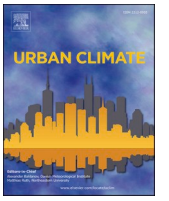
Urban Climate 38 (2021) 100890

Contents lists available at ScienceDirect

Urban Climate

journal homepage: www.elsevier.com/locate/uclim



Improvement of air quality index prediction using geographically weighted predictor methodology

Narathep Phruksahiran [*,1]

Department of Electrical Engineering, Chulachomklao Royal Military Academy, Nakhonnayok 26001, Thailand

ARTICLE INFO

Keywords:
Air quality index machine-learning algorithm prediction performance

ABSTRACT

An air quality index (AQI) is calculated based on other atmospheric pollutants and interprets how polluted the air currently is. With increasing air pollution, implementing efficient air quality monitoring models that collect information about air pollution concentration and assess air pollution in each area is necessary. Prediction methods for air quality forecasting have attracted much attention due to their performance and flexible capability. In this paper, an ensemble prediction methodology called the geographically weighted predictor method (GWP) combines the best version of machine learning algorithms and uses the additional predictor variables for prediction at the hourly level. The proposed method is employed on the Bangkok Air Quality dataset. Compared with regular machine learning models, the proposed method has better prediction performance in all prediction horizons. The results obtained suggest that such a novel model may help to enhance the accuracy of AQI prediction.

1. Introduction

Nowadays, the characteristics of health problems change according to the global environment and air pollution. Especially the diseases that occur in the respiratory system. From issues that were not so important in some countries, it must be accepted that urban society and the expansion of industrial cause problems that all sectors must cooperate to solve. Especially the government mechanism that must create methods and regulations to reduce these problems. The publications have repeatedly confirmed that environmental pollution is a significant cause of health and harms the well-being of the population (Kumari and Toshniwal, 2020; Lee et al., 2020; Zhong et al., 2019; Karimian et al., 2019).

Six pollutants important for determining the air quality index are $PM_{2.5}$, PM_{10}, O_3, CO, NO_2, and SO_2. Each has the following characteristics and initial health risks: PM stands for particulate matter and comprises dense particles and fluid droplets located in the atmosphere. $PM_{2.5}$ has suspended particularly smaller than 2.5 μm in aerodynamic diameter, and PM_{10} has suspended particularly smaller than ten mu in aerodynamic diameter. Both can cause dangerous health dilemmas within the lungs and bloodstream. Of these, particles less than 2.5 μm in diameter affect the most significant wellness hazard. Ozone (O_3) is a highly reactive gas comprised of three oxygen atoms. Ozone has two characteristics of concern to human health. First, it causes skin cancer and cataracts. Second, it reacts chemically with many biological molecules in the respiratory tract, leading to several adverse health effects. The well-being influences of CO depend on the CO consistency and period of exposure, and specific individual's health situation. For illustration, most people will

* Corresponding author.
E-mail address: narathepp@gmail.com.
[1] Present/permanent address. Department of Electrical Engineering Chulachomklao Royal Military Academy Nakhonnayok 26,001 THAILAND

https://doi.org/10.1016/j.uclim.2021.100890
Received 26 October 2020; Received in revised form 1 May 2021; Accepted 13 June 2021
Available online 17 June 2021

not encounter any indications from increased exposure to CO levels of about 1 to 70 ppm, but some heart cases might feel a rise in chest injury. Nitrogen Dioxide (NO_2) is a group of highly reactive gases. Inhaling air with an unusual gathering of NO_2 can aggravate airways in the human respiratory system. Sulfur dioxide (SO_2) can injure the human respiratory system and cause breathing difficulty. It can react with other mixtures in the atmosphere to form small particles, contributing to particulate matter (PM) pollution and health problems.

The most critical issue in establishing regulations is analytical data and predictive mechanisms, including an efficient intelligence system that can support people to decrease air pollution. Therefore, investigation about air quality forecasting with modern and reliable technology is required. These technologies and studies about the consequences of air pollution have been developed continuously during the past year (Ma et al., 2019). Various models and architectures regarding the $PM_{2.5}$ concentration prediction were presented, and many researchers have been working on proposing a better method to improve prediction accuracy. For example, Wang et al. (2018) proposed a statistically reliable and interpretable national modeling framework to calibration of the daily ground $PM_{2.5}$ concentrations. Kang et al. (2018) investigated big-data and machine learning techniques for air quality forecasting. Masih (2019) report that machine learning techniques are mainly conducted in continent Europe and America, and the pollution estimation is generally performed by using ensemble learning and regression-based approaches. Wang et al. (2019) analyzed the acquired data and predicted the following data using a neural network. The above studies show that the research approach will use machine learning algorithms to analyze and forecast pollution values more accurately.

1.1. Related work

The precise prediction of AQI and other atmospheric pollutant levels is crucial for policymakers to give early warning and develop control strategies. Because of the nonlinearity of the time series and the volatility characteristics of atmospheric pollutants, it isn't simple to predict accurately (Qiao et al., 2019; Liu et al., 2019; Soh et al., 2018). Due to the complexity of land use and weather conditions in the respective large cities, it is essential to generate your town model using the parameters and other requirements. Table 1 presents the research results on the value of each pollutant by country and the variables studied.

RF stands for the Random Forest, XGBoost is eXtreme Gradient Boosting, DNN indicates Deep Neural Network, LR means Linear Regression, SE denotes for Stacked Ensemble, RL implies Reinforcement Learning, EN signifies Elastic-Net Regression, GB is Gradient Boosting, and NN means Neural Network.

From Table 1, it can be seen that most research studies used a single algorithm for analysis. And feature design for machine learning algorithms will be the science and art of the program designer. The highlight of this research is that the measuring station's geographic coordinates are interpolated with algorithms to determine the value of each feature. In addition, three types of algorithms have been applied for comparison. And we are also taking into account the importance of each pollution type variable that must be analyzed together.

1.2. Motivation and contributions

Although various attempts have been made to predict $PM_{2.5}$ concentration, the correlation between features that influence $PM_{2.5}$ concentration prediction is still not well recognised. Only a few investigations, of limited extent, have examined the significance of these features on $PM_{2.5}$ concentration prediction (Hu et al., 2019). It is essential to identify the relationship between the various influencing factors and the $PM_{2.5}$ concentration before forming the predictive model, which ensures that the model uses the proper input prognostic features for prediction (Tao et al., 2019; Hvidtfeldt et al., 2018). But current research is still limited to using a single

Table 1
Previous studies and related work.

