OPEN

Combination of data-driven models and interpolation technique to develop of PM₁₀ map for Hanoi, Vietnam

Dung Anh Nguyen¹, Son Hong Duong², Phuong Anh Tran^{2,3}, Hai Hoang Cao^{2,3} & Bang Quoc Ho^{4,5}✉

The degradation of air quality is the most concerned issue of our society due to its harmful impacts on human health, especially in cities with rapid urbanization and population growth like Hanoi, the capital of Vietnam. This study aims at developing a new approach that combines data-driven models and interpolation technique to develop the PM₁₀ concentration maps from meteorological factors for the central area of Hanoi. Data-driven models that relate the PM₁₀ concentration with the meteorological factors at the air quality monitoring stations in the study area were developed using the Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) algorithms. Models' performance comparison showed that ANN models yielded better goodness-of-fit indices than MLR models at all stations in the study area with average coefficient of correlation (*r*) and Nash–Sutcliffe Efficiency Index (*NSE*) of 0.51 and 0.34 for the former, and 0.7 and 0.49 for the latter. These indices indicate that the ANN-based data-driven models outperformed the MLR-based models. Thus, the ANN-based models and the Inverse Distance Weighting (IDW) interpolation technique were then combined for mapping the monthly PM₁₀ concentration with a spatial resolution of 1 km from global meteorological data. With this combination, the PM₁₀ concentration maps account for both local PM₁₀ concentration and impacts of spatio-temporal variations of meteorological factors on the PM₁₀ concentration. This study provides a promising method to predict the PM concentration with a high spatio-temporal resolution from meteorological data.

Air pollution is one of the most concerned issues of our society due to its impacts on human health and the economy^{1,2}. Many studies have shown that the mortality rate increases with increasing particulate matter (PM) concentration in the air. For example, Dockery et al.³ reported that an increase of 10 µg/m³ in the PM_{2.5} concentration would make the death rate increase by 1.5% in 6 cities in the United States⁴. A high amount of PM in the air also showed a strong correlation with lung cancer and other cardiovascular and lung related diseases which makes it one of the known causes for deaths in adults in the United States⁵. An increased concentration of 10 µg/m³ PM₁₀ would result in a 0.5% increase in the number of deaths per day in 20 big cities with more than 50 million inhabitants^{6,7}. A similar outcome was also confirmed by Katsouyanni et al.⁸ in a study of over 29 cities. Not only does it have negative effects on human health, PM air pollution also causes huge economic loss. Because of air pollution, each year around 4.1 billion Euro loss was reported in Switzerland⁹ and a loss of 40–50 million USD was revealed by a statistic made by The National Center for Health Statistics of the United States in 2001¹⁰. According to the World Bank, in 2016 air pollution cost the world's economy around 5.11 trillion USD in welfare losses. In South Asia and East Asia and the Pacific, losses were around 7.4 percent and 7.5 percent of the regional gross domestic product (GDP), respectively¹¹. It has been long recognized that meteorological

¹Department of Science and Technology, Ministry of Natural Resources and Environment, 10 Ton That Thuyet Street, My Dinh 2 Ward, Nam Tu Liem District, Hanoi City, Vietnam. ²Water Resources Institute, 8 Phao Dai Lang Street, Lang Thuong Ward, Dong Da District, Hanoi City, Vietnam. ³Department of Water Resources Engineering and Technology, Water Resources Institute, 8 Phao Dai Lang Street, Lang Thuong Ward, Dong Da District, Hanoi City, Vietnam. ⁴Air Pollution and Climate Change Research Center (APAC), Institute for Environment and Resources (IER), 142 To Hien Thanh, Ward 14, District 10, Ho Chi Minh City, Vietnam. ⁵Vietnam National University-Ho Chi Minh City (VNU-HCM), Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam. ✉email: bangquoc@yahoo.com

conditions and PM concentration have a close relationship. Zhao et al.¹², in a study on PM_{2.5} pollution from 2005 to 2007, highlighted a clear seasonal variation in the concentration of PM_{10-2.5} in which the PM concentration at the rural area was at the minimum level in winter and maximum level in spring and summer, while the urban region experienced the opposite trend. This study also pointed out that precipitation had an important contribution to the seasonal pattern of PM_{2.5} in the urban area, while monsoon was the main factor for that in the countryside. In a similar study, Duo et al.¹³ concluded that temperature in Lhasa, Tibet was the dominant factor governing all air pollutants including PM₁₀ and PM_{2.5} in spring, while relative humidity and atmospheric pressure were the major meteorological drivers during summertime. Spatial distribution of fine to coarse PM showed an inverse relationship with wind speed^{14,15}. Srimuruganandam and Nagendra¹⁶ revealed that low wind speed was highly correlated with PM concentration. It was however reported by Giri et al.¹⁷ that PM₁₀ concentration in Kathmandu, Nepal increased with wind speed and atmospheric pressure. Wang and Ogawa¹⁸ also reported that PM_{2.5} was positively proportional to wind speed higher than 3 m/s, and negatively proportional to the wind speed lower than that level.

