Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Environ. Sci.: Adv.*, 2026, **5**, 1116

# Information-based approach to $PM_{2.5}$ estimation and air quality assessment using statistical and deep learning models

Sehrish Khan,[a] Maqbool Ahmad,[b] Bahadar Zeb,[c] Shahla Nazneen,[a] Beenish Ali,[d] Mubarak Ahmad,[e] Khan Alam*[f] and Allah Ditta *[g]

Received 25th October 2025
Accepted 16th February 2026

DOI: 10.1039/d5va00383k

rsc.li/esadvances

In Pakistan, Peshawar City is persistently experiencing high concentrations of fine particulate matter ($PM_{2.5}$), frequently surpassing national as well as international air quality standards. For this purpose, the present study aims to enhance the accuracy of $PM_{2.5}$ estimation at the city scale through a data-driven and interdisciplinary modeling framework. To achieve this, a series of predictors, such as air pollutants (nitrogen dioxide ($NO_2$) and sulphur dioxide ($SO_2$)), meteorological conditions (temperature, wind speed, humidity), and satellite-based aerosol optical depth (AOD), were used to construct a multiple linear regression (MLR) model. Similarly, the Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) were modeled to estimate $PM_{2.5}$ using historical ground-level $PM_{2.5}$ data in the year 2021, leveraging their capabilities to model temporal trends. The results revealed that estimated $PM_{2.5}$ levels using the CNN model were almost in the same range as the available measured concentrations, whereas MLR and LSTM models showed some variations against measured values. The insights about their comparative analysis showed that the CNN model could achieve better estimation than MLR and LSTM models. The CNN model achieved a root mean square error (RMSE) of 34.89 μg m$^{-3}$ and coefficient of determination ($R^2$) of 0.79, indicating higher estimation accuracy. Both the LSTM ($R^2$ = 0.74 and RMSE = 51.93 μg m$^{-3}$) and MLR ($R^2$ = 0.46 and RMSE = 44.35 μg m$^{-3}$) models underperformed. Based on the air quality index (AQI), the study region has experienced extremely unhealthy and healthy conditions, which may lead to the formation of visible haze and ultimately to the particulate component of smog. Generally, this study highlights the superior performance of deep learning approaches for urban air quality assessment. In conclusion, this study breaks new ground by applying and integrating MLR, CNN, and LSTM models in the study region. It will help in opening a promising direction for city-specific air quality modeling in any regional or local urban environment.

## Environmental significance

The relevance of temporal correlations in air pollution data and DNN is still poorly understood, despite notable advancements in air quality prediction. The present study estimated $PM_{2.5}$ concentrations in an urban environment in Peshawar using MLR, two machine learning architectures of CNN, and LSTM. To portray a more pertinent and optimistic current scenario, this study concentrates on the estimation and level of precision of $PM_{2.5}$ concentration in 2021. Furthermore, it provides information on how many dependent parameters affect $PM_{2.5}$ concentration. By comparing several deep learning algorithms, it adopts a novel approach to better accurately estimate $PM_{2.5}$ concentrations. The results would enhance model interpretation and open a promising direction for city-specific air quality modeling in any regional or local urban environment.

## 1 Introduction

The persistent contribution of harmful air pollutants has almost transformed the atmosphere into a global health crisis worldwide.[1–3] The spatiotemporal properties of these pollutants depend on air pollution. However, different meteorological conditions make it different from place to place. Particulate matter (PM) and trace gases are significant contributors to air pollution. They exist as an aggregate of solid and liquid particles. Furthermore, they exhibit heterogeneous sizes and morphologies suspended in the lower