Study	Location	Pollutants types	Methodology
Aliyu and Botai (2018)	Zaria, Nigeria	$PM_{2.5}$, PM_{10}, CO, SO_2	Statistical analysis
Amini et al. (2019)	Tehran, Iran	$PM_{2.5}$, NO_2	Statistical analysis
Joharestani et al. (2019)	Tehran, Iran	$PM_{2.5}$	RF, XGBoost, DNN
Lin et al. (2020)	Taiwan	$PM_{2.5}$	Linear data fusion
Belis et al. (2019)	Western Balkans	$PM_{2.5}$	Statistical analysis
Lim et al. (2019)	Seoul, South Korea	Air particles	LR, RF, SE
Chang et al. (2019)	China	$PM_{2.5}$	RL
Chen et al. (2019a, 2019b)	China	$PM_{2.5}$	Statistical analysis
Liu et al. (2020)	China	O_3	RF
Xue et al. (2019)	China	$PM_{2.5}$	EN
Zhang et al. (2019)	China	$PM_{2.5}$	GB
Zhao et al. (2019)	China	$PM_{2.5}$	RF
Chen et al. (2019a, 2019b)	Europe	$PM_{2.5}$, NO_2	Model comparison
Meng et al. (2018)	United State	$PM_{2.5}$	RF
Requia et al. (2019)	United State	$PM_{2.5}$	Statistical analysis
Deters et al. (2017)	Quito	$PM_{2.5}$	Regression
Stafoggia et al. (2019)	Italy	$PM_{2.5}$, PM_{10}	RF
Araki et al. (2020)	Japan	$PM_{2.5}$	NN
This study	Bangkok, Thailand	$PM_{2.5}$, PM_{10}, O_3, CO, NO_2, SO_2	RF, XGBoost, NN

machine-learning algorithm to process all parameters of AQI prediction. Mathematical fundamentals and variables change with time resulting in different programs being more effective in their analysis and forecasting accuracy.

It is well known that pollution problems are increasing rapidly and directly impacting people's livelihoods and health. But the number of measurement equipment for pollution estimation in some areas is insufficient and comprehensive, especially in remote areas and not a big city, which is the state of the problem found today. But in fact, the pollution problem is distributed in all units, and residents not only in the big cities.

This research aims to enhance the efficiency of measuring and alarm systems to cover all areas. Especially if there are no measuring instruments in a specific area, it is also necessary to forecast the pollution value. It is essential to build a suitable model based on the weighted features from that particular area based on surrounding available parts for the reasons mentioned above. Then the parameters of the respective algorithm should also be adjusted so that it can deliver improved results.

This paper aims to develop processing methods that can improve the prediction accuracy at the specific area without measuring measurement and test its performance using the air pollution dataset in the Bangkok Metropolitan Region, Thailand. The main contributions of this paper are summarized as follows.

1. This study developed an ensemble model that integrated multiple machine learning algorithm and interpolated gridded variable in the model as separate predictor parameter to estimate hourly atmospheric pollutions and AQI at an exciting area that has no measurement instruments.
2. A novel ensemble model called geographically weighted predictor method (GWP) is proposed for multivariate time series forecasting. In this study, Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Deep Neural Network (NN) machine learning methods were used to investigate atmospheric pollutant concentration prediction. The performance of the prophecy was evaluated using R^2, root mean square error (RMSE), and mean absolute error (MAE) matrics.

The remainder of this paper is organized as follows: Section 2 outlines the study area and materials used in this paper. The proposed method is introduced in section 3. The experiment and analysis are given in section 4. Finally, the conclusion and the outlook for the future are presented in section 5.

2. Study area and materials

2.1. Study area

The study area is the Bangkok Metropolitan Region. Bangkok is the capital and most populous city of Thailand. Bangkok is located at the center of Thailand (Coordinates: 13°45′09″N 100°29′39″E). The Bangkok Metropolitan Region has a tropical savanna climate and is under the influence of the South Asian monsoon system. The weather of Bangkok is hot throughout the year, with temperatures

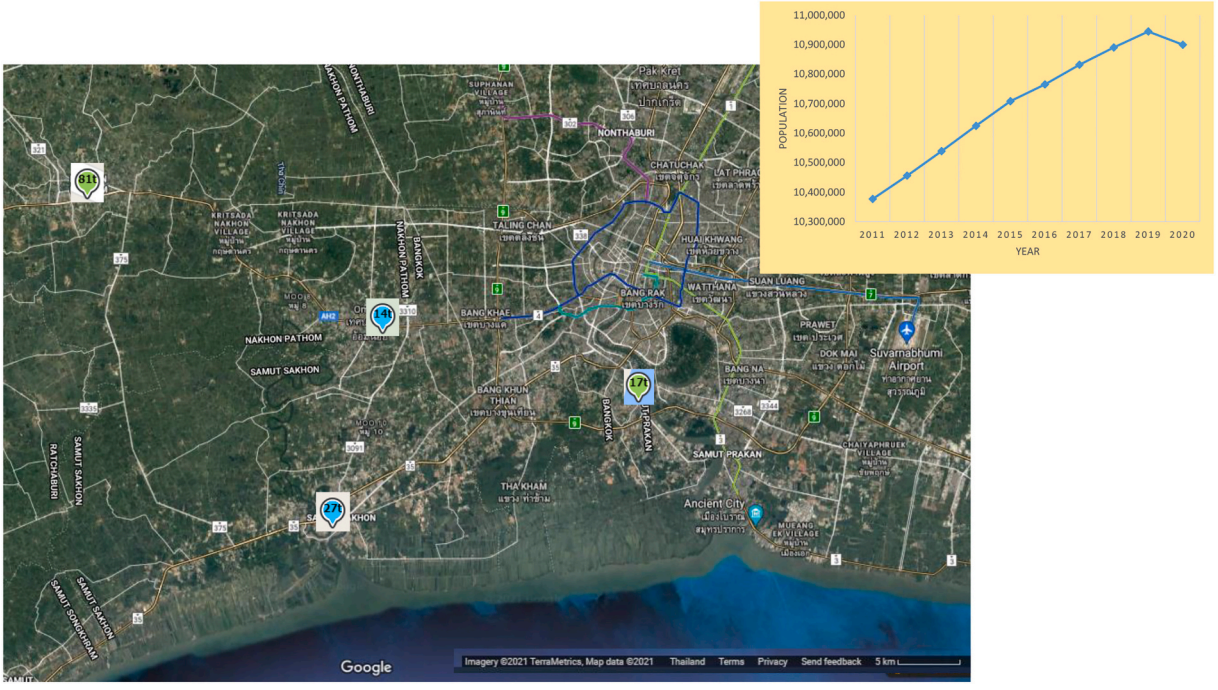


Fig. 1. The population of Bangkok metropolitan region and air quality monitoring stations from Google Map.

over 30°C, annual rainfall of 1500 mm, and average relative humidity of 78.0%. Bangkoks urban extent and population have been increasing over the last few decades. This has led to inadequate infrastructure and a haphazard layout with limited roads, terrible traffic, and severe air pollution. According to the National Statistical Office (Number of Population from Registration by Age Group Province and Region: 2011–2020, n.d), the population increase in the Bangkok metropolitan area has increased from 10.376 million in 2011 to 10.899 million in 2020, as shown in Fig. 1.

2.2. Datasets

The experimental datasets were retrieved from the Pollution Control Department of the Ministry of Natural Resources and Environment (Air Quality and Noise Management Division, 2020). This analysis is performed using air pollution (suspended particulates smaller than 2.5 μm in aerodynamic diameter ($PM_{2.5}$), suspended particulates smaller than 10 μm in aerodynamic diameter (PM_{10}), ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2)) data for the period March–August 2020, were collected at an hour sampling rate from four air quality monitoring stations. Table 2 presents the real location in latitude and longitude of the air quality monitoring stations in this study, as depicted in Fig. 1.

2.3. Air quality index

The air quality index (AQI) is a universal format that is generally used in many countries. The AQI level in Thailand is based on the level of six atmospheric pollutants, namely $PM_{2.5}$, PM_{10}, CO, O_3, NO_2, and SO_2 measured at the monitoring stations throughout each city.