Based on the relationship between the PM₁₀ concentration and meteorological factors, efforts have been made to construct data-driven models for predicting the PM₁₀ concentration from meteorological data using different statistical methods¹⁹⁻²¹. Among these methods, the Multiple Linear Regression (MLR)-based data-driven models are the most popular and have been widely used. The MLR algorithm is usually used to formulate the linear relationship between meteorological factors (including temperature, relative humidity, wind speed, and wind direction) with the PM₁₀ or PM_{2.5} concentration. Although there are some studies that reported good prediction results²², it is generally seen that the MLR-based models are yet to present consistently satisfactory results due to the linearizing of the non-linear system as reviewed by Shahraiyini and Sodoudi²³. With the ability to represent complex non-linear problems, the ANN-based data-driven model with different architectures has been extensively used to estimate the PM concentration. Several studies reported that the Artificial Neural Network (ANN) models produced satisfactory prediction results²⁴. Although some studies showed that the ANN models performed better than the MLR models²⁴, the ANN models have a more complicated structure and still present some limitations in terms of handling high dimensional input variables, local minima or interpretability (the black-box model problem). As a result, for each case study, it is necessary to compare these two models to select the more suitable model.

PM₁₀ concentration recorded in air quality monitoring stations only cover an area surrounding those stations, thus it is necessary to use interpolation techniques for mapping the PM₁₀ concentration. There are multiple interpolation techniques for this purpose. Wong et al.²⁵ provided an excellent review on these techniques and divided them into four groups, namely spatial averaging, nearest neighbor, inverse distance weighting and kriging. These interpolation techniques have been successfully employed to construct PM₁₀ maps in many studies. For example, Perez²⁶ applied the nearest neighbor technique to provide a forecasting map of PM₁₀ in Santiago, Chile. Li et al.²⁷ developed two IDW-based spatiotemporal interpolation techniques to evaluate the spatial variation of PM_{2.5} concentration over the contiguous United States. Kim et al. employed the ordinary kriging to interpolate the PM₁₀ concentration from 226 urban-ambient monitoring sites in South Korea²⁸. Raja et al.²⁹ used spatiotemporal kriging with the external drift to explore spatio-temporal variations of PM₁₀ concentrations in Ankara, Turkey. The main drawback of these studies was that they only used the PM₁₀ concentration measured at the air quality stations for interpolation without considering the impacts of the spatio-temporal variations of the meteorological factors on PM₁₀ variation. As a result, it is crucial to develop an interpolation technique that can account for information from both air quality stations and meteorological data.

With the increasing population and rapid urbanization, air quality in Vietnam, especially in large cities like Hanoi has been significantly degraded. Hopke et al.³⁰ indicated that Hanoi was one of the cities which had the worst air quality in Asia. Saksena et al.³¹ showed that the average value of PM₁₀ concentrations in the streets in Hanoi could reach up to 455 µg/m³, which is much higher than the Vietnamese daily standard for the PM₁₀ concentration (150 µg/m³). This has posed a negative effect on the city's public health. It was reported that ambient and in-house air pollution was becoming the major reason for deaths related to the environment in Vietnam, just second to smoking³². As a result, there has been increasing attention and demand from both the local community and the government of Hanoi for a study on air quality and its controlling factors with PM_{10-2.5} concentration prediction being the top priority.

There are two objectives to this study. The first objective is to develop a hybrid mapping approach that combines the data-driven model and IDW interpolation to produce the PM₁₀ concentration maps from global meteorological data. The second objective is to employ this approach to construct the monthly PM₁₀ maps for the central districts of Hanoi and analyze its spatio-temporal variations.

Methodology and material

In this study, we developed a hybrid approach that combines data-driven models and interpolation techniques to construct the PM₁₀ concentration maps from global meteorological data. As shown in Fig. 1, this approach consists of two main steps, namely, (1) development of data-driven models at each air quality monitoring station and (2) construction of PM₁₀ maps from global meteorological data using the IDW interpolation technique. Details of these two steps are presented below.

Development of data-driven models. Before developing data-driven models, a set of input features derived from meteorological factors were constructed. After that, the data-driven models linking these selected input features and the PM₁₀ concentration using two machine-learning algorithms, namely, MLR and ANN were developed. Finally, based on their performance, the more accurate models were selected for mapping the PM₁₀ concentration.

Figure 1. Hybrid data-driven and interpolation scheme.

Construction of input features. Multiple meteorological factors can be considered in data-driven models to predict the PM_{10} concentration. However, based on the availability of meteorological data at the air quality monitoring stations and their correlation with the PM_{10} concentration, the following variables were taken into account: mean daily temperature, maximum daily temperature, minimum daily temperature, mean daily humidity, mean daily wind speed and mean daily atmospheric pressure. Next, we assumed that the PM_{10} concentration was linked to meteorological factors by a quadratic function as follows:

$$PM_{10} = f(X_i, X_i^2, X_i X_j) \quad i, j = 1, 2, \dots, 6, \quad i \neq j \quad (1)$$

in which X_i ($i = 1, 2, \dots, 6$) is the mean daily pressure (X_1), mean daily temperature (X_2), mean daily humidity (X_3), mean daily wind speed (X_4), maximum daily temperature (X_5), minimum daily temperature (X_6). Totally, there are 27 features considered in Eq. (1). Since the size of features is considerably large, features with low correlation coefficients with the PM_{10} concentration or close correlation with other previously-selected features were removed from the equation.

Next, the selected input features and PM_{10} concentration were standardized to avoid the effects of differences in scale of features that could significantly influence the performance of regression models. After standardization, both input features and PM_{10} concentration were unitless values with a mean of 0 and a standard deviation of 1. The standardized features and PM_{10} concentration were used as inputs and outputs for the data-driven models.

