaluation.

Another method to evaluate the performance of the machine learning model used for this study is the confusion matrix. The confusion matrix analysis is used to further strengthen the reliability in evaluating the predictive efficiency of the selected model—Extra Trees Regression. The confusion matrix with testing data (Fig. 6) is presented for the prediction results of the best model—Extra Trees Regression: where Fig. 6(a) is a matrix with the number of predictive samples and Fig. 6(b) is a normalized matrix. The two axes are the four classifications in the $PM_{2.5}$ breakpoints, the horizontal is the forecast result, the vertical is the observed value, and the diagonal of the matrix is the number of forecast samples matching the number of observed samples. $PM_{2.5}$ pollution in HCM City is reached at the levels of the 4th classification (Unhealthy) of the $PM_{2.5}$ breakpoint. Fig. 6(a) shows 356 correctly predicted samples at $PM_{2.5}$ level 1, which corresponds to an accurate prediction rate of 29.1% (Fig. 6(b)). At $PM_{2.5}$ level 2, the highest sample number and prediction rate were 3796 samples reaching 90.7%, $PM_{2.5}$ level 3, and level 4, which averaged 62.5% and 46.3% rate, respectively. For false predictions, $PM_{2.5}$ breakpoint level 1 has the highest rate of wrong predictions (70.2%), with 859 actual samples being "Good" but being predicted higher at "Moderate". The second highest false prediction rate is the highest level—"Unhealthy", 44.3% the wrong prediction becomes the lower level "Unhealthy for sensitive". This result proves the conclusion in section 3.1 about the model's forecast trend for high and low $PM_{2.5}$ concentrations. At $PM_{2.5}$ breakpoint level 1, the model predicts higher than reality, but the number of samples only accounts for about a quarter of level 2 samples. Besides, predictions higher than reality still have an exemplary meaning in pre-warning for the future. At the "Unhealthy" level, although it has the second wrong prediction rate, the number of samples is relatively small, so it does not affect the accuracy of the model much. The forecast results for these four classification levels have an accuracy of 0.74, which means that the correct prediction rate of the Extra Trees

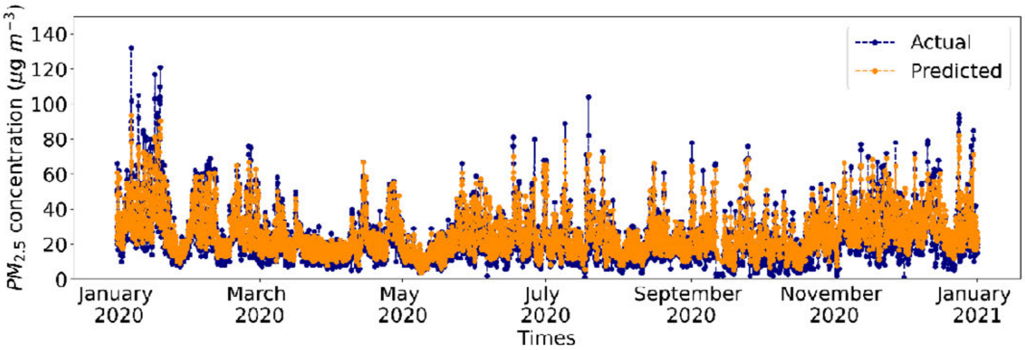


Fig. 5. Actual and predicted $PM_{2.5}$ concentration variation graph for testing period with Extra Trees Regression model.