Air quality index used in Thailand is determined by comparing the air quality standards in the general atmosphere of 6 types of air pollutants, namely $PM_{2.5}$ averaged 24 h, PM_{10} averaged 24 h, CO averaged 8 h, O_3 averaged 8 h, NO_2 an average of 1 h, and SO_2 an average of 1 h. The previously calculated air quality sub-index value of any air pollutant has the highest amount that will be used as the air quality index of that day. Air quality index criteria for Thailand are divided into five levels: AQI 0–25 very good quality, AQI 26–50 good quality, AQI 51–100 medium, AQI 101–200 began to affect health, and AQI $\geq$ 201 affecting health. The quality index of each air pollutants will be calculated from the concentration measurement data and level concentration equivalent by using the following formula:

$$I = \frac{I_j - I_i}{X_j - X_i}\ (X - X_i) + I_i \tag{1}$$

where I is air quality sub-index value, X is the concentration of air pollutants from the measurement results, X_i, X_j are the lowest and the highest value of the range with the value X, and I_i, I_j are the lowest and the highest air quality sub-index value that is the highest value of the range with that I value. In Table 3, the index of air pollution and AQI classification used in Thailand are categorized according to their concentration.

2.4. Performance comparison

In this study, the mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2) can be used to evaluate the performance of prediction model. These parameter can evaluate the degree of change and accuracy of data, and can measure the prediction quality of the developed machine learning models. The calculation formula are

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid o_i - p_i \mid \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \tag{3}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2} \tag{4}$$

Table 2
Air quality monitoring stations.

Station	Latitude	Longitude
14 t	13.705458	100.315675
17 t	13.652154	100.531840
27 t	13.550498	100.264252
81 t	13.832076	100.057961

Table 3
Thailand's AQI classification.

AQI	$PM_{2.5}$	PM_{10}	O_3	CO	NO_2	SO_2
	($\mu g/m^3$)	($\mu g/m^3$)	(ppb)	(ppm)	(ppb)	(ppb)
0–25	0–25	0–50	0–35	0–4.4	0–60	0–100
26–50	26–37	51–80	36–50	4.5–6.4	61–106	101–200
51–100	38–50	81–120	51–70	6.5–9.0	107–170	201–300
101–200	51–90	121–180	71–120	9.1–30.0	171–340	301–400
$\geq$ 201	$\geq$ 91	$\geq$ 181	$\geq$ 121	$\geq$ 30.1	$\geq$ 341	$\geq$ 401

where n is the sample size, p is the predicted time series, o and $\bar{o}$ are the observed and the mean of the real observation set. A smaller value of MAE and RMSE means better performance in prediction.

3. Method

3.1. Geographically weighted predictor

The proposed geographically weighted predictor method (GWP) combine the representation of variables using the spatial interpolated variable based on the gridded known variable from other measurement stations. This developed GWP method creates a coordinate system as a grid-based according to the actual position of surrounding air quality monitoring stations in the geographic structure. The data entered into processing is the real hourly value of each air pollutions obtained from the measurements of the surrounding stations and are not average mean. The distance referenced by geographic information is also taken into account by interpolation calculation. In this study, the pollution concentrations of station 17 t, 27 t, and 81 t were used to estimate the interpolated value for the station 14 t, assume that the real data from the measurement is unknown. Fig. 2 presents the determination of spatial interpolated variable. The x-axis and y-axis stand for longitude and latitude, respectively. The station names are plotted according to its position. The colour range was used to represent the intensity of the interpolated value for each row of input dataset.

In this study, the AQI prediction procedure using the developed geographically weighted predictor method was established following four steps, as shown in Fig. 3.

3.2. Step 1: data preprocessing

Before the development of the model, the collected data were preprocessed to recognise whether it needs to be cleaned. The

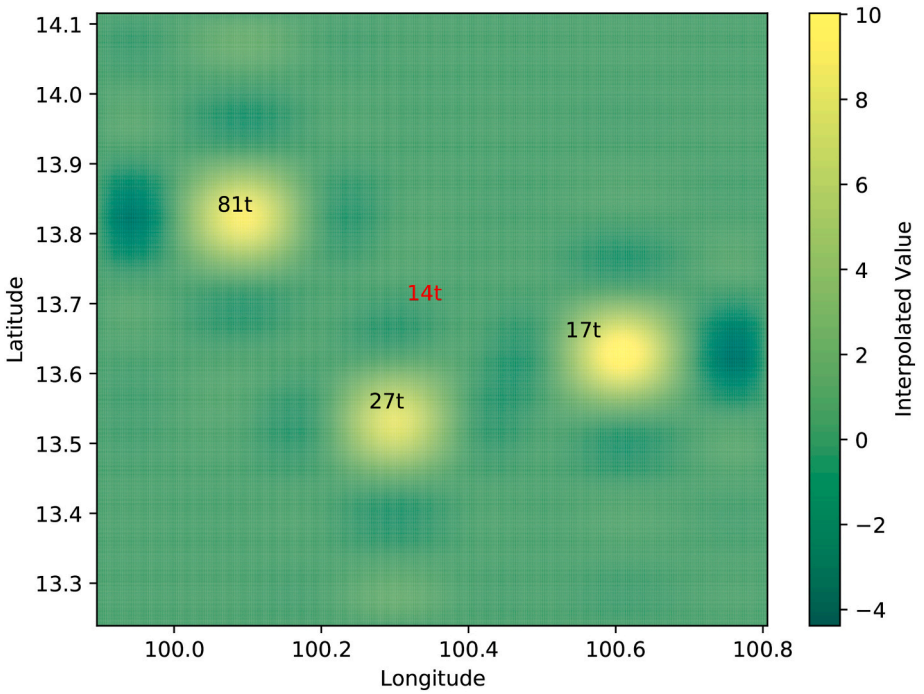


Fig. 2. Hourly geographically weighted predictor estimation of each feature.

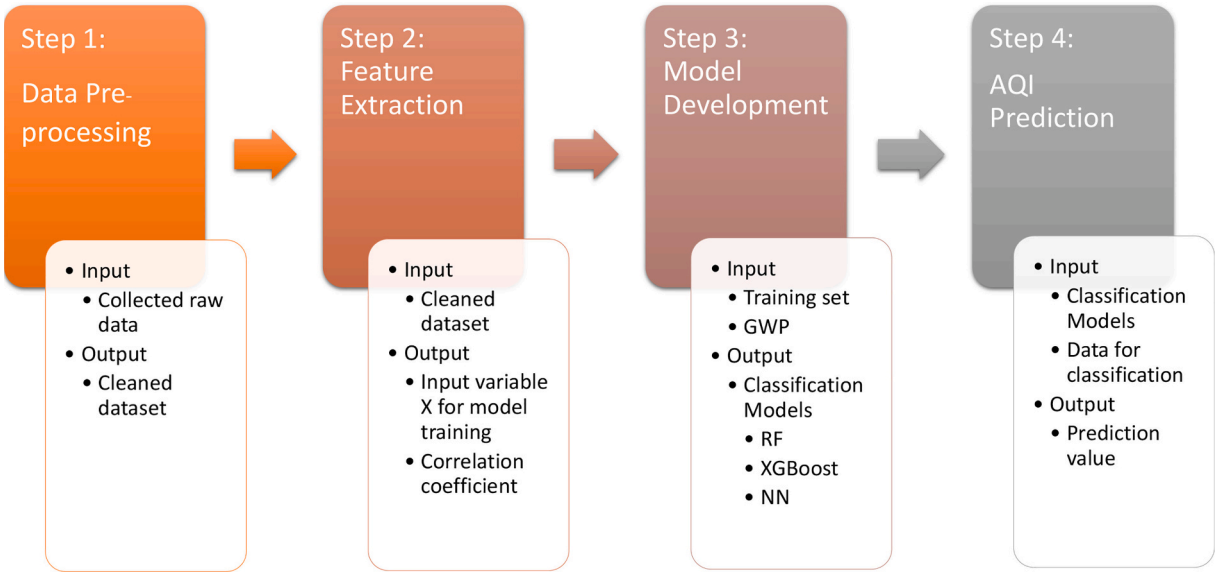


Fig. 3. Processing flowchart.

cleaning of data was based on two requirements, which are particularly incorrect and missing data periods. Some method is expected to fill the missing values before model development. Interpolation was implemented to estimate and fill each variable with missing values.