Development of data-driven models. **Multiple linear regression model.** MLR is a statistical technique used to find a linear relationship between a response variable (dependent variable) and explanatory variables (independent variables). It is one of the most common methods to generalize the relationship of PM concentration with its determinants²³. Generally, the MLR model is defined by the following equation:

$$y_i = c_0 + c_1 x_{i1} + c_2 x_{i2} + \dots + c_n x_{in} + \epsilon \quad (2)$$

in which, y_i is dependent variable/response variable; x_i = independent/explanatory variables; c_0 is the intercept; c_n is the slope coefficient for each independent variables; ϵ is the error term. The y_i , in this study, is the standardized PM_{10} concentration, and x_i is the standardized features constructed in the previous step. At each air quality monitoring station, a data-driven model based on the MLR algorithm was developed from measurement data of PM_{10} concentration and meteorological-derived input features using the method of least squares.

Artificial neural network model. In this study, a feed forward neural network (FFNN) model, a common architecture of ANN model, was employed to build data-driven models for each air quality monitoring station using the same input and output data as in the MLR models for comparison. The structure of a FFNN model consists

of three layers (input layer, hidden layer and output layer). We referred to Sanger³³ for more detailed information about the FFNN algorithm.

In order to develop the FFNN data-driven models, the input and output data were randomly sampled into 3 sub-sets with 70% of data for training, 15% for validation and 15% for testing. Since the number of nodes in the input and output layers was determined, the determination of the FFNN structure focused on determining the number of hidden nodes. In this study, the trial-and-error method was used to find the number of hidden nodes for each air quality monitoring station.

Development of a hybrid interpolation approach for PM₁₀ concentration mapping. Based on the data-driven models developed in the previous section, this study constructed the monthly PM₁₀ concentration maps from meteorological data using a new approach based on the IDW interpolation method. In order to predict the PM concentration at a given location (interpolated location) from surrounding air quality monitoring stations, the IDW method determines the weighting factors of each station as below:

$$w_{ik} = \frac{1/d_{ik}^2}{\sum_{i=1}^N d_{ik}^2} \quad (3)$$

in which w_{ik} is the weighting factor of station i th at interpolated location k th. d_{ik}^2 is the distance from station i th to interpolated location k th. N is the total number of air quality monitoring stations used for interpolation. Using the weighting factors, the PM₁₀ concentration was estimated as below:

$$PM_{10}^k = w_{ik} f_i^{PM_{10}}(X_k) \quad (4)$$

in which $f_i^{PM_{10}}$ is the data-driven model developed for the station i th; X_k is the input feature vector which was derived from meteorological data at interpolated location k th. The novel of this approach is that instead of using the PM₁₀ concentration values at the air quality monitoring stations like in traditional approaches, it used meteorological data at the interpolated location to feed the data-driven models. This hybrid approach allows us to consider the impacts of both local conditions (via the data-driven model developed for each station) and spatio-temporal variations of meteorological factors.

All the above steps including feature construction, development of MLR and ANN-based models and mapping the PM₁₀ concentration were programmed on Matlab (coding of this program is provided in the Supplementary document). This programming language allows for quick implementation of the algorithms and easy visualization of the results without using any other additional software.

The method developed in this study can be well applied for other cities where meteorological and PM₁₀ concentration observations are available. However, the selection of meteorological factors and development of data-driven models in this study were purely relied on data from air quality monitoring stations in the city of Hanoi. As a result, the data-driven models may not be applicable to other cities. The data-driven models should be developed for each city based on the availability of observation data in that city.

Study area. Hanoi is the capital city of Vietnam with an area of 3358 km² (following the administrative expansion in 2008) and more than 7.4 million people in 2017³⁴. The study area consists of eleven central districts of Hanoi (Fig. 2). This is the most crowded area of Hanoi where 41% population of the city resides in 7.7% of the total city area. With a high population density and a large number of vehicles and construction activities, this area has been severely suffered from air pollution. As for the weather conditions, Hanoi has four distinct seasons including spring (March–May), summer (June–August), autumn (September–November) and winter (December–February). In winter, the weather is cold and dry, while summer has high humidity and rainfall³⁵. With a lower temperature and low humidity, the PM₁₀ concentration in winter is much higher than the other seasons (Fig. 3). While the 24 h PM₁₀ concentration is mostly below the National Technical Regulation on Ambient Air Quality (QCVN 05:2013/BTNMT, PM₁₀ = 150 µg/m³), there are many days in which PM₁₀ concentration is above this standard in winter. Due to the harmful impacts of high PM₁₀ concentration on human health and the importance of the study area, it is necessary to construct high resolution PM₁₀ concentration maps for this area in order to provide more detail air quality information for local residents who are most likely to be affected.