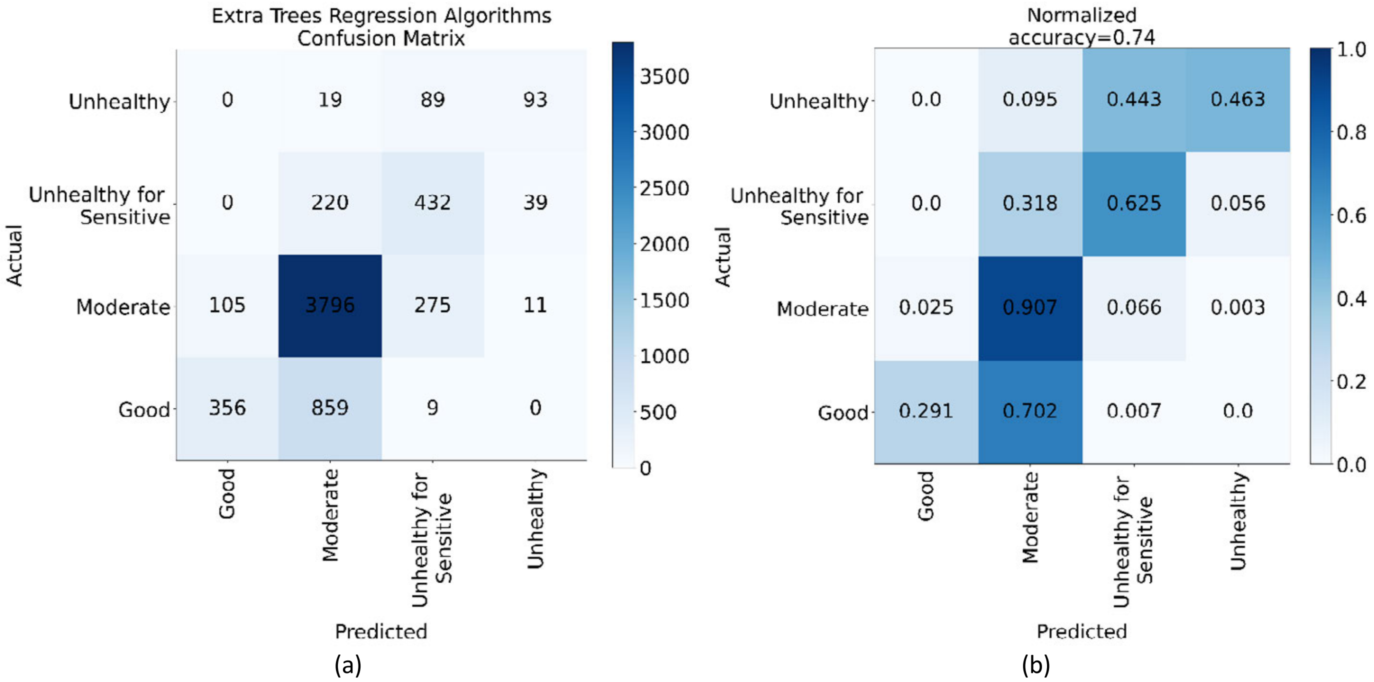


Fig. 6. Extra Trees Regression model's confusion matrix (a) with samples (b) normalized confusion matrix (the horizontal is the forecast result, the vertical is the observed value, and the diagonal of the matrix is the number of predicted equal with observed).

Regression model is 74%. The above confusion matrix analysis results and statistical evaluations show that the Extra Trees Regression model can predict $PM_{2.5}$ concentration reasonably.

This paper focuses on evaluating the predictability of machine learning models, so the model evaluation results show that the Extra Trees Regression model has good predictability and can be applied in specific predictive studies more than in Vietnam. Section 3.2 below presents the verification and accreditation of the Extra Trees Regression machine learning model running with observed meteorological data and simulated by the WRF model for the predictions.

3.2 Machine Learning Predict $PM_{2.5}$ Using WRF Model Meteorological Data

This section presents the $PM_{2.5}$ prediction results by the Extra Trees Regression trained model for the past two periods using two types of meteorological data: 1) observation and 2) forecast by WRF model. The objective is to accreditation the model and to evaluate the prediction results of the machine learning model with the forecast meteorological data from the WRF model. The period selected to run the forecast for the past period is two months in 2 distinct seasons of HCM City, the rainy season and the dry season. Continue, the input data used for 24-hour $PM_{2.5}$ prediction (day-1) is the forecasted meteorological data of that day and all data (including $PM_{2.5}$, temperature, wind direction, wind speed) of the previous five days. Prediction for the next day (day-2) using data from the forecasted day-1 and all data for the previous four days.

Fig. 7 shows the prediction results by the Extra Trees Regression model with observed and simulated meteorological data by WRF of 2 months representing two rainy and dry seasons (September 2020 and January 2021). The blue bars are the observed $PM_{2.5}$ concentration; the green markers are the forecast result with observed meteorological data, and the red colors with the WRF model data. Considering the seasonal trend, $PM_{2.5}$ concentrations on rainy days (Fig. 7(a)) are also often lower than in the dry season due to the deposition of dust particles in the air by rain. In the rainy season, there are times of the day when the concentration of $PM_{2.5}$ spikes, which the reduced convection can explain in the air, the pollution does not diffuse but accumulates in the atmosphere. When there is rain, these accumulated pollution particles are carried by rainwater and deposited, the air becomes cleaner, the concentration of $PM_{2.5}$ decreases again. This result is also similar to the study of Yu *et al.* (2021) when analyzing the correlation between dry deposition and $PM_{2.5}$ concentration. The study of Yu *et al.* (2021) also concluded that reducing convection will reduce the diffusion of substances in the air, so the concentration of $PM_{2.5}$ will