Standardisation was used to normalise the features. The entire dataset is normalised by deducting the mean and scaling by the variance of each element as

$$s = \frac{x - \bar{x}}{\sigma} \tag{5}$$

where x_i, $\bar{x}$, and σ_x are the sample values, the mean, and variance of each characteristic variable, respectively.

3.3. Step 2: feature selection and correlation analysis

It is interested in whether there is any relationship or connection between these quantities. Suppose one characteristic time series is the vector $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, the other time series is vector $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$. If there is such dependency between the two variables x and y, from this point of view, one speaks of a correlation between $\mathbf{X}$ and $\mathbf{Y}$ as (Papula, 1999)

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} \tag{6}$$

where r is the correlation coefficient, x_i, $\bar{x}$, y_i, and $\bar{y}$ are the sample values, and the mean of each characteristic variable, respectively. The gab between $\mathbf{X}$ and $\mathbf{Y}$ is smaller and the correlation is greater when the absolute of r is closer to 1. The features used for machine learning algorithms in this study are latitude, longitude, weekday, hour, $PM_{2.5}$, PM_{10}, O_3, CO, NO_2, and SO_2.

3.4. Step 3: model development using GWP

The model development aims to estimate each atmospheric pollution concentration at a specific area, which has no measuring instrument in hourly intervals. In this study, models were developed using the machine learning tool of Python (Scikit-learn, 2020). Machine learning algorithms are described as learning a target function $\mathbf{f}$ that best maps input variables $\mathbf{X}$ to an output variable $\mathbf{Y}$. This objective is expressed in a machine learning algorithm as

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}) \tag{7}$$

with

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{8}$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ x_m \end{bmatrix} \tag{9}$$

Dataset records from station 17 t, 27 t, and 81 t were used for train the models. In this study, Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Deep Neural Network (NN) machine learning methods were used to investigate atmospheric pollutant concentration prediction, because of the extensive usage and performance superiority.

In GWP, the six geographically weighted predictors ($PM_{2.5,\ g}$, $PM_{10,\ g}$, $O_{3,\ g}$, CO_g, $NO_{2,\ g}$, and $SO_{2,\ g}$) were added to $\mathbf{X}$ in Eq. (8) to develop the new model and to predict the hourly air pollution concentrations at the station 14 t and new input variable $\mathbf{X}_g$ as

$$\mathbf{X}_g = \begin{bmatrix} x_{11} & \cdots & x_{1n} & x_{1,g1} & \cdots & x_{1,gn} \\ x_{21} & \cdots & x_{2n} & x_{2,g1} & \cdots & x_{2,gn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m1} & \cdots & x_{mn} & x_{m,g1} & \cdots & x_{m,gn} \end{bmatrix} \tag{10}$$

where index g indicates the geographically weighted predictors of each selected features. The prediction models based on Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Deep Neural Network (NN) machine learning methods have been trained again using input variable $\mathbf{X}_g$.

From the original dataset of three stations, 70% of the data are randomly sampled to train the three machine learning models, and the remaining 30% are later used to validate the trained machine learning models. RMSE was used as the optimization criteria. The parameter of each machine learning algorithm are optimized by running through significant parameter for respective air pollutions.

3.5. Step 4: AQI prediction

After training to convergence, the normal prediction models and the optimal weights of GWP prediction models are obtained. The evaluations were conducted using the hourly air pollution concentration of station 14 t, which has not been used in training dataset.

4. Results and discussion

The correlation coefficient between each atmospheric pollution was calculated. As shown in Table 4, PM_{10}, CO, NO_2, and SO_2 are positive correlation with $PM_{2.5}$ concentration, while O_3, hour, and weekday are negative correlation with $PM_{2.5}$ concentration. It is found that all the time series variables are correlated with atmospheric pollution concentration, which indicates that there should be used as the input of the prediction model.

The predicted and observed pollution concentrations using RF and RF-GWP are presented in Fig. 4. The solid line with | stands for observed pollution concentrations. The dashed line with ∘ presents the prediction results using common RF algorithm. The dashed line with ⋄ represents prediction results by applying the GWP variable to the RF machine learning algorithm. It can be observed from the figure that the RF-GWP model produced results which can follow the fluctuations of real values during the test set better than conventional RF.

Fig. 5 shows the predicted and observed times series data using XG and XG-GWP. The solid line with | stands for observed pollution concentrations. The dashed line with Δ presents the prediction results using standard XG algorithm. The dashed line with ▽ represents prediction results by applying the GWP variable to the XG machine learning algorithm. It is evident that the variation of time series nad overall trends for both prediction results based on common XG and prediction results based on XG-GWP seem to be similar through the experiment duration. However, the values are different. It can be seen from the figure that the prediction results using XG-GWP are closer to the observed values in every atmospheric pollution.

Table 4
Correlation coefficient (r) between air pollutions.

r	$PM_{2.5}$	PM_{10}	O_3	CO	NO_2	SO_2	Hour	Weekday
$PM_{2.5}$	1.0000	0.5747	−0.1836	0.2132	0.3622	0.2582	−0.1153	−0.0139
PM_{10}	0.5747	1.0000	−0.1275	0.1394	0.3453	0.3035	−0.0638	−0.0275
O_3	−0.1836	−0.1275	1.0000	−0.1246	−0.2391	0.0453	0.2092	0.0125
CO	0.2132	0.1394	−0.1246	1.0000	0.3776	0.2629	−0.0439	0.0196
NO_2	0.3622	0.3453	−0.2391	0.3776	1.0000	0.4473	−0.0675	−0.0073
SO_2	0.2582	0.3035	0.0453	0.2629	0.4473	1.0000	−0.0774	−0.0024
Hour	−0.1153	−0.0638	0.2092	−0.0439	−0.0675	−0.0774	1.0000	0.0128
Weekday	−0.0139	−0.0275	0.0125	0.0196	−0.0073	−0.0024	0.0128	1.0000