Data availability. The data used in this study was collected from two sources corresponding to two objectives. For the development of the data-driven models, input data was collected from 11 air quality monitoring stations located across the study area (Fig. 2). These stations include three fixed stations (Minh Khai, Trung Yen 3 and Nguyen Van Cu) and eight sensor stations (Table 1). Nguyen Van Cu station is under the management of the Vietnam Environment Administration. The remaining stations are under the management of the Hanoi Department of Environmental Protection. Hourly PM₁₀ concentration and meteorological data (atmospheric pressure, temperature, humidity, wind speed) at all stations from 01/06/2017 to 31/12/2018 were collected. This dataset covers one and a half year, and therefore, can represent the temporal variations of the PM₁₀ concentration and meteorological factors over four seasons of the year. In addition to PM₁₀, other air pollutants were also collected, although they were not considered in the scope of this study. The hourly PM₁₀ concentration and meteorological data were averaged to generate a daily dataset to reduce the measurement errors and remove their diurnal variation.

Since the data collected from 11 meteorological stations were limited, the PM₁₀ concentration calculated from these data was not representative for its spatial variation in the study area. Therefore, high spatial resolution maps of the PM₁₀ concentration were needed. For mapping the monthly PM₁₀ concentration, we used the

Figure 2. Location of 11 air quality stations used in the study.

Figure 3. Daily PM_{10} in 2017 in Trung Yen 3 air quality monitoring station.

global meteorological data from the WorldClim 2.0 database (<https://www.worldclim.org/>), which contains temperature (mean, maximum, minimum), precipitation, solar radiation, vapor pressure, and wind speed data with a spatial resolution of 1 km². This is a reliable data source that was validated with gauged data (correlation coefficient with gauged data $r \geq 0.99$ for temperature and vapor pressure, $r \geq 0.86$ for precipitation and $r \geq 0.76$ for wind speed). After the global meteorological data was downloaded, they were extracted for the region of the study area. Because the relative humidity was not available, it was calculated from actual and saturated vapor pressure. The atmospheric pressure was calculated from the location altitude and air temperature. Figure 4 shows the temperature, wind speed, relative humidity and air pressure in February obtained from the WorldClim database as an example. As shown in the figure, the data extracted from the WorldClim can well reflect the spatial variation of meteorological factors.

No	Monitoring station	Longitude	Latitude	Site
Fixed stations				
1	Minh Khai	21°02'57.8"	105°44'31.0"	Minh Khai Ward People's Committee
2	Trung Yen 3	21°00'55.1"	105°47'59.6"	Hanoi Department of Environmental Protection
3	Nguyen Van Cu	21°2'58.43"	105°52'55.8"	Northern Center for Environmental monitoring
Sensor stations				
4	Phạm Van Dong	21°03'04.3"	105°46'54.6"	Ha Noi Centre for Environmental and Natural Resources monitoring
5	Tay Mo	21°00'45.2"	105°44'48.1"	Tay Mo Ward People's Committee
6	My Dinh	21°01'10.5"	105°46'13.3"	My Dinh 1 Ward People's Committee
7	Hang Dau	21°02'25.7"	105°50'44.5"	Hang Ma District Police
8	Hoan Kiem	21°01'35.5"	105°51'11.9"	Hoan Kiem District Police
9	Kim Lien	21°00'24.8"	105°50'11.3"	Kim Lien Kindergarten
10	Thanh Cong	21°01'11.0"	105°48'53.0"	Thanh Cong Lake Park
11	Hoang Mai	20°59'18.0"	105°51'17.6"	Hoang Van Thụ Ward People's Committee

Table 1. Summary information of air quality monitoring stations.

Figure 4. Spatial distribution of Temperature, Humidity, Atmospheric Pressure, and Wind Speed derived from WorldClim 2.0 in January.

Results

Construction of input features for data-driven models. The meteorological data collected from 01/06/2017 to 31/12/2018 were used to construct the input features for the data-driven models. The total number of features considered in this study was 27. In order to reduce this number of features, the correlation coefficients between each feature with the PM_{10} concentration and between features were estimated. Figure 5 presents the correlation matrix which indicates the correlation coefficients of input features with each other and with the PM_{10} concentration. The figure shows that the correlation coefficients between the input features and the PM_{10} concentration range from -0.46 to 0.46 . All six meteorological factors (mean daily atmospheric pressure, mean daily humidity, mean daily wind speed, maximum daily temperature and minimum daily temperature) have a relatively high correlation with the PM_{10} concentration with absolute correlation coefficients greater than 0.24 . It is interesting that of these meteorological factors, only the mean daily pressure is positively correlated with the PM_{10} concentration, while the other factors have negative correlation coefficients. This implies that the PM_{10} concentration increases with increasing mean daily pressure and decreasing other factors. The features with the highest correlation with the PM_{10} concentration are the mean daily pressure (X_1) and its quadratic term (X_1^2) (correlation coefficients = 0.46), while the product of the mean daily pressure and the mean daily humidity (X_1X_3) has the lowest correlation (correlation coefficient = -0.22).

Figure 5. Coerrelation matrix of features and between features with the PM₁₀ concentration. X_1 is the mean daily pressure; X_2 is the mean daily temperature; X_3 is the mean daily humidity; X_4 is the mean daily wind speed; X_5 is the maximum daily temperature; X_6 is the minimum daily temperature.