[a] Department of Environmental Sciences, University of Peshawar, Peshawar 25120, Khyber Pakhtunkhwa, Pakistan
[b] Department of Elementary and Secondary Education, Peshawar, Khyber Pakhtunkhwa, Pakistan
[c] Department of Mathematics, Shaheed Benazir Bhutto University, Sheringal, Dir (Upper), Pakistan
[d] Department of Geology, Bacha Khan University Charsadda, Charsadda 24420, Khyber Pakhtunkhwa, Pakistan
[e] School of Automation, Wuxi University, 333 Xishan Avenue, Xishan District, Wuxi, Jiangsu Province, 214105, China
[f] Department of Physics, University of Peshawar, Peshawar, 25120, Pakistan. E-mail: khanalam@uop.edu.pk
[g] Department of Environmental Sciences, Shaheed Benazir Bhutto University, Sheringal, Dir (U), Khyber Pakhtunkhwa 18000, Pakistan. E-mail: allah.ditta@sbbu.edu.pk

atmosphere, arriving from different sources.[2,4,5] $PM_{2.5}$, or fine particles having an aerodynamic diameter of 2.5 μm, can enter the respiratory system deeply and cause adsorption.[6,7] It reduces visibility and causes hazy conditions in the lower atmosphere when its levels are elevated. According to recent research, particulate matter is one of the main air contaminants. Numerous factors, including industries, agriculture, urban growth, transportation networks, and rising infrastructure, contribute considerably to air pollution in the urban areas. Predicting air pollution in places without monitoring stations is crucial for the creation and application of preventative measures, in addition to measuring sites that facilitate air quality monitoring.[8,9]

The statistical methods do not account for the physical and chemical mechanisms of pollutants. There remains a strong relationship between meteorological variables and historical input data to estimate future concentrations.[2,10] They are mostly employed for analytical purposes, and as such, their ability to adequately represent the nonlinear dynamics in large datasets is restricted.[11] Rising $PM_{2.5}$ levels are significantly correlated with an increased mortality rate.[12–14] This makes it important to estimate air pollution concentrations to investigate an appropriate response. A significant number of uncertainties still exist and confront the problems of confounding variables and estimates. Uncertainties arise from the existence of confounding variables and unreliable estimates of fine exposures in a large urban environment.[15] To date, several investigations about the health impacts of long-term exposure have significantly relied on comparative assessments across different cities, limited by a small number of monitoring stations in the same premises. Still, these comparisons are subject to misclassification and estimations of exposures.

Deep learning has recently gained widespread recognition as a method that goes beyond the traditional boundaries of prediction by enabling the general-purpose extraction of complicated knowledge from large data.[16] Zheng *et al.*[17] predicted the concentration of $PM_{10}$ and nitrogen dioxide ($NO_2$) using the spatial characteristics of highways, factories, and parks. Ahmad *et al.*[11] estimated $PM_{2.5}$ concentration using an aerosol optical depth (AOD) and the normalized difference vegetation index (NDVI) as independent parameters using an artificial neural network (ANN) and a multiple linear regression (MLR) model.

Zohreh and Jamshid[18] used a different neural network with meteorological data to forecast $PM_{2.5}$ concentrations. Feng *et al.*[19] forecasted the next 24-hour $PM_{2.5}$ concentrations of air pollutants in Hangzhou, China, using both random forest (RF) and recurrent neural network (RNN). Although statistical methods can estimate $PM_{2.5}$ concentrations, machine-learning techniques tend to yield estimates that are more accurate. Additionally, by using different models, the estimation results may vary per study area. Comparing the accuracy of several machine learning methods for $PM_{2.5}$ estimation is therefore of great importance. It is crucial to acknowledge that, despite the publication of various machine-learning frameworks for estimating variant air pollutants, few of the models currently in use in air pollution research have been assessed in the intricate environments of emerging nations.

