ISSN: 1648-6897 / eISSN: 1822-4199

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of ENVIRONMENTAL ENGINEERING & LANDSCAPE MANAGEMENT

2024
Volume 32
Issue 4
Pages 292–304
https://doi.org/10.3846/jeelm.2024.22361

STATISTICAL AND MACHINE LEARNING APPROACHES FOR ESTIMATING POLLUTION OF FINE PARTICULATE MATTER ($PM_{2.5}$) IN VIETNAM

Tuyet Nam Thi NGUYEN[1]✉, Tan Dat TRINH[2], Pham Cung Le Thien VU[2], Pham The BAO[2]

[1]*Faculty of Environment, Saigon University, Ho Chi Minh, Vietnam*
[2]*Faculty of Information Science, Saigon University, Ho Chi Minh, Vietnam*

Highlights:

- $PM_{2.5}$ pollution was predicted by ARIMA, machine learning and deep learning models;
- the best model was CNN+Bi-LSTM with the lowest prediction errors and the highest R^2;
- LSTM and Bi-LSTM models showed a similar performance, with comparable errors and R^2;
- ARIMA had the worst performance due to information loss during differencing data;
- the predicted air quality indexes for $PM_{2.5}$ matched the observed ones up to 96%.

Article History:

- received 22 April 2024
- accepted 11 September 2024

Abstract. This study aims to predict fine particulate matter ($PM_{2.5}$) pollution in Ho Chi Minh City, Vietnam, using autoregressive integrated moving average (ARIMA), linear regression (LR), random forest (RF), long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM), and convolutional neural network (CNN) combining Bi-LSTM (CNN+Bi-LSTM). Two experiments were set up: the first one used data from 2018–2020 and 2021 as training and test data, respectively. Data from 2018–2021 and 2022 were used as training and test data for the second experiment, respectively. Consequently, ARIMA showed the worst performance, while CNN+Bi-LSTM achieved the best accuracy, with an R^2 of 0.70 and MAE, MSE, RMSE, and MAPE of 5.37, 65.4, 8.08 μg/m³, and 29%, respectively. Additionally, predicted air quality indexes (AQIs) of $PM_{2.5}$ were matched the observed ones up to 96%, reflecting the application of predicted concentrations for AQI computation. Our study highlights the effectiveness of machine learning model in monitoring of air pollution.

Keywords: $PM_{2.5}$, machine learning, ARIMA, univariate time series, Ho Chi Minh City.

✉Corresponding author. E-mail: ***ntnam@sgu.edu.vn***

1. Introduction

Fine particulate matter ($PM_{2.5}$) is atmospheric particles which have diameter less than 2.5 μm. $PM_{2.5}$ has been regarded as one of the criteria air pollutants because of toxicity and tiny size of $PM_{2.5}$, facilitating its transport into the human respiratory system (Filonchyk et al., 2017). Exposure to $PM_{2.5}$ pollution can cause several symptoms, including nose or throat irritation, coughing, asthma, and lung diseases (Nguyen et al., 2023a; Chlebnikovas et al., 2023). It is, therefore, suggested that $PM_{2.5}$ should be frequently monitored for human health protection. The air quality related to $PM_{2.5}$ can be reported as the air quality index (AQI), which has descriptions, for instance, of good or unhealthy air quality, and health advice for public related to several pollution levels (Vietnam Environment Administration [VEA], 2019).

To monitor $PM_{2.5}$ concentrations, real-time monitoring using sensors or specialized instruments can be employed to continuously observe mass concentrations of $PM_{2.5}$ in the atmosphere. Apart from monitoring, it is also essential to predict $PM_{2.5}$ concentrations to support decision-making related to reducing $PM_{2.5}$ pollution and its adverse influence on human health. Recently, statistical models, such as autoregressive models, and machine learning or deep learning models, have been increasingly applied to estimate concentrations of air pollutants, including $PM_{2.5}$ (Clark et al., 2024; Upadhya et al., 2024; Harishkumar et al., 2020; Wang et al., 2017; Tong et al., 2019). In this regard, the input to these models is frequently a time series of variables representing a sequence of values over time. These variables include pollutant concentrations and meteorological parameters, such as wind speed, ambient air temperature, and rainfall levels. Additionally, the time series of variables can be classified as a univariate or multivariate series depending on the number of variables. The former and latter consist of single (e.g., pollutant concentrations) and multiple variables (e.g., pollutant concentrations and meteorological data), respectively. Due to the simultaneous consideration of variable relations, multivariate time

series data may yield more accurate predictions. However, univariate time series data are recommended for prediction, especially when multiple variables are not available.

Several statistical and machine learning (ML) models have been widely used to predict concentrations of air pollutants, such as $PM_{2.5}$, based on time series data (Filonchyk et al., 2018). Some of these models include an autoregressive integrated moving average (ARIMA) (Wang et al., 2017; Bhatti et al., 2021), one of the most widely used statistical approaches for univariate time-series forecasting, linear regression (LR) (Zhao et al., 2018), random forest (RF) (Kumari & Singh, 2023; Xu et al., 2020), support vector regression (SVR) (Wang et al., 2017), and ensemble models combining multiple ML model (Ejohwomu et al., 2022). In addition, deep learning (DL) approach, a subset of ML, has been increasingly applied for the prediction of air pollution because DL models can outperform some traditional ML models (Zamani Joharestani et al., 2019; Nath et al., 2021) due to their ability to learn the large and complex data. Particularly, some DL models specialized for time series data, such as long short-term memory networks (LSTM) (Sherstinsky, 2020; Rakholia et al., 2022; Barthwal & Goel, 2024), bidirectional LSTM (Bi-LSTM) (Tong et al., 2019), and convolutional neural network (CNN) (Rabie et al., 2024), have been widely selected to predict $PM_{2.5}$ concentrations because these models have memory cells storing long-term information (e.g., variation trends of concentrations), thus contributing to the outperformance of DL in comparison to statistical and ML models (Wu et al., 2021).

To develop a predictive model using ML and/or DL techniques, acquiring historical data is essential. This data serves as the training set for model learning and is also used to predict new as well as unseen data. Consequently, the performance of ML and DL models heavily depend on the characteristics of the dataset. Additionally, tuning the model's architecture may be necessary to optimize performance for specific data types, reflecting the dependencies on the data's statistical properties, such as trends and seasonality (Bontempi et al., 2013). The practical value of using ML and DL models to predict $PM_{2.5}$ concentrations lies in their ability to provide accurate and real-time forecasts. These models can offer hourly updates based on recent data, enabling air quality predictions up to an hour in advance (Bai & Li, 2023; Feng et al., 2020; Cai et al., 2023). This approach minimizes reliance on extensive historical records and real-time data collection, making it particularly effective for real-time air quality management. However, its long-term effectiveness depends on the stability and consistency of the input data. Moreover, long-term predictions may require periodic retraining of the model to adapt to new air quality data.

This study aims to predict the pollution of $PM_{2.5}$, including its concentrations and air quality index (AQI), at hourly intervals in Ho Chi Minh City (Vietnam) using statistical and machine learning models based on univariate time series concentrations of $PM_{2.5}$. The models considered in this study include ARIMA, LR, RF, LSTM, Bi-LSTM, and a hybrid DL model combining CNN and Bi-LSTM (CNN+Bi-LSTM). In fact, some statistical and ML models, such as LR and RF, were used to predict concentrations of air pollutants in Ho Chi Minh (HCM) City, Vietnam, at daily intervals based on multivariate time series data, which are datasets covering pollutant concentrations and meteorological parameters such as wind speed, rainfall levels, and air temperature (Rakholia et al., 2022; Minh et al., 2021; Le et al., 2022; Phung et al., 2020). The collection of these auxiliary data is sometimes a challenge, especially at hourly intervals, due to conditions of data access and monitoring stations. The prediction of $PM_{2.5}$ concentrations at hourly intervals in HCM City has not been considered in previous studies. In addition, the use of predicted $PM_{2.5}$ concentrations for further applications, such as the prediction of AQI, has not been widely considered. Therefore, the results of this study would improve the understanding of applying statistical and ML models for air pollution monitoring, providing more information on air pollutants and thereby supporting decision-making related to the mitigation of air pollution.