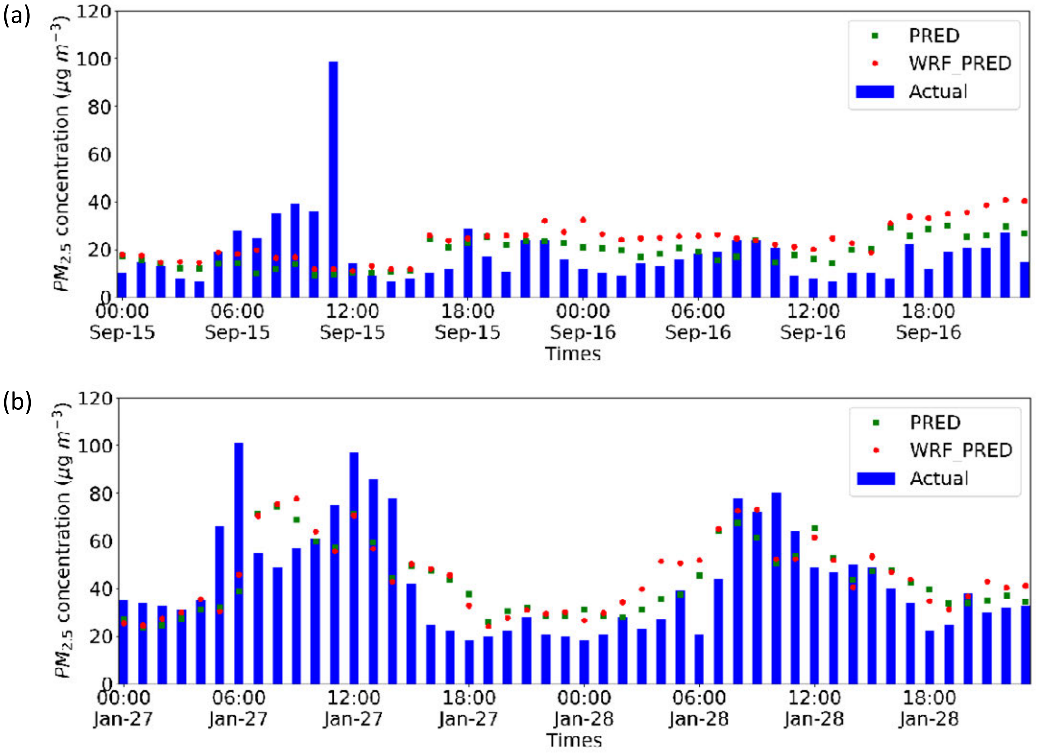


Fig. 7. Actual and predicted $PM_{2.5}$ concentration with observation and simulated meteorological data; (a) in the rainy season (September 2020), (b) in the dry season (January 2021).

increase at these times. At the time of the sudden increase in pollution, the model did not accurately predict; the predicted results were lower than the observed values at these times of high pollution concentration. This prediction result is also consistent with the slope and residual analyzes in the regression equation of machine learning models (see section 3.1). On the dry season days (Fig. 7(b)), the concentration of $PM_{2.5}$ will usually be higher than in the rainy season, and the weather usually has little change; this concentration will not have much change. This result shows that both predictions with two meteorological data sets have a perfect similarity and consistent trends with reality. It is concluded that the forecasting model can give predictions with the input meteorological data simulated by the WRF model.

To be more certain of the above conclusion, the article further analyzes the confusion matrix results normalized for the two $PM_{2.5}$ prediction results and observation. Fig. 8 compares the results of normalized confusion matrices between actual and predicted with two meteorological data sets (in January 2021). The accuracy and the proportion of accurate predictions compared to reality in each class are presented in Figs. 8(a) and 8(b). Accuracy compared to reality is about 60% for both types of meteorological data. The proportion of accurate predictions for classes reached an average of approximately 54.5% to 68.0%. The confusion matrix in Fig. 8(c) shows the accuracy of two prediction results from two meteorological datasets with high accuracy of 85% and a high percentage of correct predictions in all three classifications (respectively. 73.7%; 88.9%, and 100%). This confusion matrix results confirm that the meteorological data predicted by the WRF model is good to use for future forecasts.

3.3 Future Predictions

This section will apply the above analysis results to use the machine learning model and the Extra Trees Regression algorithm for future predictions with no $PM_{2.5}$ and meteorological data observed. First, set up and run a forecast model on May 1st, 2021; at this time, the NCEI website has only updated meteorological data until the end of April 29th, 2021, and the AirNow website updated $PM_{2.5}$ data until 23:00 April 30th, 2021. Next, make a short-term forecast for 72 hours (from April 30th to May 3rd, 2021) and a 7-day *medium-term* forecast (from April 30th to May 7th,