$PM_{2.5}$ concentration levels and emission patterns are much higher in nations like Pakistan. Despite existing advanced techniques of air quality assessment, the impacts and significance of correlations in air pollution data and deep neural networks (DNNs) are rarely known. Similarly, Pakistani megacities have not investigated the use of DNNs for this purpose. The purpose of this study is to estimate $PM_{2.5}$ concentrations in an urban environment of Peshawar City using three different methods: multiple linear regression (MLR), two machine learning architectures of convolution neural network (CNN), and long-short term memory (LSTM). To portray a more pertinent and optimistic scenario in terms of the current situation, the present study concentrates on the estimation and level of precision of $PM_{2.5}$ concentration in 2021. By acquiring the model validation values, the accuracy levels of selected models were compared. Furthermore, it differs from earlier research in that it provides information on how many dependent parameters, such as AOD, wind speed (WS), humidity, temperature (temp.), sulfur dioxide ($SO_2$), and $NO_2$, affect $PM_{2.5}$ concentration. By comparing several deep learning algorithms (MLR, CNN, and LSTM), this study adopts a novel approach to accurately estimate $PM_{2.5}$ concentrations in the study region. The results would preserve model quality while enhancing model interpretation. The results will open a promising direction for city-specific air quality modeling in any regional or local urban environment.

## 2 Experimental

### 2.1 Study area

The capital of Pakistan's Khyber Pakhtunkhwa (KPK) province, Peshawar, is one of the nation's largest cities. Due to its geographical location, industrial activities, and increasing vehicular emissions, Peshawar faces significant air quality challenges, particularly concerning the concentrations of $PM_{2.5}$.[4,8] Therefore, this region has been selected to estimate and predict the time series variation of $PM_{2.5}$ and its effects on different meteorological parameters and air pollutants. The geographic location of the study region was visualized in ArcMap 10.5, as illustrated in Fig. 1. The study area includes Peshawar's urban and peri-urban areas, with an emphasis on areas with heavy traffic and population density. Being one of the hottest places in Pakistan, Peshawar experiences hot summers, mild winters, and a monsoon season (from July to September), which can influence $PM_{2.5}$ levels.[20,21] The main reason for the accumulation of air pollutants in the study region is anthropogenic sources.[4,8] The major anthropogenic sources in the study region comprise vehicle emissions, brick kiln emissions, industrial activities, construction dust, biomass burning, and fossil fuel combustion, as well as natural sources of dust storms. Fig. 2 illustrates the detailed daily averaged values of selected air pollutants in this study across the study region. They also extend their role in the validation of various numerical models and approaches to improve the accuracy of output estimation.

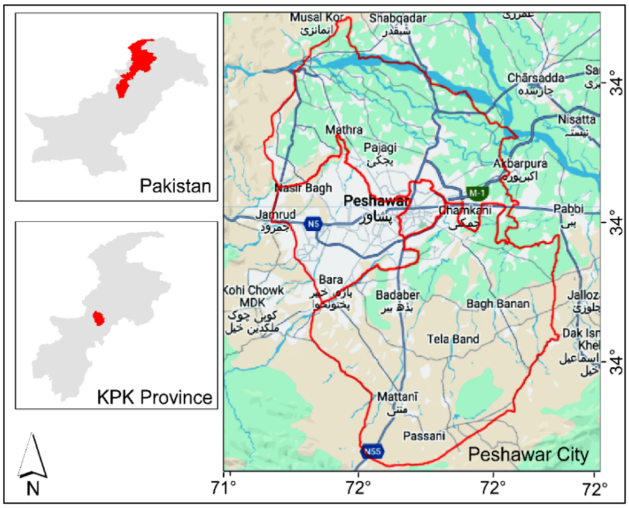

Fig. 1 The geographic location of Peshawar City. The closed red circle depicts the monitoring site for $PM_{2.5}$ and meteorological datasets.

### 2.2 Ground-based $PM_{2.5}$ measurement

The US Consulate in Peshawar's air quality monitoring network provided hourly and daily average $PM_{2.5}$ concentrations between January 2021 and December 2021.

To estimate dynamic and real-time $PM_{2.5}$ levels, the NowCast method was applied. This method requires at least 8 out of the previous 12 hourly valid readings to produce a reliable concentration value.[22] Furthermore, it enables the calculation of <24-hour average concentrations and supports the short-term assessment of air quality standards. Based on these levels, the NowCast-based $PM_{2.5}$ values are grouped into corresponding air quality index (AQI) categories. The necessary information can be obtained from the website https://www.airnow.gov. The meteorological data of humidity, temperature, and wind speed were obtained from the Pakistan Meteorological Department (PMD). Fig. 2 displays the temperature, humidity, $NO_2$, $SO_2$, $PM_{2.5}$, AOD, and AOD time series fluctuations for the year 2021.