2. Data and methods

2.1. Study area

The study area is Ho Chi Minh (HCM) City, one of the metropolises of Vietnam. HCM City has a tropical climate with two seasons: the dry season (November to April of the subsequent year) and the rainy season (May to October) (Nguyen et al., 2023a; H. C. M. C. P. s. Committee, 2022). The ambient air temperatures of the two seasons tend to be relatively similar, ranging from 30 °C to 35 °C during the daytime throughout the year (H. C. M. C. P. s. Committee, 2022; Department of Natural Resources and Environment, 2021). Additionally, HCM City experiences significant variation in rainfall levels between the dry and rainy seasons, with the bulk of the rainfall occurring during the rainy season (May–October). Because of this weather condition, concentrations of atmospheric pollutants, such as $PM_{2.5}$, in HCM City are likely to be lower in the rainy season, as an increase in rainfall levels would wash out air pollutants from the atmosphere, leading to a decline in the concentration of air pollutants, including $PM_{2.5}$ (Nguyen et al., 2023a).

Air pollutants (e.g., $PM_{2.5}$) in HCM City were reported to be mainly emitted from vehicle emission (Nguyen et al., 2023a; B. Q. Ho et al., 2021; Q. B. Ho et al., 2019), which contributes from motorcycles accounting for over 90% (Ho, 2017). Other emission sources include household activities (Ho et al., 2019) and industrial production (Nguyen et al., 2023a; Ho et al., 2019), especially textile and food industries (Ho, 2017). Moreover, regarding hourly variation, $PM_{2.5}$ concentrations in HCM City tend to increase during the morning and evening rush hours (Hien et al., 2019), lasting from 7–9 a.m. and 5–7 p.m., respectively, due to the higher density of transportation vehicles during these periods (Nguyen et al., 2023a; Hien et al., 2019; Hoa, 2023).

2.2. Collection and preprocessing of $PM_{2.5}$ concentrations

This study considered hourly mass concentrations of $PM_{2.5}$ downloaded from the AirNow network (https://www.airnow.gov). The monitoring site is at the US Consultant in HCM City (latitude: 10.78° N and longitude: 106.7° E), surrounding by several prominent roads of District 1, HCM City (Figure 1). The data period considered in this study is from January 1, 2018 to December 31, 2022. To ensure data quality, only valid values passing quality control were further used in this study. The invalid concentrations, including null, outliers, and negative values, were treated as missing data, which accounted for approximately 3% of the dataset. The missing concentrations were filled in using linear regression algorithm, meaning that they were estimated using linear regression algorithm built from the series of $PM_{2.5}$ concentrations over the whole study period. Particularly, monitoring days with over 10 missing values were eliminated for data quality assurance.

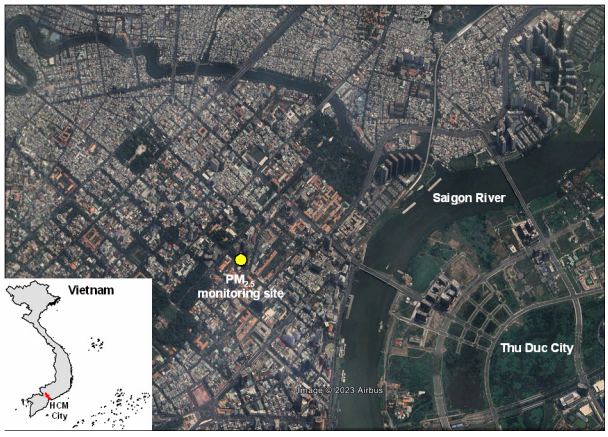

Figure 1. Map of the $PM_{2.5}$ monitoring site in HCM City, Vietnam

2.3. Model overview

Figure 2 illustrates the approach for predicting $PM_{2.5}$ concentrations used in this study. We first utilized the MinMaxScaler technique to scale the original input data (i.e., $PM_{2.5}$ concentrations over the study period) to a range between 0 and 1. Next, the normalized data were transformed into a supervised learning format, enabling the training of ML or DL models. Finally, we proposed and implemented the statistical and machine learning models to train and evaluate the performance of the system based on the processed input data.

To be more specific, we presented a univariate approach for predicting $PM_{2.5}$ concentrations. The input for our prediction models consisted only of $PM_{2.5}$ concentrations, which also served as the target variable. To represent the normalized data for regression-based supervised learning, we generated lagged versions of the features from the input $PM_{2.5}$ time series data, allowing us to capture temporal dependencies.

Specifically, $PM_{2.5}$ concentrations from previous time steps (i.e., the preceding 24 hours) are utilized as input features to predict concentrations for the following hours. We then combine these lagged features to construct a supervised learning dataset. Each row in this dataset comprises the input features from previous time steps alongside the corresponding target value, which is the $PM_{2.5}$ concentration for the following hour. Figure 3 describes an example of the transformation from the original univariate data to the input feature set for supervised learning using a timestep of 5.

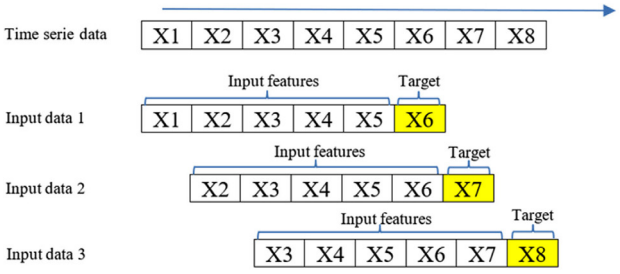

Figure 3. An example of the transformation from univariate data to the input feature set for supervised learning with a timestep of 5

Regarding the statistical and ML models, several models were applied to predict $PM_{2.5}$ concentrations, including ARIMA, ML (LR and RF), and DL (LSTM, Bi-LSTM, and hybrid model combining CNN and Bi-LSTM).

2.3.1. Autoregressive integrated moving average (ARIMA)

The ARIMA model integrates autoregressive (AR) and moving average (MA) processes. Regarding the *ARIMA (p, d, q)* process, *d* represents the number of times the series Y_t (e.g., $PM_{2.5}$ concentrations) needs to be differenced to become a stationary series, *p* denotes the autoregressive order, and *q* denotes the moving average order. Both *p* and *q* are orders corresponding to the differenced stationary series. In this study, the *(p, d, q)* parameters for the ARIMA model are derived from the experimental data-

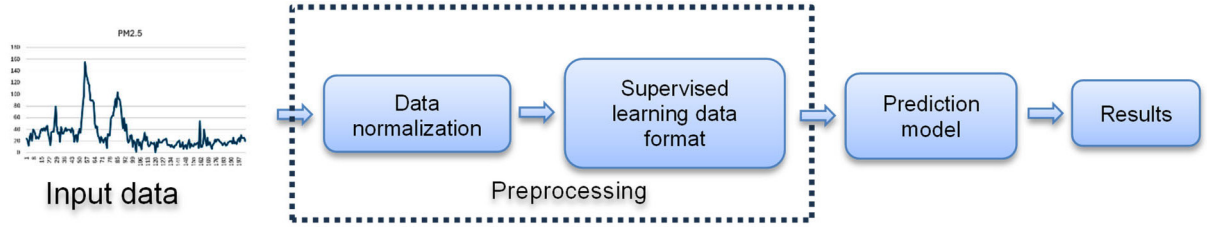

Figure 2. Approach for the prediction of $PM_{2.5}$ concentrations using statistical and machine learning models in this study

set itself. The estimation of the ARIMA model is achieved through the method of maximum likelihood estimation (MLE). More details on the ARIMA model can be found elsewhere (Wang et al., 2017; Nath et al., 2021).

2.3.2. Linear regression (LR) and random forest (RF)

We also examined two fundamental ML approaches for time series analysis, including linear regression (LR) and random forest (RF). The LR model examines the relationship between a dependent variable and one or more independent variables over time, aiming to find the best-fitting linear line that minimizes the difference between the observed data and the predicted values (Harishkumar et al., 2020).

For the RF, it is a well-known ensemble modelling method in ML utilizing bagging (bootstrap aggregating) to decrease variance and enhance the stability of predictions (Feng et al., 2020). This is achieved by training several decision tree models on varying subsets of the training data, accomplished by repetitively resampling the training dataset with replacement. For regression tasks, the average of all predictions is taken as the final output of the ensemble. To be more specific, the RF builds an ensemble of decision trees based on the bagging technique to predict future values within a time series. Each tree makes predictions at a specific point in the future, which are then combined using methods such as averaging to generate the final ensemble prediction. However, it is worth noting that the RF model can be complex and require higher computational cost (Bontempi et al., 2013).