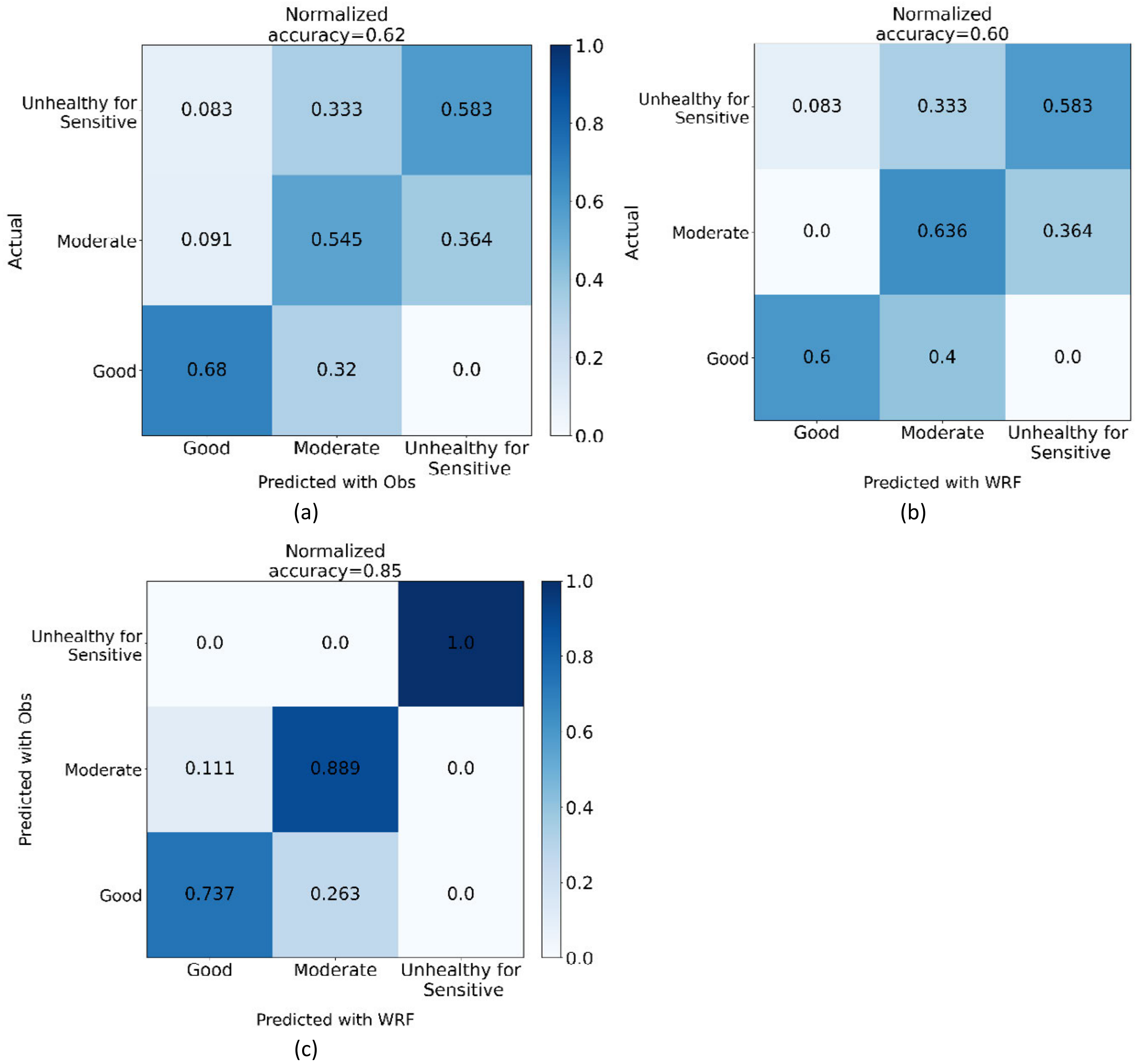


Fig. 8. Normalized confusion matrix (January 2021); (a) Actual and Predicted with meteorological observation data, (b) Actual and Predicted with simulated meteorological data, (c) Predicted with observations and Predicted with WRF simulations.

2021). It is necessary to forecast every 24 hours and then use the result just predicted to generate the next 24-hour forecast. In case the input data for the machine learning model to predict for April 30th, 2021 (first 24 hours) are hourly observation data for the previous five days (from April 26th to April 29th, 2021) and WRF's forecast meteorological data for times when no data is available (April 30th). The following days will take both the predicted results and the meteorological data of the WRF model as input and repeat the forecast every 24 hours. All WRF's forecast meteorological data need for all days with no data is from April 29th to May 7th, 2021.