In order to select input features for the data-driven models, we evaluated the correlation of each feature with the PM₁₀ concentration and with the other features. The mean daily pressure (X_1), mean daily temperature (X_2), mean daily humidity (X_3), mean daily wind speed (X_4) are well correlated with the PM₁₀ concentration and are independent from each other. As a result, they were added to the input features. The maximum (X_5) and minimum daily temperature (X_6) are well-correlated to the mean daily temperature with correlation coefficients of 0.98. Hence, these two features and their associated features (X_1X_5 , X_2X_5 , X_3X_5 , X_4X_5 , X_5X_5 , X_5X_6 ; X_1X_6 , X_2X_6 , X_3X_6 , X_4X_6 , X_5X_6 , X_6X_6) were not included in the input features. Features X_1X_1 , X_1X_2 , X_1X_3 , X_1X_4 , X_1X_5 and X_1X_6 that are functionally correlated with X_1 were not considered either. Of the features associated with the mean daily temperature (X_2) and mean daily humidity (X_3), only features X_2X_3 , X_2X_4 , and X_3X_4 are relatively independent on the others factors and have a high correlation with the PM₁₀ concentration. Therefore, these features were selected as inputs for the data-driven models. In total, the input features consist of X_1 , X_2 , X_3 , X_4 , X_2X_4 , and X_3X_4 .

Development of data-driven models. *Multiple linear regression model.* Using the input features selected in the previous section, the MLR-based model (Eq. 2) was written as below:

$$PM_{10} = c_1 + c_2X_1 + c_3X_2 + c_4X_3 + c_5X_4 + c_6X_2X_3 + c_7X_2X_4 + c_8X_3X_4 \quad (5)$$

in which the coefficient c_i ($i = 1 \dots 8$) was determined from measurement data using the method of least squares. Comparing with previous studies that usually used the MLR algorithm to construct the relationship between the PM₁₀ concentration and meteorological factors, this study considered both the meteorological factors (X_1 , X_2 , X_3 , X_4) and their combinative quadratic terms (X_2X_3 , X_2X_4 , X_3X_4), and therefore, could account for the nonlinear quadratic form of this relationship. In addition, to account for the seasonal dependence of the PM₁₀ concentration on meteorological factors, this study built two MLR models corresponding to the winter and spring period and the summer and autumn periods.

Artificial neuron network model. Although the MLR-based data-driven models developed in this study accounted for the nonlinear and seasonal relationship between the PM₁₀ concentration and meteorological factors, they only considered the quadratic nonlinear relationship. For comparison, we employed the ANN algorithm, which could formulate this relationship in a more complicated manner. The ANN model included three layers (input, hidden and output) in which the number of nodes in the input layer and output layer was equal to 7 (number of input features) and 1 (PM₁₀), respectively. After using the trial-and-error method, we found that the number of the hidden layers with 12 nodes was the most suitable for all air quality monitoring stations. Figure 6 illustrates the architecture of the ANN model used in this study.

In order to build the ANN model for each station, the measurement data were divided into three sub-datasets (70% for training, 15% for validation and 15% for testing) to avoid overfitting. Figure 7 compares the PM₁₀

Figure 6. Schematization of the ANN model structure. The model includes 3 layers: input (7 nodes), hidden (12 nodes) and output (1 node).

concentration between the ANN modeling and measurement in each sub-dataset and the whole dataset at the Trung Yen 3 station as an example. The figure shows that the ANN model could well simulate the PM_{10} concentration at all datasets, and therefore, could be reliably used for predicting PM_{10} concentration.

Comparison of model performance. To evaluate the performance of the MLR and ANN models, we used three statistical indices, namely, Root Mean Squared Error ($RMSE$), correlation coefficient (r) and Nash–Sutcliffe Efficiency (NSE), which are formulated as below:

$$r = \frac{\sum_{i=1}^N (Y_o^i - \bar{Y}_o)(Y_m^i - \bar{Y}_m)}{\sqrt{\sum_{i=1}^N (Y_o^i - \bar{Y}_o)^2} \sqrt{\sum_{i=1}^N (Y_m^i - \bar{Y}_m)^2}} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_m^i - Y_o^i)^2} \quad (7)$$

$$NSE = 1 - \frac{\sum_{i=1}^N (Y_m^i - Y_o^i)^2}{\sum_{i=1}^N (Y_o^i - \bar{Y}_o)^2} \quad (8)$$

where N is total number of data points; Y_m is the modeled data, Y_o is the observed data. The correlation coefficient ranges from -1 to 1 in which the higher value corresponds to the closer positive relationship between the modeling and measurement. $RMSE$ measures the differences between modeling and measurement. The lower $RMSE$ indicates a better agreement between the modeling and measurement. NSE varies in the range from $-∞$ to 1 in which $NSE = 1$ indicates a perfect match between the modeling and measurement, $NSE ≤ 0$ implies that the model predictions have the same or lower accuracy than the mean of measurements.

Table 2 compares the indices for both data-driven models. The table indicates that the ANN models performed much better than the MLR models in all eleven stations. The ANN outputs are well correlated with the measurement data with an average correlation coefficient of 0.7 . No station has a correlation coefficient below 0.65 . Meanwhile, the average correlation coefficient of the MLR outputs with measurements is 0.58 in which the Hoan Kiem station has the lowest coefficient ($r = 0.51$). As for the modeling errors, the average $RMSE$ of the ANN and MLR models are 14.1 and 15.6 , respectively. The NSE criterion ranges from 0.41 to 0.57 for the ANN models and from 0.26 to 0.53 for the MLR models. It is clear that the differences between modeling and measurement of the ANN models are lower than those of the MLR models. The reason for this fact is that the ANN algorithm

Figure 7. ANN models results for Trung Yen 3 air quality monitoring station.