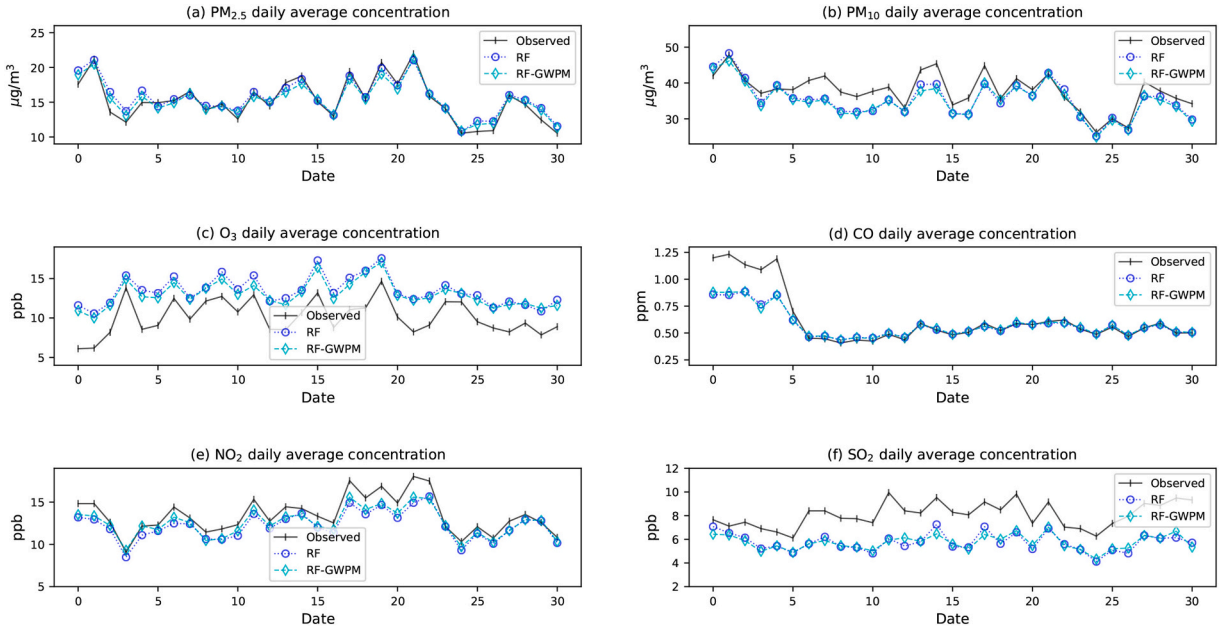


Fig. 4. Daily average pollution concentration of observation (|), prediction results using common RF (∘), and prediction results using RF-GWP (⋄): (a) $PM_{2.5}$, (b) PM_{10}, (c) O_3, (d) CO, (e) NO_2, and (f) SO_2 concentrations.

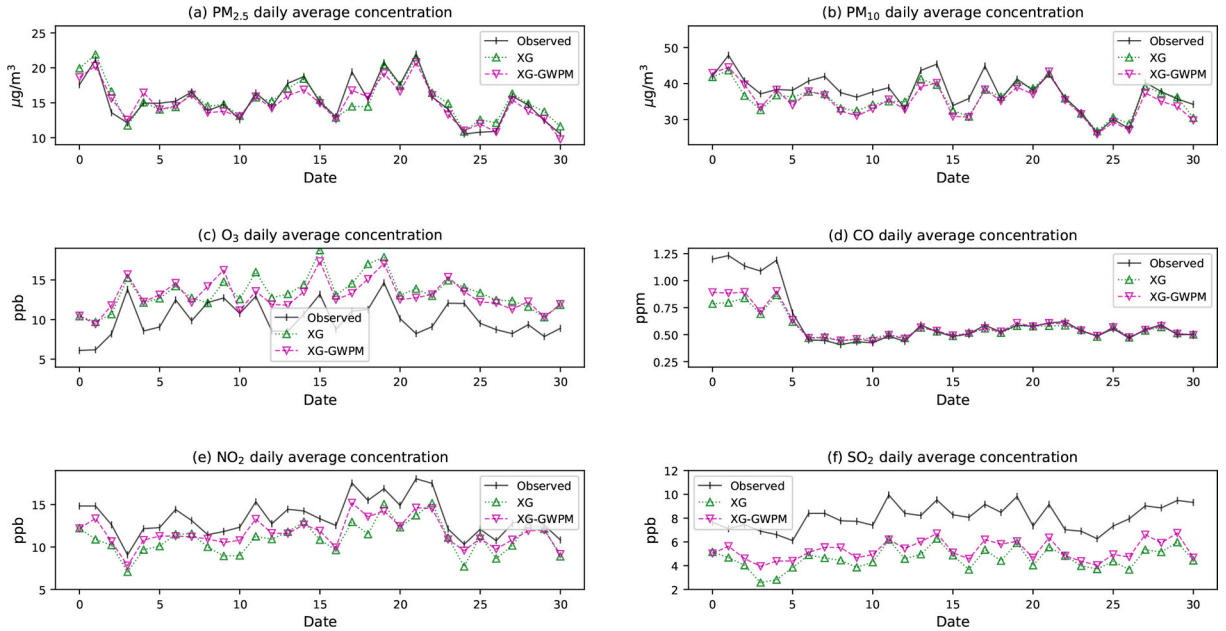


Fig. 5. Daily average pollution concentration of observation (|), prediction results using common XG (Δ), and prediction results using XG-GWP (▽): (a) $PM_{2.5}$, (b) PM_{10}, (c) O_3, (d) CO, (e) NO_2, and (f) SO_2 concentrations.

The simulation results based on NN and NN-GWP were shown in Fig. 6. The solid line with | stands for observed pollution concentrations. The dashed line with + presents the prediction results using conventional NN algorithm. The dashed line with × represent prediction results by applying the GWP variable to the NN machine learning algorithm. From the prediction results, it can be seen that NN and NN-GWP have similar performance. Still, the prediction error of NN-GWP is lower than common NN, which shows that the application of GWP can bee used to improve the performance of traditional NN algorithm.

The AQI prediction results are shown in Fig. 7 with observed values (|), prediction results using RF (∘), prediction results using RF-GWP (⋄), prediction results using XG (Δ), prediction results using XG-GWP (▽), prediction results using NN (+), and prediction results