The result in Fig. 9(a) shows historical $PM_{2.5}$ concentrations (blue line) and forecast results for the next three days (orange line). With the dissemination of $PM_{2.5}$ pollution information, people's concern is not the specific concentration number but just the pollution level. Fig. 9(b) shows the pollution level through the color scale of warning effects on health. The yellow bar is the pollution in the normal range and the orange bar being the warning affecting sensitive people. These results can be visualized on a web platform or a mobile phone application to provide immediate information. However, for the information on these applications to be reliable, it is necessary to build a database system or, in particular, to set up a system of monitoring stations for the

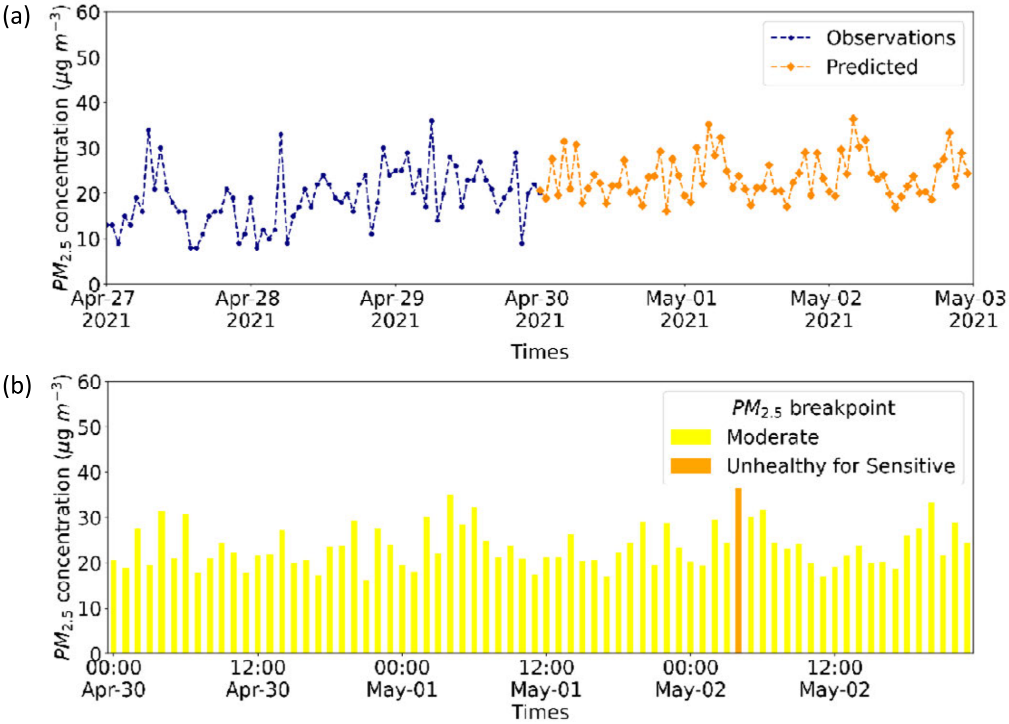


Fig. 9. $PM_{2.5}$ concentration 72 hours predictions (from April 30th to May 3rd, 2021) (a) observations and predicted; (b) $PM_{2.5}$ breakpoint color scale.

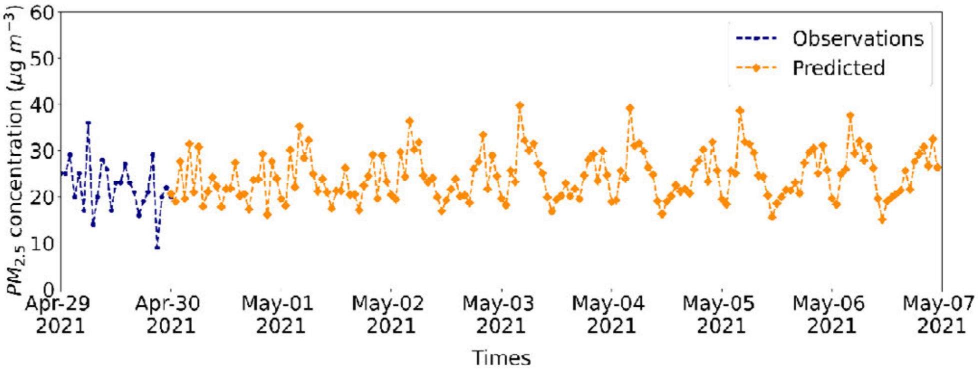


Fig. 10. $PM_{2.5}$ concentration medium-term prediction for seven days.

forecast area. The work required the continuation of other projects longer, but within the framework of this study, a basic model for good prediction was proposed.

The machine learning models can predict for more days in the future; the results of the medium-term forecast for seven days are shown in Fig. 10. WRF model's input data (initial conditions, boundary conditions) used in the study is predictable for 384 hours (16 days). Hence, the machine learning model is predictable for many more days. However, it is necessary to test more closely for these forecasting results to have the basis to correct the model and increase the forecasting performance.