2.3.3. Long-short term memory (LSTM) and bidirectional LSTM (Bi-LSTM)

Long Short-Term Memory (LSTM) falls within the category of recurrent networks, a category of artificial neural networks in which node connections create a directed graph over a sequential progression. One notable benefit of the LSTM model is its capacity to capture patterns and relationships from both past and future contexts in a data sequence. In other words, LSTM model can learn long-term dependencies within the input data (Wang et al., 2017, 2021; Hamami & Dahlan, 2020). An ordinary LSTM unit consists of a cell, along with an input gate, an output gate, and a forget gate. The cell retains information across various time spans, whereas the three gates control the data movement into and out of the cell. More details on LSTM model can be seen elsewhere (Sherstinsky, 2020; Hamami & Dahlan, 2020).

Furthermore, combining both past and future contexts within the LSTM can yield improved outputs and thus provide better information to neighboring elements. Consequently, we can combine two LSTMs in two opposite directions, for instance, one forward and one backward, and combined them into a unified framework known as as bidirectional LSTM (Bi-LSTM). This allows for the exploitation of information from the input feature sequence or series in both directions (Siami-Namini et al., 2019).

2.3.4. Hybrid convolutional neural network and Bi-LSTM model

To enhance the performance of the $PM_{2.5}$ concentration prediction system, we introduced a hybrid model combining the convolutional neural network (CNN) and Bi-LSTM, called CNN+Bi-LSTM. Because the input $PM_{2.5}$ concentration data were univariate (one-dimensional time series data), we used a one-dimensional CNN (1D-CNN) to learn features from the input data. The Bi-LSTM was then fed these features to take advantage of the past and future contexts of the input $PM_{2.5}$ concentration time series data. The network configuration summary for $PM_{2.5}$ concentration prediction is displayed in Figure 4.

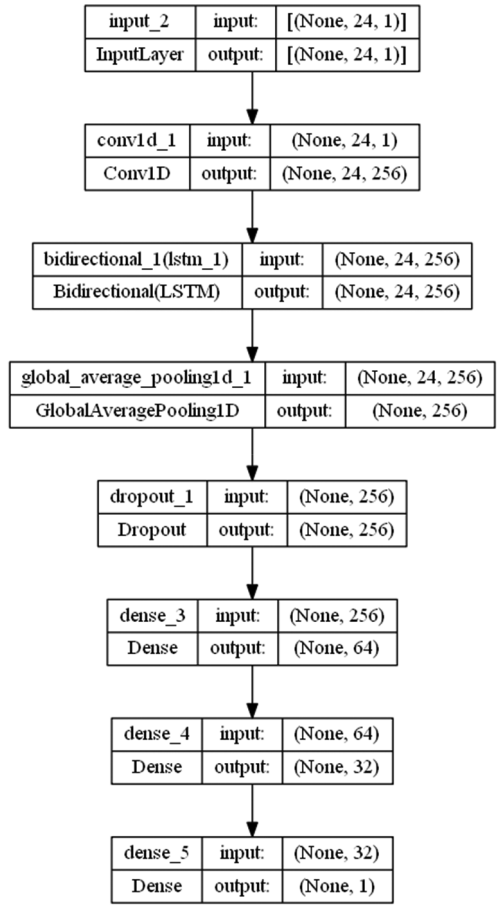

Figure 4. A hybrid CNN+Bi-LSTM model using Keras framework

2.4. Experiment setup

The Auto ARIMA and Grid Search functions were utilized to identify the optimal parameters for the ARIMA and ML models. (i.e., LR and RF models), respectively. The ARIMA, ML (LR and RF), and DL (LSTM, Bi-LSTM, and CNN+Bi-LSTM) models were implemented using the 'auto_arima', 'sklearn', and Keras frameworks, respectively. A summary of these model hyperparameters is shown in Table 1. In addition, the dataset was divided into training and test sets for model running. Two experiments were set up in this study. For the first one, the hourly $PM_{2.5}$ concentra-

tions from 00h January 1, 2018, to 23h December 31, 2020 were used as training data and the $PM_{2.5}$ concentrations in 2021 (00h January 1 – 23h December 31) were considered as test data. Regarding the second experiment, the hourly $PM_{2.5}$ concentrations from 00h January 1, 2018 to 23h December 31, 2021 were used as training data. The $PM_{2.5}$ concentrations in 2022 (00h January 1 – 23h December 31) were used as test data in the second experiment.

Table 1. Hyperparameters of the models used on this study

Hyperparameter	Value
ARIMA	
Number of time-lags (*p*)	1–5
The order of moving average (*q*)	1–5
The order of first differencing (*d*)	1
LR	
fit_intercept	True
n_jobs	–1
RF	
n_jobs	–1
max_depth	9
max_features	'auto'
estimators	100
CNN+Bi-LSTM	
Number of filters of 1D Conv layer	256
Kernel size of Conv layer	1
Number of units in the Bi-LSTM layers	256
Dropout rate	0.2
Number of units of dense layers	64, 32, 1
Optimizer	'adam'
Loss functions	'mae'

Hyperparameter	Value
LSTM	
Model Initialization	Sequential
Number of units in the LSTM layer	64
Dropout	0.2
Number of units Dense Layer	1
Optimizer	'adam'
Loss functions	'mae'
Bi-LSTM	
Model Initialization	Sequential
Number of units in the Bi-LSTM layers	128, 64
Dropout	0.2
Number of units Dense Layer	1
Optimizer	'adam'
Loss functions	'mae'

2.5. Evaluation metrics

For the evaluation of model performance, several evaluation metrics were used, including coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Equations of these evaluation metrics are expressed below:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}; \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|; \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2; \quad (3)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}; \quad (4)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{n}, \quad (5)$$

where y_i is the monitored $PM_{2.5}$ concentrations; $\bar{y}_i$ denotes the arithmetic mean concentration of the monitored $PM_{2.5}$; $\hat{y}_i$ is $PM_{2.5}$ concentrations predicted by the models. In general, the higher R^2 values (i.e., close to a unity) or the lower error values suggest better model performance.

2.6. Computation of air quality index (AQI)

To test the suitability of $PM_{2.5}$ concentrations estimated by the best performance model for further applications, the modeled results were used to compute the air quality index (AQI), representing pollution levels of the ambient air associated with human health. The hourly and daily AQIs of $PM_{2.5}$ were calculated following the technical guideline for calculation of Vietnamese AQI (VN_AQI) (VEA, 2019). In short, according to the value, VN_AQI is classified into six levels of concern, which are: good (<50), moderate (51–100), unhealthy for sensitive group (101–150), unhealthy (151–200), very unhealthy (201–300), and hazardous (301–500). The VN_AQI were computed using the Nowcast concentrations, the weighted $PM_{2.5}$ concentrations calculated from those at the preceding 12 hours of the considered hour. More details on the VN_AQI computation are described elsewhere (VEA, 2019).

3. Results and discussion

3.1. Description of $PM_{2.5}$ concentrations and AQI

The statistical analysis of $PM_{2.5}$ concentrations in HCM City used in this study is described in Table 2.

As shown in Table 2, $PM_{2.5}$ concentrations in HCM City experienced a decreasing trend from 2018 (mean±SD: 26.31±12.13 μg/m^3) to 2022 (mean±SD: 23.80±11 μg/m^3). However, no statistically significant difference in the $PM_{2.5}$ concentrations of the target years was found (one-way ANOVA on ranks, $p > 0.05$). The SD, indicating variability in the data, showed the highest variability in hourly data (SD: 13.44–17.65 μg/m^3) compared to 24-hour averages (SD: 9.31–12.36 μg/m^3), reflecting the more fluctuation of $PM_{2.5}$ concentrations in HCM City regrading hourly variation. In addition, $PM_{2.5}$ concentrations in HCM City met the $PM_{2.5}$ concentration threshold issued by Vietnamese's government (24-hour level: 50 μg/m^3) (Ministry of Natural Resources and Environment, 2013). However, compared to the WHO's guideline on air quality, concentrations of $PM_{2.5}$ in HCM City mostly exceeded the threshold (24-hour

Table 2. Statistical description of $PM_{2.5}$ concentrations (μg/m^3) in HCM City over the study period

Year	24-h average					Hourly				
	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
2018	26.31	12.13	23.46	3.40	68.08	26.31	17.65	22.00	1.00	168.00
2019	25.79	12.36	22.73	3.25	98.32	25.82	17.20	21.00	1.00	196.00
2020	23.17	9.31	21.31	7.46	63.58	23.17	13.90	20.00	1.00	132.00
2021	22.37	10.86	20.04	6.96	87.42	22.38	14.89	19.00	1.00	167.00
2022	23.80	11.00	20.88	7.29	125.00	23.64	13.44	20.02	1.00	147.00
2018–2022	24.32	11.32	21.75	3.25	125.00	24.43	16.05	20.00	1.00	196.00

Note: SD: standard deviation.

level: 15 μg/m^3) (World Health Organization, 2021) over the study period.