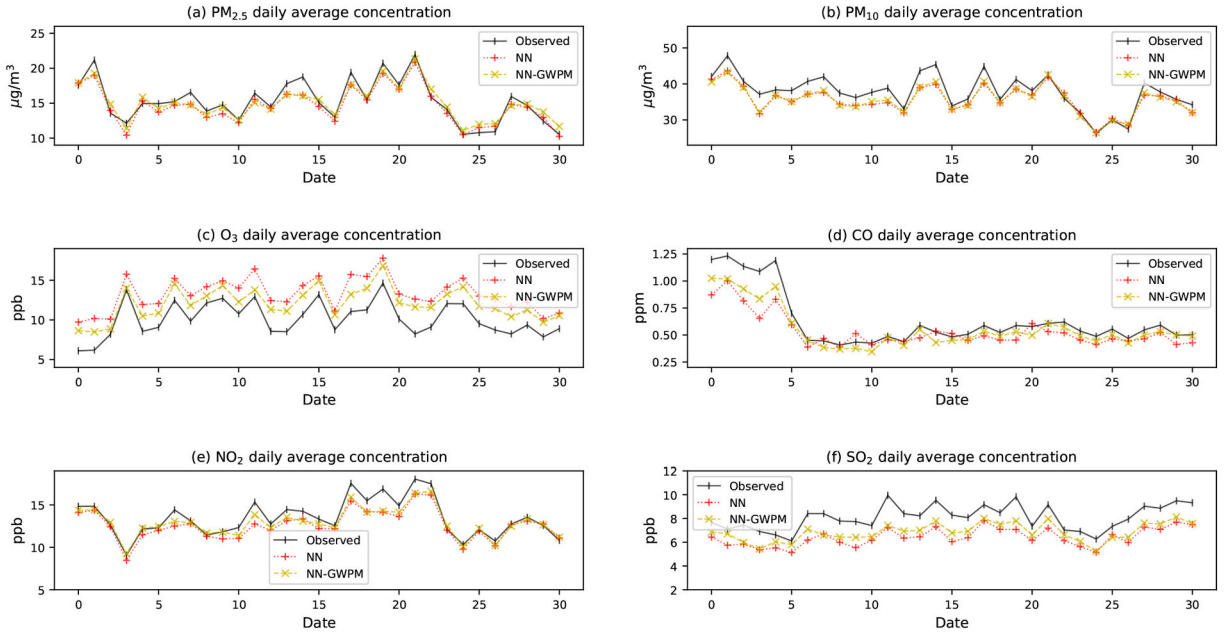


Fig. 6. Daily average pollution concentration of observation (|), prediction results using common NN (+), and prediciton results using NN-GWP (×): (a) $PM_{2.5}$, (b) PM_{10}, (c) O_3, (d) CO, (e) NO_2, and (f) SO_2 concentrations.

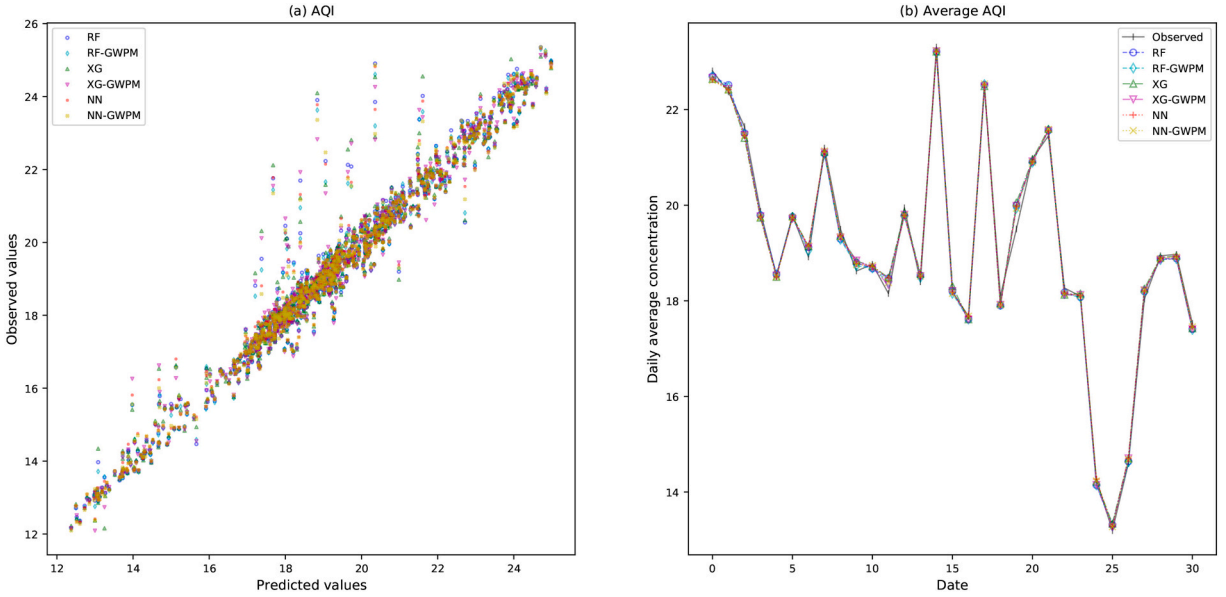


Fig. 7. (a) AQI scatter plot with the comparison model, (b) Daily average AQI of observation: observed values (|), prediction results using RF (∘), prediciton results using RF-GWP (⋄), prediction results using XG (Δ), prediciton results using XG-GWP (∇), prediction results using NN (+), and prediciton results using NN-GWP (×).

using NN-GWP (×). Fig. 7(a) presents the scatter plot with the comparison models, which indicated the stability of the model on a temporary order. Fig. 7(b) shows the model performance at the daily level. The AQI prediction results were quite accurate in obtaining the hourly and daily average scale.

Table 5 lists the quantitative results by RMSE, R^2, and MAE, which gives a comparative analysis of RF, XG, NN, and their combinations with the proposed model of GWP. The index *G* indicate the quantitative results using GWP of each pollution. This table shows a comparison of the performance of the models applied to each variable. This table makes it potential to see the impact and change resulting from the application of the GWP. For illustration, the RMSE value of $PM_{2.5}$ will have less error from 0.3315 to 0.3262 when

Table 5
Quantitative results by RMSE, R^2, and MAE.

	RMSE			R^2			MAE		
	RF	XG	NN	RF	XG	NN	RF	XG	NN
$PM_{2.5}$	0.3315	0.3501	6.5693	0.9888	0.9877	0.9894	0.2559	0.2745	0.2518
$PM_{2.5,\ G}$	0.3262	0.3385	6.5853	0.9892	0.9884	0.9897	0.2531	0.2626	0.2441
PM_{10}	0.3339	0.3272	22.7485	0.9841	0.9846	0.9858	0.2549	0.2554	0.2381
$PM_{10,\ G}$	0.3257	0.3300	22.7484	0.9849	0.9846	0.9857	0.2480	0.2541	0.2385
O_3	1.4561	1.4781	1.3373	0.9194	0.9171	0.9332	1.0772	1.0972	0.9987
$O_{3,\ G}$	1.4100	1.4454	1.2800	0.9222	0.9198	0.9369	1.0487	1.0786	0.9545
CO	0.1870	0.1965	0.2532	0.9902	0.9900	0.9793	0.1472	0.1529	0.1928
$CO_{,G}$	0.1848	0.1831	0.2007	0.9902	0.9902	0.9868	0.1460	0.1431	0.1502
NO_2	1.6400	2.0622	1.5684	0.3361	0.2460	0.3573	1.2100	1.5459	1.1289
$NO_{2,\ G}$	1.5800	1.8923	1.5503	0.3569	0.2713	0.3456	1.1548	1.3861	1.1165
SO_2	0.8142	1.0859	0.6829	0.2375	0.1459	0.2577	0.6436	0.9141	0.4957
$SO_{2,\ G}$	0.7773	0.9397	0.6220	0.2999	0.1685	0.2728	0.6128	0.7626	0.4305
AQI	0.5591	0.5648	0.5178	0.9558	0.9546	0.9616	0.3234	0.3281	0.2966
$AQI_{,G}$	0.5121	0.5309	0.4994	0.9626	0.9599	0.9643	0.3053	0.3173	0.2929

using GWP methodology.

To analyze the prediction effect of each model more intuitively, Table 6 illustrates the improving percentage of pollution concentration prediction during the whole test set by using GWP predictor and standard machine learning algorithm. It can be seen from the table that the combination of the machine learning algorithm and GWP can afford more increased accuracy in most cases, in particular, the application of RF and GWP deliver 8.4036% better RMSE by AQI prediction. Besides, the RF-GWP can provide a higher coefficient of determination between observed and predicted AQI with 0.7134% better than conventional RF. Furthermore, the method of GWP with RF machine learning algorithm can deliver 5.6182% improvement of MAE compared to purely RF algorithm.