4 CONCLUSIONS

This study initially shows an effective, fast, and accessible machine learning model for predicting future $PM_{2.5}$ concentrations. It is simple to build a stable performance predictive model with a long enough input data set (over three years) and a machine learning single algorithm. Combined with the meteorological data forecasted by the WRF model, the machine learning model can predict $PM_{2.5}$ concentration in short to medium term. The predicted results are hourly $PM_{2.5}$ concentrations. The model also gives the necessary information to the people's concern about $PM_{2.5}$ impact on health (the classified model according to $PM_{2.5}$ breakpoint). The study has achieved some positive results as follows:

- After testing model training with six single algorithms, the best model to predict $PM_{2.5}$ concentration is with the Extra Tree Regression algorithm. The selected model performance evaluation indicators include R-squared = 0.68, and the confusion matrix accuracy is 74%.
- The future $PM_{2.5}$ concentration prediction model can be combined well with meteorological data from the WRF model. The predicted results are similar to those predicted by observed meteorological data.
- The model predicts short-term (48 hours) and medium-term (7 days) $PM_{2.5}$ concentrations in the period from April 30th to May 7th, 2021.

The machine learning algorithms selected in model training are all single algorithms, and the daily data sample frequency makes the predictive model achieve only fundamental optimism. The new research results are the first step to test the possibility of combining machine learning and WRF to forecast $PM_{2.5}$ for the HCMC area. To increase the performance of the machine learning model, we can use data per minute or per second for training and testing to improve the correlation between meteorological data and pollution data. In addition, it is possible to combine machine learning algorithms to increase predictive performance for the model. To increase the performance of the machine learning model, we can use data per minute or per second for training and testing to improve the correlation between meteorological data and pollution data. In addition, it is possible to combine machine learning algorithms to increase predictive performance for the model.

ACKNOWLEDGMENTS

This research is funded by University of Science, VNU-HCM under grant number T2020-23.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.210108

REFERENCES

AIRNow Program (U.S) (2000). AIRNow. [Research Triangle Park], N.C.: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards. United States. https://www.airnow.gov/international/us-embassies-and-consulates/ (accessed 10 April 2020).

Amnuaylojaroen, T., Inkom, J., Janta, R., Surapipith, V. (2020). Long range transport of southeast Asian $PM_{2.5}$ pollution to northern Thailand during high biomass burning episodes. Sustainability 12, 10049. https://doi.org/10.3390/su122310049

Brassington, G. (2017). Mean absolute error and root mean square error: Which is the better metric for assessing model performance? Eur. Geosci. Union Gen. Assem. 19, EGU2017-3574. https://meetingorganizer.copernicus.org/EGU2017/EGU2017-3574.pd (accessed 6 January 2021).

Chai, T., Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, F., Dudhia, J. (2001). Coupling an advanced land surface–Hydrology model with the penn state–NCAR MM5 modeling system. Part II: Preliminary model validation. Mon. Weather Rev. 129, 587–604. https://doi.org/10.1175/1520-0493(2001)129<0587:CAALSH>2.0.CO;2

Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C.H., Liu, R. (2019). XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access 7, 13149–13158. https://doi.org/10.1109/ACCESS.2019.2893448

Choubin, B., Khalighi-Sigaroodi, S., Malekian, A., Kişi, Ö. (2016). Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. Hydrol. Sci. J. 61, 1001–1009. https://doi.org/10.1080/02626667.2014.966721

Criminisi, A. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends Comput. Graph. Vis. 7, 81–227. https://doi.org/10.1561/0600000035

Crippa, M., Janssens-Maenhout, G., Guizzardi, D., Van Dingenen, R., Dentener, F. (2019). Contribution and uncertainty of sectorial and regional emissions to regional and global $PM_{2.5}$ health impacts. Atmos. Chem. Phys. 19, 5165–5186. https://doi.org/10.5194/acp-19-5165-2019

Croft, D.P., Zhang, W., Lin, S., Thurston, S.W., Hopke, P.K., Masiol, M., Squizzato, S., van Wijngaarden, E., Utell, M.J., Rich, D.Q. (2019). The association between respiratory infection and air pollution in the setting of air quality policy and economic change. Ann. Am. Thorac. 16, 321–330. https://doi.org/10.1513/AnnalsATS.201810-691OC

de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F. (2016). Mean absolute percentage error for regression models. Neurocomputing 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Delavar, M., Gholami, A., Shiran, G., Rashidi, Y., Nakhaeizadeh, G., Fedra, K., Hatefi Afshar, S. (2019). A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran. ISPRS Int. J. Geo-Inf. 8, 99. https://doi.org/10.3390/ijgi8020099

Diez, P. (2018). Chapter 1 - Introduction, in: Diez, P. (Ed.), Smart Wheelchairs and Brain-Computer Interfaces, Academic Press, pp. 1–21. https://doi.org/10.1016/B978-0-12-812892-3.00001-7