The monthly mean concentrations of $PM_{2.5}$ from 2018 to 2022 are illustrated in Figure 5. The highest concentrations of $PM_{2.5}$ tended to be found in the dry season, especially in November, January, and February. Added to this, the $PM_{2.5}$ concentrations were lower in the rainy season, especially the period of July to September (Figure 5). During the rainy season, the rainfall levels in HCM City increase significantly, and the higher rainfall amount would result in wet deposition of $PM_{2.5}$, which contribute to a decline of $PM_{2.5}$ concentrations in the rainy season (Hien et al., 2019).

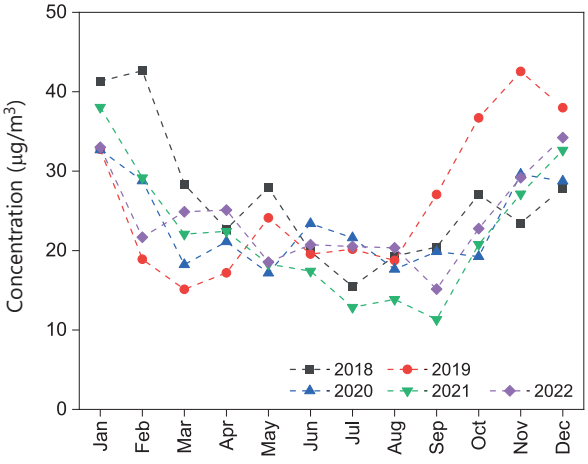

Figure 5. Monthly mean concentrations of $PM_{2.5}$ in HCM City in 2018–2022

Regarding hourly variation, the concentrations of $PM_{2.5}$ generally followed a similar trend throughout the year. Specifically, the concentrations reached their lowest values around 1–2 a.m., then continuously rose and peaked at 9 a.m. (Figure 6). Moreover, the concentrations of $PM_{2.5}$ decreased gradually after the peak, which could be attributed to a fall in traffic density (i.e., the number of vehicles) after the morning rush hours. Additionally, the lower $PM_{2.5}$ concentrations might be influenced by meteorological conditions, such as the expansion of the planetary boundary layer in the afternoon, which enhances the dispersion of air pollutants and contributes to the decrease in $PM_{2.5}$ concentrations (Nguyen et al., 2023a; Hien et al., 2019).

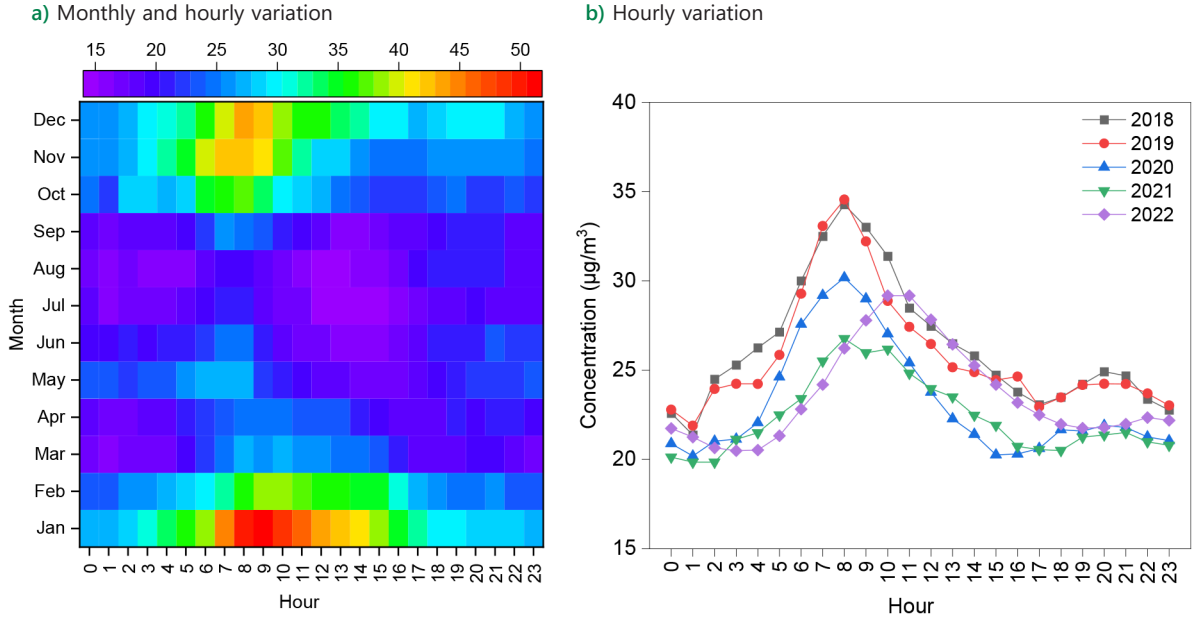

Figure 6. Mean concentrations of $PM_{2.5}$ with respect to monthly and hourly variations

Notably, hourly mean concentrations of $PM_{2.5}$ in several months, consisting of January, February, October, November, and December, were obviously higher than those in the others (Figure 6). An explanation would be because these months tended to have the lower rainfall amount, leading to the prolonged occurrence of atmospheric pollutants (e.g., $PM_{2.5}$). Furthermore, the lower ambient air temperature in such months (Nguyen et al., 2023a, 2023b) would possibly contribute to an increase of $PM_{2.5}$ concentrations because air pollutants tend to more accumulate near the ground surface as a result of the lower air mixing height derived from the lower ambient air temperature (Nguyen et al., 2023a).

Figure 7 illustrates the hourly VN_AQI values of $PM_{2.5}$ in HCM City shown for each month. The VN_AQI values below 50 and in the range of 50–100 indicate good and moderate air quality, respectively. As shown in Figure 7, the AQI values were generally lower in the period from April to October and ranged from 30 to 60, indicating better air quality in HCM City. Additionally, from November to February of the following year, AQI values increased, reflecting worse air quality during this period. These observations were also in line with the temporal variation of $PM_{2.5}$ concentrations mentioned above.

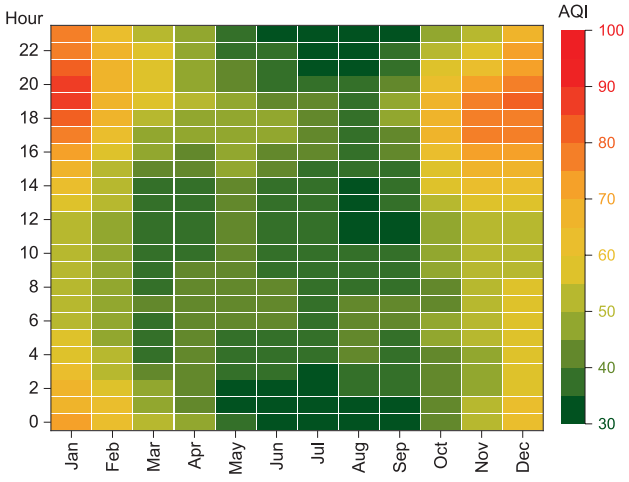

Figure 7. Hourly mean values of $PM_{2.5}$ air quality index (VN_AQI) shown for each month over the study period

3.2. Prediction of $PM_{2.5}$ concentrations

A total of six models, including ARIMA, LR, RF, LSTM, Bi-LSTM and hybrid CNN+Bi-LSTM, were trained and tested for the prediction of $PM_{2.5}$ concentrations. The performance of these models is described in Table 3 and Figure 8.

As shown in Table 3, the predicted results of all models would be regarded as having reasonable prediction accuracy because their MAPE values ranged between 29% and 40% (Kumari & Singh, 2023). Traditional models like ARIMA and LR proved less effective in handling the nonlinear patterns in $PM_{2.5}$ concentration data, as indicated by their higher error metrics and percentage errors. The ARIMA model showed the worst performance because it had the lowest R^2 value (0.58–0.64) and the highest error values (MAE: 5.55–6.00., MSE: 73.1–77.7., RMSE: 8.55–8.82., and MAPE: 30%–41%). This would be because prior to being added into the ARIMA, the $PM_{2.5}$ concentrations were differentiated to obtain stationary status, reflecting a time-independent state. The concentration differentiation could lead to information loss, leading to the worse performance of ARIMA compared to that of the other models (Kumari & Singh, 2023). Moreover, there were more underestimated concentrations in ARIMA than in the other models (Figure 8), indicating that ARIMA may not be an appropriate univariate time series model for predicting $PM_{2.5}$ concentrations in HCM City. Several previous studies also reported the underperformance of ARIMA, compared to ML models, in estimating pollutant concentrations (Kumari & Singh, 2023; Wu et al., 2021; Ma et al., 2020).