One strength of our studies is that we developed a geographically weighted predictor method that combines features using the spatial interpolated variable. As demonstrated in Table 1, most of the previous work considers only one pollutant type in the machine learning algorithm. Therefore the results between studies are difficult to compare with this study. From the relative level of each pollution type and the period of the significant level of pollution, all features as variables result in a more accurate analysis and prediction of pollution. Our study gives new insight into the configuration of the training table using geographically interpolated data and the improvement of the air quality index prediction.

From the experimental results shown in Fig. 7, we can see that the developed GWP algorithms can produce predictions closer to the actual observed value than using conventional algorithms. The use of geographic properties, location coordinates, and use of periods as supplementary variables in addition to statistical analysis will provide a noticeable increase in forecasting accuracy.

Another highlight of this research is that we compared the results produced by three different algorithms: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Deep Neural Network (NN) machine learning methods. It shows the efficiency of using each algorithm with different pollutant data. This basis goes back to the physics and variations of each pollutant during the day, including processing methods to model each algorithm, which can be used to select appropriate algorithms in actual use.

Although the algorithm that has been developed can work efficiently, there are some disadvantages: The measuring station also has no comprehensive information on climates, such as wind and air pressure, which can affect the spatial distribution of pollutants. By incorporating these variables into the calculation, it may be possible to obtain more accurate results.

5. Conclusions

To conclude, this paper proposed a new methodology framework that combined machine learning algorithm and GWP techniques to predict the air pollution concentration at the particular area based on hourly surrounding observed atmospheric pollutions. To validate the effectiveness of the proposed framework, a case study in the Bangkok metropolitan region, Thailand, was conducted. The prediction performance of GWP technique for the precisely observed station was evaluated.

From the industry's perspective, based on this study, we can see that developing an integrated air quality instrumentation will

Table 6
Performance comparison in %.

	RMSE			R^2			MAE		
	RF	XG	NN	RF	XG	NN	RF	XG	NN
$PM_{2.5}$	1.6185	3.2901	−0.2434	0.0344	0.0738	0.0369	1.1168	4.3097	3.0668
PM_{10}	2.4413	−0.8504	0.0005	0.0883	−0.0046	−0.0068	2.7197	0.5460	−0.1802
O_3	3.1653	2.2108	4.2894	0.3032	0.2946	0.3976	2.6503	1.6961	4.4204
CO	1.1561	6.8374	20.7436	0.0047	0.0248	0.7640	0.8227	6.4171	22.0944
NO_2	3.6602	8.2388	1.1526	6.2048	10.2898	−3.2757	4.5637	10.3353	1.0920
SO_2	4.5336	13.4638	8.9168	26.2380	15.5405	5.8916	4.7893	16.5724	13.1577
AQI	8.4036	6.0005	3.5651	0.7134	0.5507	0.2766	5.6182	3.2761	1.2517

significantly enhance the efficiency of the pollution alarm system, including the development of a small measuring tool that covers the area of interest. Much research aims to develop artificial intelligence from the academic perspective, both in algorithm development and data analysis processes, including selecting suitable features for creating a model. This research is one example that has shown great success in the development of algorithms that incorporate geographic factors into the training processing to obtain a more accurate prediction model for air quality forecasting.

Concerning atmospheric dynamic, the high-temporal resolution observations of air pollutions contribute valuable insight into the influence of chaotic changes. Overall, the proposed methodology framework as able to provide valid results in AQI and air pollutions prediction problems. Further experiments can be conducted to confirm its practicability in investigating other pollution origins in other areas. On the other hand, this paper has some limitations. Due to the data availability, this study only applied the historic air pollution data and did not involve the meteorological and metropolitan information. In future, more related characteristics should be collected to improved prediction and analysis performance and to cover a more significant area of interest.

Authorship statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the Urban Climate.

Authorship contributions Please indicate the specific contributions made by each author (list the authors' initials followed by their surnames, e.g., Y.L. Cheung). The name of each author must appear at least once in each of the three categories below.

Category 1 Conception and design of study: Narathep Phruksahiran.

acquisition of data: Narathep Phruksahiran.

analysis and/or interpretation of data: Narathep Phruksahiran.

Category 2 Drafting the manuscript: Narathep Phruksahiran.

revising the manuscript critically for important intellectual content: Narathep Phruksahiran.

Category 3 Approval of the version of the manuscript to be published Narathep Phruksahiran.

Declaration of Competing Interest

None.

Acknowledgements

The author is grateful to the Air Quality and Noise Management Division Bangkok of the Pollution Control Department of the Ministry of Natural Resources and Environment, Thailand.

References

Air Quality and Noise Management Division Bangkok of the Pollution Control Department of the Ministry of Natural Resources and Environment, Thailand, 2020. http://www.air4thai.net/webV2/history/ (accessed 10 August 2020).

Aliyu, Y.A., Botai, J.O., 2018. Reviewing the local and global implications of air pollution trends in Zaria, northern Nigeria. Urban Clim. 26, 51–59. https://doi.org/10.1016/j.uclim.2018.08.008.

Amini, H., Nhung, N.T.T., Shindler, C., Yunesian, M., Hosseini, V., Shamsipour, M., Hassanvand, M.S., Mohammadi, Y., Farzadfar, F., Vicedo-Cabrera, A.M., Schwartz, J., Henderson, S.B., Künyli, N., 2019. Short-term associations between daily mortality and ambient particulate matter, nitrogen dioxide, and the air quality index in a middle eastern megacity. Environ. Pollut. 254, 113121. https://doi.org/10.1016/j.envpol.2019.113121.

Araki, S., Shima, M., Yamamoto, K., 2020. Estimating historical $PM_{2.5}$ exposures for three decades (1987-2016) in Japan using measurements of associated air pollutants and land use regression. Environ. Pollut. 263, 114476. https://doi.org/10.1016/j.envpol.2020.114476.

Belis, C.A., Pisoni, E., Degraeuwe, B., Peduzzi, E., Thunis, P., Monforti-Ferrario, F., Guizzardi, D., 2019. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. Environ. Int. 133, 105158. https://doi.org/10.1016/j.envint.2019.105158.

Chang, S.W., Chang, C.L., Li, L.T., Liao, S.W., 2019. Reinforcement learning for improving the accuracy of $PM_{2.5}$ pollution forecast under the neural network framework. IEEE Access. 8, 9864–9874. https://doi.org/10.1109/ACCESS.2019.2932413.

Chen, J., Hoogh, K.D., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., Donkelaar, A.V., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H., Martin, R. V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019a. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934. https://doi.org/10.1016/j.envint.2019.104934.

Chen, S., Li, D., Zhang, H., Yu, D., Chen, R., Zhang, B., Tan, Y., Niu, Y., Duan, H., Mai, B., Chen, S., Yu, J., Luan, T., Chen, L., Xing, X., Li, Q., Xiao, Y., Dong, G., Niu, Y., Aschner, M., Zhang, R., Zheng, Y., Chen, W., 2019b. The development of a cell-based model for the assessment of carcinogenic potential upon long-term $PM_{2.5}$ exposure. Environ. Int. 131, 104943. https://doi.org/10.1016/j.envint.2019.104943.