Doreswamy, Harishkumar, K.S., Yogesh, K.M., Gad, I. (2020). Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. Procedia Comput. Sci. 171, 2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

Du, J., Qiao, F., Yu, L. (2019). Temporal characteristics and forecasting of $PM_{2.5}$ concentration based on historical data in Houston, USA. Resour. Conserv. Recycl. 147, 145–156. https://doi.org/10.1016/j.resconrec.2019.04.024

Dudhia, J. (1989). Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. J. Atmos. Sci. 46, 3077–3107. https://doi.org/10.1175/1520-0469(1989)046%3C3077:NSOCOD%3E2.0.CO;2

General Statistics Office of Vietnam (2020). Completed results of the 2019 Viet Nam population and housing census 2019. Statistical Publishing House, Viet Nam.

Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gupta, P., Christopher, S.A. (2009). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. J. Geophys. Res. 114, D14205. https://doi.org/10.1029/2008JD011496

Hesketh, H.E. (1979). Introduction to Air Pollution, in: Wang, L.K., Pereira, N.C. (Eds.), Air and Noise Pollution Control, Humana Press, Totowa, NJ, pp. 3–39. https://doi.org/10.1007/978-1-4612-6236-7_1

Hien, T.T., Chi, N.D.T., Nguyen, N.T., Vinh, L.X., Takenaka, N., Huy, D.H. (2019). Current status of fine particulate matter ($PM_{2.5}$) in Vietnam's most populous City, Ho Chi Minh City. Aerosol Air Qual. Res. 19, 2239–2251. https://doi.org/10.4209/aaqr.2018.12.0471

Ho Chi Minh City Statistical Office (2020). Statistical Yearbook of Ho Chi Minh City 2019. Ho Chi Minh City General Publishing House, Viet Nam.

Hong, S.Y., Noh, Y., Dudhia, J. (2006). A new vertical diffusion package with an explicit treatment of entrainment processes. Mon. Weather Rev. 134, 2318–2341. https://doi.org/10.1175/MWR3199.1

Janjic, Z.I. (2003). A nonhydrostatic model based on a new approach. Meteorol. Atmos. Phys. 82, 271–285. https://doi.org/10.1007/s00703-001-0587-6

Kain, J.S. (2004). The Kain–Fritsch convective parameterization: An update. J. Appl. Meteorol. Climatol. 43, 170–181. https://doi.org/10.1175/1520-0450(2004)043%3C0170:TKCPAU%3E2.0.CO;2

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S. (2019). Evaluation of different machine learning approaches to forecasting $PM_{2.5}$ mass concentrations. Aerosol Air

Qual. Res. 19, 1400–1410. https://doi.org/10.4209/aaqr.2018.12.0450

Kessler, E. (1969). On the Distribution and Continuity of Water Substance in Atmospheric Circulations. American Meteorological Society, Boston, MA. https://doi.org/10.1007/978-1-935704-36-2_1

Kiernan, D. (2014). Natural Resources Biometrics. Open SUNY Textbooks, Milne Library, State University of New York at Geneseo.

Krzyzanowski, M., Apte, J.S., Bonjour, S.P., Brauer, M., Cohen, A.J., Prüss-Ustun, A.M. (2014). Air pollution in the megacities. Curr. Environ. Health Rep. 1, 185–191. https://doi.org/10.1007/s40572-014-0019-7

Lagesse, B., Wang, S., Larson, T.V., Kim, A.A. (2020). Predicting $PM_{2.5}$ in well-mixed indoor air for a large office building using regression and artificial neural network models. Environ. Sci. Technol. 54, 15320–15328. https://doi.org/10.1021/acs.est.0c02549

Lary, D.J., Lary, T., Sattler, B. (2015). Using machine learning to estimate global $PM_{2.5}$ for environmental health studies. Environ. Health Insights 9s1, EHI.S15664. https://doi.org/10.4137/EHI.S15664

Luong, L.T.M., Dang, T.N., Thanh Huong, N.T., Phung, D., Tran, L.K., Van Dung, D., Thai, P.K. (2020). Particulate air pollution in Ho Chi Minh city and risk of hospital admission for acute lower respiratory infection (ALRI) among young children. Environ. Pollut. 257, 113424. https://doi.org/10.1016/j.envpol.2019.113424

Marlier, M.E., Jina, A.S., Kinney, P.L., DeFries, R.S. (2016). Extreme air pollution in global megacities. Curr. Clim. Change Rep. 2, 15–27. https://doi.org/10.1007/s40641-016-0032-z

Marsden, E., Bathan, G., Tsevegjav, B., Velez, M.C.R. (2019). Making Urban Asia's Air Cleaner (ADB Briefs). Asian Development Bank, Manila, Philippines. https://doi.org/10.22617/BRF190609-2

Mlawer, E.J., Taubman, S.J., Brown, P.D., Iacono, M.J., Clough, S.A. (1997). Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. J. Geophys. Res. 102, 16663–16682. https://doi.org/10.1029/97JD00237

Monin, A., Obukhov, A. (1954). Basic laws of turbulent mixing in the atmosphere near the ground. Tr. Geofiz. Inst., Akad. Nauk SSSR 24, 163–187.