The evaluation metrics of the LR and RF models were approximately 0.60–0.70, 5.45–5.65, 66.4–68.8, 8.15–8.30, and 30%–40% for the R^2, MAE, MSE, RMSE, and MAPE, respectively (Table 3). Added to this, the prediction of these two models was more accurate in the early stages, then declined in the rest of the testing period, especially in July–October (Figure 9). The LR model is appropriate for predicting continuous data (i.e., hourly $PM_{2.5}$ concentrations); thus, the performance of the LR model was better than that of ARIMA. The RF model showed better performance compared to LR because RF is an ensemble model,

Table 3. Evaluation metrics of the models

Model	R^2	MAE	MSE	RMSE	MAPE
Experiment 1 (Training data: 2018–2020, Test data: 2021)					
ARIMA	0.64	6.00	77.7	8.82	41%
LR	0.70	5.65	66.4	8.15	40%
RF	0.69	5.64	66.7	8.17	40%
LSTM	0.70	5.51	65.8	8.11	37%
Bi-LSTM	0.70	5.51	65.6	8.09	37%
CNN+ Bi-LSTM	0.70	5.52	65.4	8.08	37%

Model	R^2	MAE	MSE	RMSE	MAPE
Experiment 2 (Training data: 2018–2021, Test data: 2022)					
ARIMA	0.58	5.55	73.1	8.55	30%
LR	0.61	5.47	68.8	8.29	30%
RF	0.62	5.44	67.2	8.20	30%
LSTM	0.62	5.36	68.4	8.27	29%
Bi-LSTM	0.62	5.38	68.6	8.28	29%
CNN+ Bi-LSTM	0.62	5.37	67.5	8.21	29%

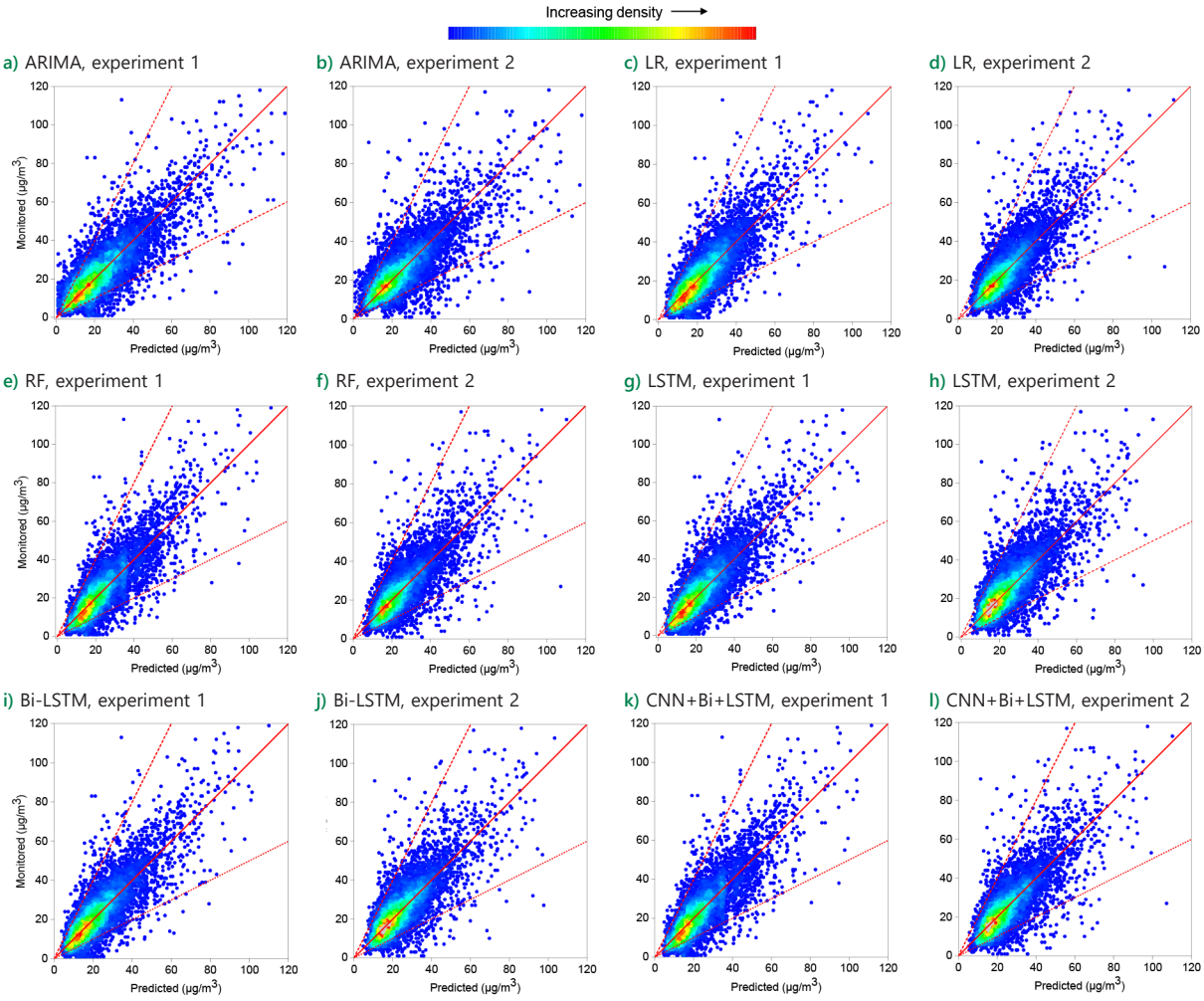

Figure 8. Scatter plots of the monitored and predicted $PM_{2.5}$ concentrations

combining multiple regressors to improve prediction accuracy.

Furthermore, models capable of capturing temporal dependencies, such as LSTM, Bi-LSTM, and CNN+Bi-LSTM, showed the best performance, particularly in reducing percentage errors (i.e., the lowest MAPE) (Table 3). All these models achieved the R^2 value of 0.62–0.70, demonstrating a robust capability to predict variance in $PM_{2.5}$ concentrations. The superior performance of DL models, compared to ML models such as RF and LR, has also been noted in several previous studies (Kumari & Singh, 2023; Rakholia, et al., 2022; Minh et al., 2021; Ma et al., 2020). The hybrid model CNN+BiLSTM showed the best performance among the models and had the lowest prediction errors (R^2: 0.62–0.70., MAE: 5.49., MSE: 64.8., RMSE: 8.04., MAPE: 37%). This is because CNN+Bi-LSTM can learn features and capture long-term $PM_{2.5}$ concentrations, meaning that it could memorize the past and present values and then use them as input for estimating the subsequent concentrations. However, the LSTM model still miscomputed $PM_{2.5}$ concentrations at some points (i.e., October 8th to 22nd, 2021) (Figures 8 and 9). Thus, an increase in data volume, such as the number of input features, is suggested in further studies to improve the model's performance.

3.3. Calculation of VN_AQI using the predicted $PM_{2.5}$ concentrations

The predicted $PM_{2.5}$ concentrations from the best-performing model (i.e., CNN+Bi-LSTM) were used to calculate the VN_AQI, considering hourly $PM_{2.5}$ concentrations (hereafter $VN\_AQI_{PM2.5}$), to check the suitability of the estimated results for further applications. The predicted $PM_{2.5}$ concentrations in two experiments were considered. Based on the $VN\_AQI_{PM2.5}$ values, the air quality was classified into several levels, including: (1) good, (2) moderate (i.e., air quality is acceptable and air pollution can negatively affect sensitive groups such as elders and children), (3) unhealthy for sensitive groups, (4) unhealthy (i.e., air pollution can negatively affect human health and sensitive groups having severe health problems), (5) very unhealthy (i.e., air pollution can potentially have a severe impact on human health), (6) hazardous (i.e., air pollution is likely to have a severe effect on human health) (VEA, 2019). The concern levels of $VN\_AQI_{PM2.5}$ identified using the predicted and monitored $PM_{2.5}$ concentrations (hereafter predicted and monitored $VN\_AQI_{PM2.5}$, respectively) were compared and visualized in Figure 10.