No	Station	ANN model			MLR model		
		RMSE	<i>r</i>	NSE	RMSE	<i>r</i>	NSE
1	Hang Dau	15.86	0.67	0.45	17.15	0.59	0.35
2	Hoan Kiem	9.30	0.65	0.42	10.53	0.51	0.26
3	Kim Lien	9.43	0.70	0.49	10.62	0.60	0.36
4	My Dinh	10.69	0.69	0.47	12.09	0.57	0.32
5	Tuong Mai	9.14	0.69	0.46	10.64	0.52	0.27
6	Thanh Cong	12.49	0.68	0.47	13.91	0.58	0.34
7	Tay Mo	11.98	0.65	0.41	13.37	0.52	0.27
8	Minh Khai	24.03	0.72	0.51	26.08	0.60	0.36
9	Trung Yen 3	15.69	0.75	0.56	16.24	0.73	0.53
10	Nguyen Van Cu	15.72	0.75	0.57	19.21	0.60	0.35
11	Pham Van Dong	20.19	0.73	0.53	21.96	0.61	0.37

Table 2. Performance comparison between ANN model and MLR model.

accounts for more complicated interactions between the input and output than the MLR model. For their better performance, the ANN models were employed for mapping the PM_{10} concentration.

Of all air quality monitoring stations, the performance of both MLR and ANN at the Trung Yen 3 3, Nguyen Van Cu and Minh Khai stations are much better than the others. Indeed, the average correlation coefficient and NSE corresponding to the ANN models for the three stations are respectively equal to 0.74 and 0.55, while these indices for the other stations are much lower (0.68 for the correlation coefficient and 0.46 for the NSE). This could be explained by the fact that Trung Yen 3, Nguyen Van Cu, Minh Khai are the main stations of Hanoi, which have been frequently checked and performed quality control. As a result, the quality of measurement data at these stations is better than the others.

Monthly PM_{10} concentration mapping. Mapping the monthly PM_{10} concentration for the study area of eleven central districts in Hanoi from the WorldClim global meteorological data was performed using the ANN models developed in the previous section and the hybrid interpolation approach (Eq. 4). Figures 8 and 9 below presents the monthly and seasonal maps of the PM_{10} concentration. The spatial resolution of these maps is equal to that of meteorological data (1 km²). The seasonal PM_{10} concentration maps were generated by assembling monthly maps.

As for the temporal variation of the PM_{10} concentration, it can be seen that the PM_{10} concentration reaches its peaks in the winter season (January, February, November and December). For example, the mean PM_{10} concentration in November is up to 71 $\mu\text{g}/\text{m}^3$. The low temperature (ranging from 16 to 21 °C in these months) and the temperature inversion phenomenon in winter are likely the causes of the high PM_{10} concentration in these months. By contrast, because the air temperature in June–August reaches its highest level (~29 °C), the concentration of PM_{10} hits a trough in this period with the PM_{10} concentration ranging from 46 to 48 $\mu\text{g}/\text{m}^3$. As regards the seasonal variation, points out that the PM_{10} concentration is highest in winter and lowest in summer. In addition to the air temperature, the humidity and wind speed, which are lowest in winter, are also the reason for the higher PM_{10} concentration in winter than in the other seasons.

As for the spatial variation, the concentration of PM_{10} in Long Bien district which is situated in the northeast of the study area is much lower than the other districts. The main reason is that compared to the other districts in the study area, the density of population in this district is lowest in the study area (around 4.5 thousand people/km², versus 11.6 thousand people/km² in other districts). This lower population density leads to less intensive traffic. Besides, due to the sparse air quality monitoring network, the PM_{10} concentration in Long Bien district strongly depends on the PM_{10} concentration at the Nguyen Van Cu station, which situates relatively far away from transportation routes. On the other hand, as pointed out by Nghiem et al.³⁶, since the inauguration of Vinh Tuy Bridge in 2010 and Nhat Tan Bridge in 2015, the flow of traffic vehicles through Nguyen Van Cu road was decreased. As a result, the annual average of PM_{10} concentration at Nguyen Van Cu station from 2010 to 2018 slightly declined which makes the PM_{10} concentration lower. On the contrary, the highest concentration is found at the Pham Van Dong station located in the southwest of the study area. This station is placed on the Pham Van Dong Street, one of the main route to access Hanoi from Noi Bai International Airport, thus the traffic in this street is normally quite intensive. Besides, a high number of active construction works in this area might be an important factor for the increased level of PM_{10} . Meteorological conditions also influence the spatial distribution of the PM_{10} concentration. The highest PM_{10} concentration in the northwest region is partly caused by the low air temperature in this region (see Fig. 4). However, Fig. 8 shows that the impact of local factors (e.g., street, population, transportation intensity) on the spatial variation of the PM_{10} concentration is larger than that of meteorological factors.