NOAA National Centers for Environmental Information (2001). Global Surface Hourly [Ho Chi Minh City]. NOAA National Centers for Environmental Information. National Climatic Data Center, NESDIS, NOAA, U.S. Department of Commerce. https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc%3AC00532/html# (accessed 29 October 2019).

NOAA National Centers for Environmental Prediction (NCEP) (2011). NOAA/NCEP Global Forecast System (GFS). https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs (accessed 1 May 2021).

Obukhov, A. (1946). Turbulence in thermally inhomogeneous atmosphere. Trudy Inst. Teor. Geofiz. Akad. Nauk SSSR 1, 95–115.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Pires, J.C.M., Martins, F.G., Sousa, S.I.V., Alvim-Ferr, M.C.M., Pereira, M.C. (2008). Prediction of the daily mean PM_{10} concentrations using linear models. Am. J. Environ. Sci. 4, 445–453. https://doi.org/10.3844/ajessp.2008.445.453

Shaddick, G., Thomas, M.L., Mudu, P., Ruggeri, G., Gumy, S. (2020). Half the world’s population are exposed to increasing air pollution. npj Clim. Atmos. Sci. 3, 1–5. https://doi.org/10.1038/s41612-020-0124-2

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Wang, W., Huang, X.Y., Duda, M. (2008). A description of the advanced research WRF Version 3. UCAR/NCAR. https://doi.org/10.5065/D68S4MVH

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Liu, Z., Berner, J., Wang, W., Powers, J.G., Duda, M.G., Barker, D.M., Huang, X.Y. (2019). A description of the advanced research WRF model Version 4. UCAR/NCAR. https://doi.org/10.5065/1DFH-6P97

Swamidass, P.M. (Ed.) (2000). MEAN ABSOLUTE PERCENTAGE ERROR (MAPE), in: Encyclopedia of Production and Manufacturing Management, Springer US, pp. 462–462. https://doi.org/10.1007/1-4020-0612-8_580

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58, 267–288. http://www.jstor.org/stable/2346178

U.S. EPA (2018). Technical assistance document for the reporting of daily air quality: The Air Quality Index (AQI), September 2018 [edition]. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Air Quality Assessment Division, Research Triangle Park, NC, USA.

Van Rossum, G., Drake, F.L. (2009). Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vu, H.N.K., Ha, Q.P., Nguyen, D.H., Nguyen, T.T.T., Nguyen, T.T., Nguyen, T.T.H., Tran, N.D., Ho, B.Q. (2020). Poor air quality and its association with mortality in Ho Chi Minh City: Case study. Atmosphere 11, 750. https://doi.org/10.3390/atmos11070750

Willmott, C., Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82. https://doi.org/10.3354/cr030079

World Bank, Institute for Health Metrics and Evaluation (2016). The Cost of Air Pollution: Strengthening the Economic Case for Action. World Bank, Washington, DC. © World Bank.

Yang, D., Ye, C., Wang, X., Lu, D., Xu, J., Yang, H. (2018). Global distribution and evolvement of urbanization and $PM_{2.5}$ (1998–2015). Atmos. Environ. 182, 171–178. https://doi.org/10.1016/j.atmosenv.2018.03.053

Yu, F., Cui, K., Sheu, H.L., Hsieh, Y.K., Tian, X. (2021). Sensitivity analysis for dry deposition and $PM_{2.5}$-bound content of PCDD/Fs in the ambient air. Aerosol Air Qual. Res. 21, 210118. https://doi.org/10.4209/aaqr.210118

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S. (2019). $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere 10, 373. https://doi.org/10.3390/atmos10070373

Zhai, B., Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China. Sci. Total Environ. 635, 644–658. https://doi.org/10.1016/j.scitotenv.2018.04.040

Zhang, J., Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong. IJERPH 14, 114. https://doi.org/10.3390/ijerph14020114

Zhang, L., Wilson, J.P., MacDonald, B., Zhang, W., Yu, T. (2020). The changing $PM_{2.5}$ dynamics of global megacities based on long-term remotely sensed observations. Environ. Int. 142, 105862. https://doi.org/10.1016/j.envint.2020.105862

Zhang, X., Kang, J., Chen, H., Yao, M., Wang, J. (2018). $PM_{2.5}$ meets blood: In vivo damages and immune defense. Aerosol Air Qual. Res. 18, 456–470. https://doi.org/10.4209/aaqr.2017.05.016