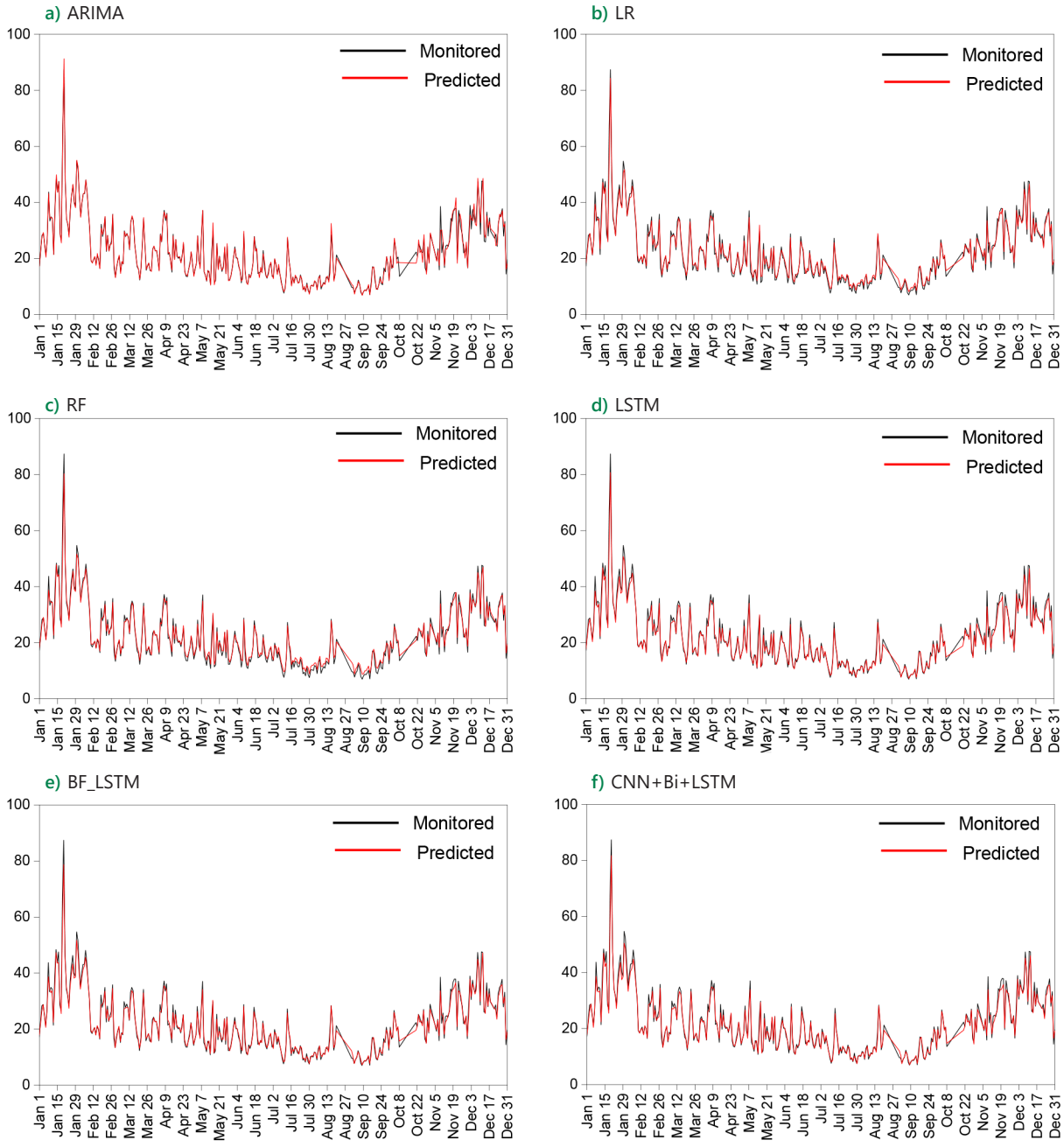

Figure 9. Comparison of the monitored and predicted $PM_{2.5}$ concentrations shown as daily averages over the test period in 2021

Figure 10 shows that the strongest degree of agreement between predicted and monitored $VN\_AQI_{PM2.5}$ was observed for the 'good air quality' level, corresponding to a $VN\_AQI_{PM2.5}$ value lower than 50. The levels of 'moderate air quality' and 'unhealthy for sensitivity group' also experienced the high and moderate degree of agreement, accounting for approximately 80% and 55%, respectively. Interestingly, the 'unhealthy air quality' level had the moderate degree of agreement (i.e., 63%) when using the predicted $PM_{2.5}$ concentrations of experiment 1 (training data: 2018–2020, test data: 2021). However, regarding the experiment 2 (training data: 2018–2021, test data: 2022), the degree of agreement at this level declined by approximately threefold (i.e., 28%). This could be due to the higher prediction errors (Table 3) and the more overestimated values produced by the CNN+Bi-LSTM model in experiment 2 (Figure 8l).

Regarding the $VN\_AQI_{PM2.5}$ calculated from the monitored $PM_{2.5}$ concentrations (i.e., the monitored $VN\_AQI_{PM2.5}$), air quality classified as 'unhealthy' and 'unhealthy for sensitive groups' was mainly observed from January to April and November to December (Figure 11). Additionally, from May to October, the air quality was likely to be better as the monitored $VN\_AQI_{PM2.5}$ was frequently determined to be good and moderate, corresponding to air quality categories 1–3.

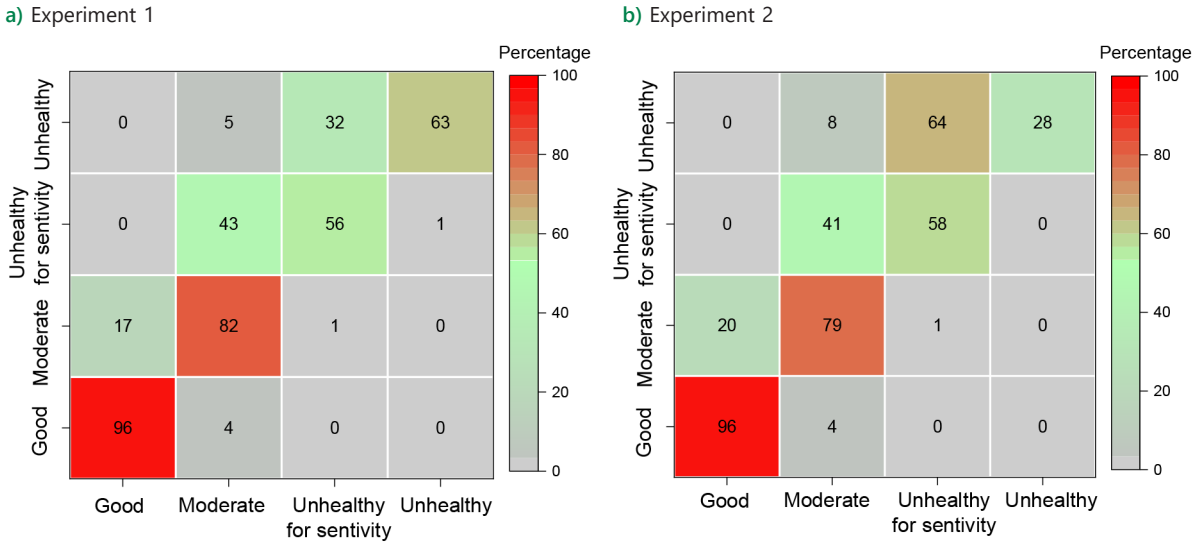

Figure 10. Heatmap of the concern levels from the predicted and monitored VN_AQI$_{PM2.5}$. The percentage represents the degree of agreement between the predicted and monitored VN_AQI$_{PM2.5}$