Conclusion

In this study, a combinative approach of data-driven models and IDW interpolation technique was developed to construct the PM_{10} concentration maps for the central area of Hanoi. The construction of data-driven models consisted of two steps, feature construction and model development. The feature construction is responsible for constructing optimal features from meteorological factors. By evaluating the correlation between the PM_{10} concentration with each feature and correlation between features, a set of features was selected as the input for the data-driven models. The model development step built the data-driven models that link the PM_{10} concentration with the input features using the MLR and ANN algorithms for each air quality monitoring station. The obtained results indicate that the ANN-based data-driven models provided much better results than the MLR-based models. In order to construct the PM_{10} concentration maps, the IDW interpolation technique was used to calculate the weighting factors for each air quality monitoring stations. While many other studies obtained the unknown PM_{10} concentration by interpolating the PM_{10} concentration at the air quality monitoring stations without considering meteorological factors, this study accounted for the meteorological factors in the data-driven models. Using this approach, both the local PM_{10} surrounding monitoring stations and the dependence of PM_{10} on meteorological factors were taken into account hence provided a better representation of the current situation in the study area.

Due to a lack of high spatial resolution of meteorological data, this study used the 1 km² resolution monthly WorldClim data as the input to predict monthly PM_{10} concentration via combination of the established data-driven models and interpolation method. The monthly PM_{10} maps were then aggregated to construct seasonal maps. The temporal analysis revealed that the PM_{10} concentration was highest in the winter months and lowest in the summer months, which was mainly caused by the negative dependence of the PM_{10} concentration on air temperature and low humidity. The spatial analysis indicated that the northeast region was the region with the lowest PM_{10} concentration because the urbanization in this region was less developed than the others. The northwest region had the highest PM_{10} concentration because of the high population and ongoing constructions

Figure 8. Monthly maps of PM_{10} concentration.

Figure 9. Seasonal maps of the PM₁₀ concentration. The red-solid lines represent the administrative boundary of eleven districts.

of new buildings and roads, which together elevated the PM₁₀ concentration. The meteorological factors also influenced the spatial variation of the PM₁₀ concentration but with a lower impact level compared to the local sources of PM₁₀ generation surrounding the monitoring stations. This study also pointed out that although the spatial variation of meteorological factors was taken into account, the low density of air quality monitoring stations might reduce the accuracy of PM₁₀ concentration maps. Hence, it is necessary to establish a denser air quality monitoring stations network to better cover the spatial variation of the PM₁₀ concentration. The approach developed in this study can be applied to provide the forecasting PM₁₀ concentration maps based on predicting meteorological information. These results could also provide a very meaningful foundation for the local authority in deriving and implementing city air quality management activities and urban planning in Hanoi.

Received: 11 June 2020; Accepted: 28 September 2020
Published online: 06 November 2020

References

1. Pope, C. A. III., Ezzati, M. & Dockery, D. W. Fine particulate air pollution and life expectancy in the United States. *N. Engl. J. Med.* **360**, 376–386 (2009).
2. Kunzli, N., Perez, L. & Rapp, R. *Air Quality and Health* (ERS, Lausanne, 2010).
3. Dockery, D. W. *et al.* An association between air pollution and mortality in six United-States cities. *N. Engl. J. Med.* **329**, 1753–1759 (1993).
4. Laden, F., Schwartz, J., Speizer, F. E. & Dockery, D. W. Air pollution and mortality: a continued follow-up in the Harvard six cities study. *Epidemiology* **12**, S81 (2001).
5. Pope, C. A. *et al.* Particulate air pollution as a predictor of mortality in a prospective study of United-States adults. *Am. J. Respir. Crit. Care Med.* **151**, 669–674 (1995).
6. Samet, J. M., Dominici, F., Currier, I., Coursac, I. & Zeger, S. L. Fine Particulate air pollution and mortality in 20 US cities, 1987–1994. *N. Engl. J. Med.* **343**(24), 1742–1749 (2000).
7. Samet, J. M. *et al.* The National Morbidity, Mortality, and Air Pollution Study. Part II: morbidity and mortality from air pollution in the United States. *Res. Rep. Health Effects Inst.* **94**, 5–70 (2000).
8. Katsouyanni, K. *et al.* Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: Results from time series data from the APHEA project. *BMJ* **314**, 1658–1663 (1997).
9. Seethaler, R. Health Costs due to Road Traffic-related Air Pollution, Air Pollution Attributable Cases. An impact assessment project of Austria, France and Switzerland, Prepared for the WHO-Ministerial Conference on Environment and Health, London, 16–18 June 1999. Synthesis. Berne, Paris (1999).
10. Fuchs, V. R. & Frank, S. R. Air pollution and medical care use by older Americans: a cross-area analysis. *Health Affairs (Millwood)* **21**(6), 207–214 (2002).
11. World Bank. The cost of air pollution. Strengthening the Economic Case for Action. The World Bank and Institute for Health Metrics and Evaluation, University of Washington, Seattle (2016).
12. Zhao, X. *et al.* Seasonal and diurnal variations of ambient PM_{2.5} concentration in urban and rural environments in Beijing. *Atmos. Environ.* **43**, 2893–2900 (2009).
13. Duo, B. *et al.* Observations of atmospheric pollutants at Lhasa during 2014–2015: pollution status and the influence of meteorological factors. *J. Environ. Sci.* **63**, 28–42 (2018).