Deters, J.K., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y., 2017. Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters. J. of Electr. and Comp. Engi. 2017, 1–14. https://doi.org/10.1155/2017/5106045.

Hu, Y., Sun, X., Nie, X., Li, Y., Liu, L., 2019. An enhanced LSTM for trend following of time series. IEEE Access. 7, 34020–34030. https://doi.org/10.1109/ACCESS.2019.2896621.

Hvidtfeldt, U.A., Sorensen, M., Geels, C., Ketzel, M., Khan, J., Tjonneland, A., Overvad, K., Brandt, J., Raaschou-Nielsen, O., 2018. Long-term residential exposure to $PM_{2.5}$, PM_{10}, black carbon, NO_2, and ozone and mortality in a Danish cohort. Environ. Int. 123, 265–272. https://doi.org/10.1016/j.envint.2018.12.010.

Joharestani, M.Z., 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. 7, 1–19. https://doi.org/10.3390/atmos10070373.

Kang, G.K., Gao, J.Z., Chiao, S., Lu, S., Xie, G., 2018. Air quality prediction: big data and machine learning approaches. Int. J. Environ. Sci. 1, 8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S., 2019. Evaluation of different machine learning approaches in forecasting $PM_{2.5}$ mass concentrations. Aerosol Air Qual. Res. 6, 1400–1410. https://doi.org/10.4209/aaqr.2018.12.0450.

Kumari, P., Toshniwal, D., 2020. Impact of lockdown on air quality over major cities across the globe during COVID-19 pandemic. Urban Clim. 34, 100719. https://doi.org/10.1016/j.uclim.2020.100719.

Lee, K., Hwang-Bo, H., Ji, S.Y., Kim, M.Y., Kim, S.Y., Park, C., Hong, S.H., Kim, G.Y., Song, K.S., Hyun, J.W., Choi, Y.H., 2020. Diesel particulate matter 2.5 promotes epithelial-mesenchymal transition of human retinal pigment epithelial cells via generation of reactive oxygen species. Environ. Pollut. 262, 1–10. https://doi.org/10.1016/j.envpol.2020.114301.

Lim, C.C., Kim, H., Vilcassim, M.J.R., Thurston, G.D., Gordon, T., Chen, L.C., Lee, K., Heimbinder, M., Kim, S.Y., 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ. Int. 131, 105022. https://doi.org/10.1016/j.envint.2019.105022.

Lin, Y.C., Chi, W.J., Lin, Y.Q., 2020. The improvement of spatial-temporal resolution of $PM_{2.5}$ estimation based on micro-air quality sensors by using data fusion technique. Environ. Int. 134, 105305. https://doi.org/10.1016/j.envint.2019.105305.

Liu, F., Cai, M., Wang, L., Lu, Y., 2019. An ensemble model based on adaptive noise reducer and over-fitting prevention LSTM for multivariate time series forecasting. IEEE Access. 7, 26102–26115. https://doi.org/10.1109/ACCESS.2019.2900371.

Liu, H., Liu, J., Liu, Y., Ouyang, B., Xiang, S., Yi, K., Tao, S., 2020. Analysis of wintertime O_3 variability using a random forest model and high-frequency observations in Zhangjiakou - an area with background pollution level of the North China plain. Environ. Pollut. 262, 114191. https://doi.org/10.1016/j.envpol.2020.114191.

Ma, J., Ding, Y., Gan, V.J.L., Lin, C., Wan, Z., 2019. Spatiotemporal prediction of $PM_{2.5}$ concentrations at different time granularities using IDW-BLSTM. IEEE Access. 7, 107897–107907. https://doi.org/10.1109/ACCESS.2019.2932445.

Masih, A., 2019. Machine learning algorithms in air quality modeling. G. J. environ. Sci. Manag 4, 515–534. https://doi.org/10.22034/gjesm.2019.04.10.

Meng, X., Hand, J.L., Schichtel, B.A., Liu, Y., 2018. Space-time trends of $PM_{2.5}$ constituents in the conterminous United States estimated by a machine learning approach, 20052015. Environ. Int. 121, 1137–1147. https://doi.org/10.1016/j.envint.2018.10.029.

Number of Population from Registration by Age Group Province and Region: 2011–2020. Thailand. http://statbbi.nso.go.th/staticreport/page/sector/en/01.aspx (accessed 1 May 2021).

Papula, L., 1999. Mathematik fuer Ingenieur und Naturwissenschaftler Band 3, third ed. Vieweg, Braunschweig/Wiesbaden.

Qiao, W., Tian, W., Tian, Y., Yang, Q., Wang, Y., Zhang, J., 2019. The forecasting of $PM_{2.5}$ using a hybrid model based on wavelet transform and an improved deep learning algorithm. IEEE Access. 7, 142814–142825. https://doi.org/10.1109/ACCESS.2019.2944755.

Requia, W.J., Jhun, I., Coull, B.A., Koutrakis, P., 2019. Climate impact on ambient $PM_{2.5}$ elemental concentration in the United States: a trend analysis over the last 30years. Environ. Int. 131, 104888. https://doi.org/10.1016/j.envint.2019.05.082.

Scikit-learn, 2020. Machine Learning in Python. https://scikit-learn.org/stable/ (accessed 10 August 2020).

Soh, P.W., Chang, J.W., Huang, J.W., 2018. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access. 6, 38186–38199. https://doi.org/10.1109/ACCESS.2018.2849820.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., Hoogh, K.D., Donato, F.D., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., Schwartz, J., 2019. Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 20132015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179. https://doi.org/10.1016/j.envint.2019.01.016.

Tao, Q., Liu, F., Li, Y., Sidorov, D., 2019. Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. IEEE Access. 7, 76690–76698. https://doi.org/10.1109/ACCESS.2019.2921578.

Wang, Y., Hu, X., Chang, H., Waller, L., Belle, J., Liu, Y., 2018. A Bayesian downscaler model to estimate daily $PM_{2.5}$ levels in the continental US. Int. J. Environ. Res. Public Health 9, 1–14. https://doi.org/10.3390/ijerph15091999.

Wang, B., Kong, W., Guan, H., Xiong, N.N., 2019. Air quality forecasting based on gated recurrent long short term memory model in internet of things. IEEE Access. 7, 69524–69534. https://doi.org/10.1109/ACCESS.2019.2917277.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. Spatiotemporal continuous estimates of $PM_{2.5}$ concentrations in China, 20002016: a machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357. https://doi.org/10.1016/j.envint.2018.11.075.

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q., Huang, L., 2019. A predictive data feature exploration-based air quality prediction approach. IEEE Access. 7, 30732–30743. https://doi.org/10.1109/ACCESS.2019.2897754.

Zhao, C., Wang, Q., Ban, J., Liu, Z., Zhang, Y., Ma, R., Li, S., Li, T., 2019. Estimating the daily $PM_{2.5}$ concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01řŒ0.01ř spatial resolution. Environ. Int. 134, 105297. https://doi.org/10.1016/j.envint.2019.105297.

Zhong, M., Chen, F., Saikawa, E., 2019. Sensitivity of projected $PM_{2.5}$ - and O_3 -related health impacts to model inputs: a case study in mainland China. Environ. Int. 123, 256–264. https://doi.org/10.1016/j.envint.2018.12.002.