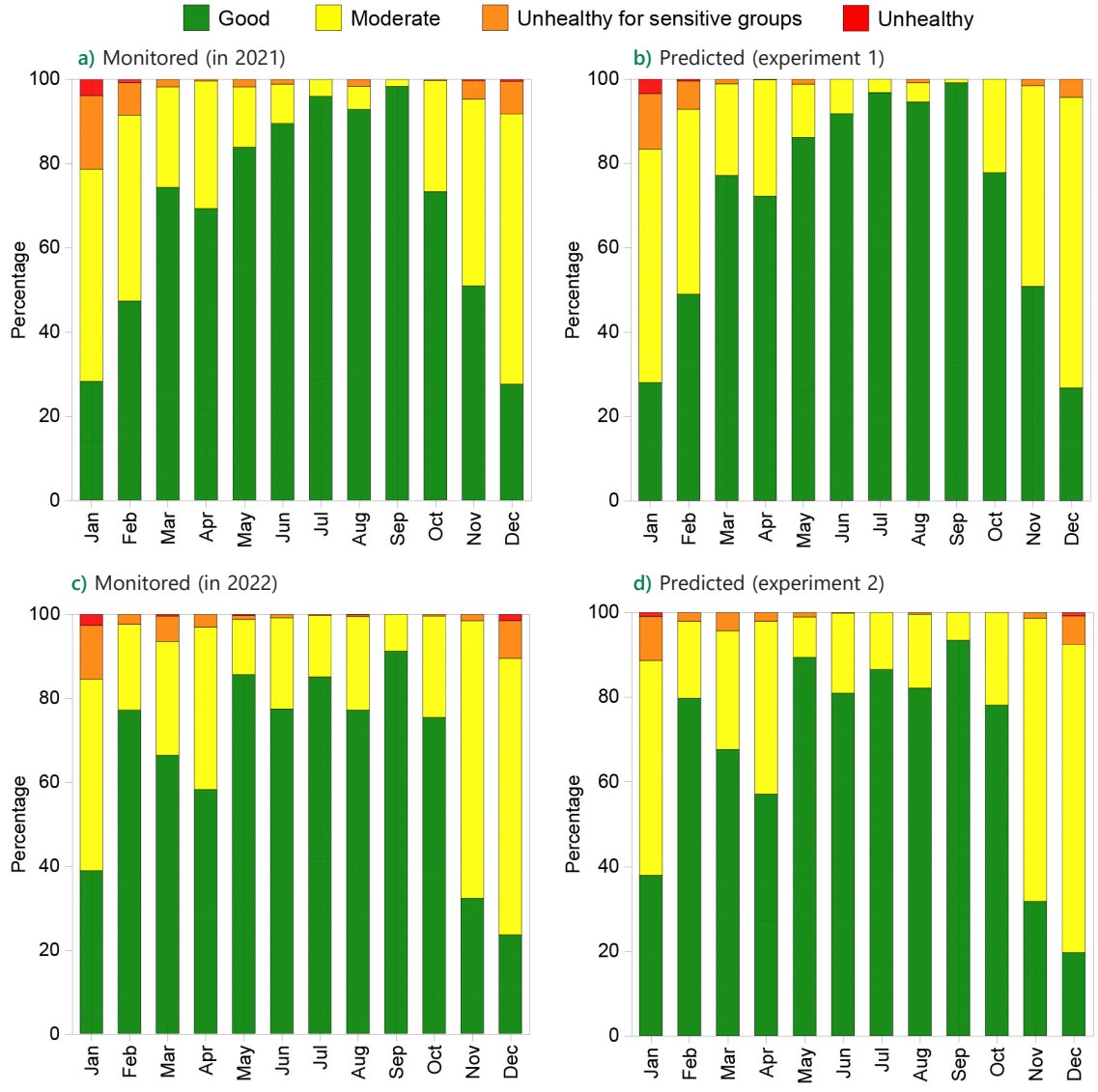

Figure 11. Percentage of the concern levels acquired from the monitored and predicted VN_AQI$_{PM2.5}$ shown in individual months of the test period

The results also revealed that the $VN\_AQI_{PM2.5}$ values calculated from the predicted $PM_{2.5}$ concentrations highly matched those from the monitored concentrations (Figures 11b and 11c). Additionally, regarding experiment 2 (training data: 2018–2021, test data: 2022), the $VN\_AQI_{PM2.5}$ values were overestimated for the 'unhealthy for sensitive groups' category in January and December (Figures 10 and 11d). This observation could be due to the differences between the monitored and predicted concentrations of $PM_{2.5}$ as mentioned previously, that is, some concentration peaks were underestimated by the CNN+Bi-LSTM model (Figure 9f). However, the results of this study reflect the capability of using $PM_{2.5}$ concentrations predicted by the ML and/or DL models for further applications, such as identifying the $VN\_AQI_{PM2.5}$ and concern levels for human health.

Moreover, an improvement in the model's performance can contribute to a more accurate determination of air quality categories. Further studies are suggested to add more input features (e.g., meteorological variables) into the models. Other DL models, as well as combinations of DL and other models, can also be considered to provide better predictions.

4. Conclusions

To sum up, this study used a total of six models, including ARIMA, LR, RF, LSTM, Bi-LSTM, and CNN+Bi-LSTM, for univariate time series prediction of hourly concentrations of $PM_{2.5}$ in HCM City, Vietnam. The results showed that CNN+Bi-LSTM, a DL model, outperformed the others because it has the capability of memorizing information over long sequences, such as time series data. The predicted concentrations from this model were also used to calculate $VN\_AQI_{PM2.5}$, the Vietnamese air quality index for $PM_{2.5}$, to check the appropriateness of the predicted results for further applications. Consequently, the average MAPE was 34%, and some air quality categories were misidentified because of errors in the concentration prediction. Based on these findings, it is suggested that the predicted values can be used for further applications, such as AQI identification. However, the model's performance should be improved by increasing the data volume, including the number of input variables. Another option is to consider other hybrid DL algorithms to enhance the prediction accuracy of the models. Overall, this study contributes to a better understanding of statistical and ML model applications in the monitoring of air pollutants.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Author contributions

TNT Nguyen: conceptualization, supervision, formal analysis, and writing; TD Trinh: methodology, formal analysis, data curation, and writing – review & editing; PCLT Vu: formal analysis and data curation; PT Bao: methodology and data curation.

Disclosure statement

The authors declare that there is no conflict of interest.

References

Bai, W., & Li, F. (2023). $PM_{2.5}$ concentration prediction using deep learning in internet of things air monitoring system. *Environmental Engineering Research*, *28*(1), Article 210456. https://doi.org/10.4491/eer.2021.456

Barthwal, A., & Goel, A. K. (2024). Advancing air quality prediction models in urban India: A deep learning approach integrating DCNN and LSTM architectures for AQI time-series classification. *Modeling Earth Systems and Environment*, *10*, 2935–2955. https://doi.org/10.1007/s40808-023-01934-9

Bhatti, U. A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., Yu, Z., & Yuan, L. (2021). Time series analysis and forecasting of air pollution particulate matter $PM_{2.5}$: An SARIMA and factor analysis approach. *IEEE Access*, *9*, 41019–41031. https://doi.org/10.1109/ACCESS.2021.3060744

Bontempi, G., Ben Taieb, S., & Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. In M.-A. Aufaure & E. Zimányi (Eds.), *Lecture notes in business information processing: Vol. 138. Business intelligence: Second European Summer School, eBISS 2012* (pp. 62–77). Springer. https://doi.org/10.1007/978-3-642-36318-4_3

Cai, P., Zhang, C., & Chai, J. (2023). Forecasting hourly $PM_{2.5}$ concentrations based on decomposition-ensemble-reconstruction framework incorporating deep learning algorithms. *Data Science and Management*, *6*(1), 46–54. https://doi.org/10.1016/j.dsm.2023.02.002

Chlebnikovas, A., Paliulis, D., Bradulienė, J., & Januševičius, T. (2023). Short-term field research on air pollution within the boundaries of the large city in the Baltic region. *Environmental Science and Pollution Research*, *30*(34), 81950–81965. https://doi.org/10.1007/s11356-022-23798-9

Clark, S. N., Kulka, R., Buteau, S., Lavigne, E., Zhang, J. J. Y., Riel-Roberge, C., Smargiassi, A., Weichenthal, S., & van Ryswyk, K. (2024). High-resolution spatial and spatiotemporal modelling of air pollution using fixed site and mobile monitoring in a Canadian city. *Environmental Pollution*, *356*, Article 124353. https://doi.org/10.1016/j.envpol.2024.124353

Department of Natural Resources and Environment. (2021). *Report of the environmental status of Ho Chi Minh city*. Ho Chi Minh City.