14. Li, X., Ma, Y., Wang, Y., Liu, N. & Hong, Y. Temporal and spatial analyses of particulate matter (PM₁₀ and PM_{2.5}) and its relationship with meteorological parameters over an urban city in northeast China. *Atmos. Res.* **198**, 185–193 (2017).
15. Hartog, J. J. *et al.* Relationship between different size classes of particulate matter and meteorology in three European cities. *J. Environ. Monit.* **7**, 302–310 (2005).
16. Srimuruganandam, B. & Nagendra, S. Impact of meteorology on roadside ambient particulate matter concentrations. *Mod. Traffic Transp. Eng. Res.* **2**(3), 141–152 (2013).
17. Giri, D., Krishna Murthy, V. & Adhikary, P. R. The influence of meteorological conditions on PM₁₀ concentrations in Kathmandu Valley. *Int. J. Environ. Res.* **2**(1), 49–60 (2003).
18. Wang, J. & Ogawa, S. Effects of meteorological conditions on PM_{2.5} concentrations in Nagasaki, Japan. *Int. J. Environ. Res. Public Health* **12**, 9089–9101 (2015).
19. Liu, Z. *et al.* Seasonal and diurnal variation in particulate matter (PM₁₀ and PM_{2.5}) at an urban site of Beijing: analyses from a 9-year study. *Environ. Sci. Pollut. Res.* **22**, 627–642 (2015).
20. Koutrakis, P. *et al.* Analysis of PM₁₀, PM_{2.5}, and PM_{2.5–10} concentrations in Santiago, Chile, from 1989 to 2001. *J. Air Waste Manag. Assoc.* **55**, 342–351 (2005).
21. Clements, N., Hannigan, M. P., Miller, S. L., Peel, J. L. & Milford, J. B. Comparisons of urban and rural PM_{10–2.5} and PM_{2.5} mass concentrations and semi-volatile fractions in northeastern Colorado. *Atmos. Chem. Phys.* **16**, 7469–7484 (2016).
22. Chellali, M., Abderrahim, H., Hamou, A., Nebatti, A. & Janovec, J. Artificial neural network models for prediction of daily fine particulate matter concentrations in Algiers. *Environ. Sci. Pollut. Res.* **23**, 14008–14017 (2016).
23. Shahraiyini, H. T. & Sodoudi, S. Statistical modeling approaches for PM₁₀ prediction in urban areas: a review of 21st-century studies. *Atmosphere* **7**(2), 15 (2016).
24. Chaloulakou, A., Kassomenos, P., Spyrellis, N., Demokritou, P. & Koutrakis, P. Measurements of PM₁₀ and PM_{2.5} particle concentrations in Athens, Greece. *Atmos. Environ.* **37**(5), 649–660 (2003).
25. Wong, D. W., Yuan, L. & Perlin, S. A. Comparison of spatial interpolation methods for the estimation of air quality data. *J. Exposure Sci. Environ. Epidemiol.* **14**(5), 404–415 (2004).
26. Perez, P. Combined model for PM₁₀ forecasting in a large city. *Atmos. Environ.* **60**, 271–276 (2012).
27. Li, L., Losser, T., Yorke, C. & Piltner, R. Fast inverse distance weighting-based spatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter PM_{2.5} in the contiguous us using parallel programming and kd-tree. *Int. J. Environ. Res. Public Health* **11**(9), 9101–9141 (2014).
28. Kim, S. Y. *et al.* Ordinary kriging approach to predicting long-term particulate matter concentrations in seven major Korean cities. *Environ. Health Toxicol.* **29**, e2014012 (2014).
29. Raja, N. B., Aydin, O., Turkoglu, N. & Cicek, I. Characterising the seasonal variations and spatial distribution of ambient PM₁₀ in Urban Ankara, Turkey. *Environ. Process.* **5**(2), 349–362 (2018).
30. Hopke, P. K. *et al.* Urban air quality in the Asian region. *Sci. Total Environ.* **404**(1), 103–112 (2008).
31. Saksena, S., Quang, T. N., Nguyen, T., Dang, P. N. & Flachsart, P. Commuters' exposure to particulate matter and carbon monoxide in Hanoi, Vietnam. *Transp. Res. Part D Transp. Environ.* **13**(3), 206–211 (2008).
32. Global Burden of Disease (GBD). Visualizations 2013. Institute for Health Metrics and Evaluation (2013).
33. Sanger, T. D. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Netw.* **2**(6), 459–473 (1989).
34. General Statistics Office of Vietnam (GSO). Vietnam Statistics Yearbook in 2017 (2018).
35. Hien, P. D., Bac, V. T., Tham, H. C., Nhan, D. D. & Vinh, L. D. Influence of meteorological conditions on PM_{2.5} and PM_{2.5–10} concentrations during the monsoon season in Hanoi, Vietnam. *Atmos. Environ.* **36**, 3473–3484 (2002).
36. Nghiem, T.-D., Mac, D.-H., Nguyen, A.-D. & Le, N. C. An integrated approach for analyzing air quality monitoring data: a case study in Hanoi, Vietnam. *Air Qual. Atmos. Health* <https://doi.org/10.1007/s11869-020-00907-6> (2020).

Acknowledgements

The authors gratefully acknowledge Vietnam Environment Administration and Hanoi Department of Natural Resources and Environment for the data providing Irish Research Council COALESCE Research Fund 2019—IRC-COALESCE-2020-31 under HealthyAIR project.

Author contributions

N.A.D. contributes as coordinator for whole research and develop air pollution map. D.H.S. contributes to develop methodology for this paper. T.A.P. contributes to analysis data for this paper. C.H.H. contributes to analysis meteorological data for this paper. H.Q.B. contributes to prepare the manuscript and develop air pollution map.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-75547-y>.

Correspondence and requests for materials should be addressed to B.Q.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020