Ejohwomu, O. A., Shamsideen Oshodi, O., Oladokun, M., Bukoye, O. T., Emekwuru, N., Sotunbo, A., & Adenuga, O. (2022). Modelling and forecasting temporal $PM_{2.5}$ concentration using ensemble machine learning methods. *Buildings*, *12*(1), Article 46. https://doi.org/10.3390/buildings12010046

Feng, L., Li, Y., Wang, Y., & Du, Q. (2020). Estimating hourly and continuous ground-level $PM_{2.5}$ concentrations using an ensemble learning algorithm: The ST-stacking model. *Atmospheric Environment*, *223*, Article 117242. https://doi.org/10.1016/j.atmosenv.2019.117242

Filonchyk, M., Yan, H., & Hurynovich, V. (2017). Temporal-spatial variations of air pollutants in Lanzhou, Gansu Province, China, during the spring–summer periods, 2014–2016. *Environmental Quality Management*, *26*(4), 65–74. https://doi.org/10.1002/tqem.21502

Filonchyk, M., Yan, H., Yang, S., & Lu, X. (2018). Detection of aerosol pollution sources during sandstorms in Northwestern China using remote sensed and model simulated data. *Advances in Space Research*, *61*(4), 1035–1046. https://doi.org/10.1016/j.asr.2017.11.037

H. C. M. C. P. s. Committee. (2022). *Climate and weather of Ho Chi Minh City*. https://hochiminhcity.gov.vn/-/khi-hau-thoi-tiet?redirect=%2Fdieu-kien-tu-nhien

Hamami, F., & Dahlan, I. A. (2020, October 20–21). Univariate time series data forecasting of air pollution using LSTM neural network. In *2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)* (pp. 1–5), Lombok, Indonesia. https://doi.org/10.1109/ICADEIS49811.2020.9277393

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

Hien, T. T., Chi, N. D. T., Nguyen, N. T., Vinh, L. X., Takenaka, N., & Huy, D. H. (2019). Current status of fine particulate matter ($PM_{2.5}$) in Vietnam's most populous city, Ho Chi Minh City. *Aerosol Air Quality Research*, *19*(10), 2239–2251. https://doi.org/10.4209/aaqr.2018.12.0471

Ho, B. Q. (2017). Modeling PM_{10} in Ho Chi Minh City, Vietnam and evaluation of its impact on human health. *Sustainable Environment Research*, *27*(2), 95–102. https://doi.org/10.1016/j.serj.2017.01.001

Ho, B. Q., Vu, H. N. K., Nguyen, T. T. T., Nguyen, T. T., Nguyen, T. T. H., Khoa, N. T. D., & Phu, V. L. (2021). Photochemical modeling of $PM_{2.5}$ and design measures for $PM_{2.5}$ reduction: A case of Ho Chi Minh City, Vietnam. *IOP Conference Series: Earth Environmental Science*, *652*(1), Article 012025. https://doi.org/10.1088/1755-1315/652/1/012025

Ho, Q. B., Vu, H. N. K., Nguyen, T. T., Nguyen, T. T. H., & Nguyen, T. T. T. (2019). A combination of bottom-up and top-down approaches for calculating of air emission for developing countries: A case of Ho Chi Minh City, Vietnam. *Air Quality, Atmosphere & Health*, *12*(9), 1059–1072. https://doi.org/10.1007/s11869-019-00722-8

Hoa, N. T. (2023). Evaluation of fine particulate matter ($PM_{2.5}$) concentrations in Ho Chi Minh City in 2021 (in Vietnamese). *Tạp chí khí tượng thủy văn*, *2023*(751), 68–77.

Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO_2 emission prediction in India. *Environmental Science and Pollution Research*, *30*, 116601–116616. https://doi.org/10.1007/s11356-022-21723-8

Le, C. D., Pham, H. V., Pham, D. A., Le, A. D., & Vo, H. B. (2022, December 20–22). A $PM_{2.5}$ concentration prediction framework with vehicle tracking system: From cause to effect. In *2022 RIVF International Conference on Computing and Communication Technologies* (pp. 714–719), Ho Chi Minh City, Vietnam. https://doi.org/10.1109/RIVF55975.2022.10013864

Ma, J., Yu, Z., Qu, Y., Xu, J., & Cao, Y. (2020). Application of the XGBoost machine learning method in $PM_{2.5}$ prediction: A case study of Shanghai. *Aerosol and Air Quality Research*, *20*(1), 128–138. https://doi.org/10.4209/aaqr.2019.08.0408

Minh, V. T. T., Tin, T. T., & Hien, T. T. (2021). $PM_{2.5}$ forecast system by using machine learning and WRF model, a case study: Ho Chi Minh City, Vietnam. *Aerosol and Air Quality Research*, *21*(12), Article 210108. https://doi.org/10.4209/aaqr.210108

Ministry of Natural Resources and Environment. (2013). *National technical regulation on ambient air quality* (QCVN 05:2013/BTNMT). Ha Noi, Vietnam.

Nath, P., Saha, P., Middya, A. I., & Roy, S. (2021). Long-term time series pollution forecast using statistical and deep learning methods. *Neural Computing and Applications*, *33*(19), 12551–12570. https://doi.org/10.1007/s00521-021-05901-2

Nguyen, T. N. T., Du, N. X., & Hoa, N. T. (2023a). Emission source areas of fine particulate matter ($PM_{2.5}$) in Ho Chi Minh City, Vietnam. *Atmosphere*, *14*(3), Article 579. https://doi.org/10.3390/atmos14030579

Nguyen, T. N. T., Nguyen, N. T., Nguyen, M. T. T., & Bao, P. T. (2023b). Characteristics and effect of the temperature inversion on concentrations of fine particulate matter ($PM_{2.5}$) in Ho Chi Minh city. *Vietnam Journal of Hydro-Meteorology*, *746*, 87–95.

Phung, N. K., Long, N. Q., Tin, N. V., & Le, D. T. T. (2020). Development of a $PM_{2.5}$ forecasting system integrating low-cost sensors for Ho Chi Minh City, Vietnam. *Aerosol and Air Quality Research*, *20*(6), 1454–1468. https://doi.org/10.4209/aaqr.2019.10.0490

Rabie, R., Asghari, M., Nosrati, H., Niri, M. E., & Karimi, S. (2024). Spatially resolved air quality index prediction in megacities with a CNN-Bi-LSTM hybrid framework. *Sustainable Cities and Society*, *109*, Article 105537. https://doi.org/10.1016/j.scs.2024.105537

Rakholia, R., Le, Q., Vu, K., Ho, B. Q., & Carbajo, R. S. (2022). AI-based air quality $PM_{2.5}$ forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. *Urban Climate*, *46*, Article 101315. https://doi.org/10.1016/j.uclim.2022.101315

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, Article 132306. https://doi.org/10.1016/j.physd.2019.132306

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 3285–3292). IEEE. https://doi.org/10.1109/BigData47090.2019.9005997

Tong, W., Li, L., Zhou, X., Hamilton, A., & Zhang, K. (2019). Deep learning $PM_{2.5}$ concentrations with bidirectional LSTM RNN. *Air Quality, Atmosphere & Health*, *12*, 411–423. https://doi.org/10.1007/s11869-018-0647-4

Upadhya, A. R., Kushwaha, M., Agrawal, P., Gingrich, J. D., Asundi, J., Sreekanth, V., Marshall, J. D., & Apte, J. S. (2024). Multi-season mobile monitoring campaign of on-road air pollution in Bengaluru, India: High-resolution mapping and estimation of quasi-emission factors. *Science of the Total Environment*, *914*, Article 169987. https://doi.org/10.1016/j.scitotenv.2024.169987

Vietnam Environment Administration. (2019). *Technical guidance for calculation and publication of Vietnamese air quality index* (VN_AQI).

Wang, P., Zhang, H., Qin, Z., & Zhang, G. (2017). A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting. *Atmospheric Pollution Research*, *8*(5), 850–860. https://doi.org/10.1016/j.apr.2017.01.003

Wang, Z., Zhou, Y., Zhao, R., Wang, N., Biswas, A., & Shi, Z. (2021). High-resolution prediction of the spatial distribution of $PM_{2.5}$ concentrations in China using a long short-term memory model. *Journal of Cleaner Production*, *297*, Article 126493. https://doi.org/10.1016/j.jclepro.2021.126493

World Health Organization. (2021). *WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Geneva.

Wu, C., Li, B., & Xiong, N. (2021). An effective machine learning scheme to analyze and predict the concentration of persis-

tent pollutants in the Great Lakes. *IEEE Access*, *9*, 52252–52265. https://doi.org/10.1109/ACCESS.2021.3069990

Xu, C., Xu, D., Liu, Z., Li, Y., Li, N., Chartier, R., Chang, J., Wang, Q., Wu, Y., & Li, N. (2020). Estimating hourly average indoor $PM_{2.5}$ using the random forest approach in two megacities, China. *Building and Environment*, *180*, Article 107025. https://doi.org/10.1016/j.buildenv.2020.107025

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). $PM_{2.5}$ prediction based on Random Forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7), Article 373. https://doi.org/10.3390/atmos10070373

Zhao, R., Gu, X., Xue, B., Zhang, J., & Ren, W. (2018). Short period $PM_{2.5}$ prediction based on multivariate linear regression model. *PLoS ONE*, *13*(7), Article e0201011. https://doi.org/10.1371/journal.pone.0201011