### 2.3 Moderate resolution imaging spectroradiometer

A key component of contemporary Earth observation is the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, which is located on board the Terra and Aqua satellite platforms. It offers round-the-clock, worldwide monitoring of atmospheric and terrestrial conditions in 36 spectral bands, from the thermal infrared (14.23 μm) to the visible (0.41 μm). According to Ali *et al.*[20] and Hsu *et al.*,[23] the sensor records data at three different spatial resolutions: 250 m (2 channels), 500 m (5 channels), and 1 km (29 channels) at nadir. MODIS provides priceless spatiotemporal data for examining dynamic environmental processes, with a temporal revisit cycle of 1–2 days.[24] In accordance with recent research approaches, this work uses the worldwide application of MODIS data products to average the spatial heterogeneity of aerosol optical depth (AOD) concentrations and related uncertainties.[25,26] The daily averaged AOD for 2021 was obtained from MODIS Terra by combining two well-known algorithms.

The Dark Target (DT) algorithm is best suited for use over dark areas with surface reflectance around 0.15, such as those found on land and in water. The Deep Blue (DB) method provided high accuracy in AOD retrievals over land, which reduces the impact of albedo over bright surfaces while omitting regions covered by snow or dust.[20]

All obtained datasets were subjected to thorough preprocessing and quality screening in MATLAB and Python, utilizing a multi-platform methodology. Publicly available MODIS data utilized in this analysis can be obtained at NASA's Earth data portal at https://www.earthdata.nasa.gov.

### 2.4 Ozone monitoring instrument

A satellite-borne device called the Ozone Monitoring Instrument (OMI) was created to track several atmospheric variables, including ozone.[8,24] It is on board the Aura satellite of the National Aeronautics and Space Administration (NASA) and is an essential component of the Earth Observing System (EOS). The launch date of the OMI was July 15, 2004. To research air quality and climate change, it examines atmospheric aerosols, including their optical thickness and absorption characteristics.[24] In this paper, measurements of atmospheric pollutants of $NO_2$ and $SO_2$ are retrieved in the form of HDF5 formats with products like OMNO2d and OMSO2e. The data sets are preprocessed in ArcMap 10.5, MATLAB, and Python. The required data can be obtained from https://www.earthdata.nasa.gov.

### 2.5 Linear regression

The statistical models of MLR can handle large datasets and can largely remove the uncertainties.[27] In this paper, MLR is used to estimate $PM_{2.5}$ as an output while MODIS-retrieved AOD, OMI-retrieved $NO_2$, and $SO_2$, and meteorological parameters of temperature, wind speed, and humidity are used as input parameters, see eqn (1). Using the MS Excel, Origin Lab, MATLAB, Python, NumPy, and Pandas libraries, the datasets displayed in this equation are first cleaned to eliminate unnecessary values, such as negative values and Not a Number (NaN).[11]

$$PM_{2.5} = \beta_0 + \beta_1(\text{AOD}) + \beta_2(\text{NO}_2) + \beta_3(\text{temp.}) + \beta_4(\text{WS}) + \beta_5(\text{H}) + \beta_6(\text{SO}_2) \quad (1)$$

In this equation, $\beta_{1-7}$ represent the regression coefficients. The results obtained from MLR are explored to assess the effects of inputs on the estimation of $PM_{2.5}$. The adequacy checking was employed for the proposed model. It ensures that the model is statistically sound and reliable for estimation. It consists of several steps, like checking linearity, independence, homoscedasticity, normality of residuals, model fitness, and significance of estimation results.[11] Visually, it can be assessed through the graphical illustration of standardized residuals against estimated values.[27–29]

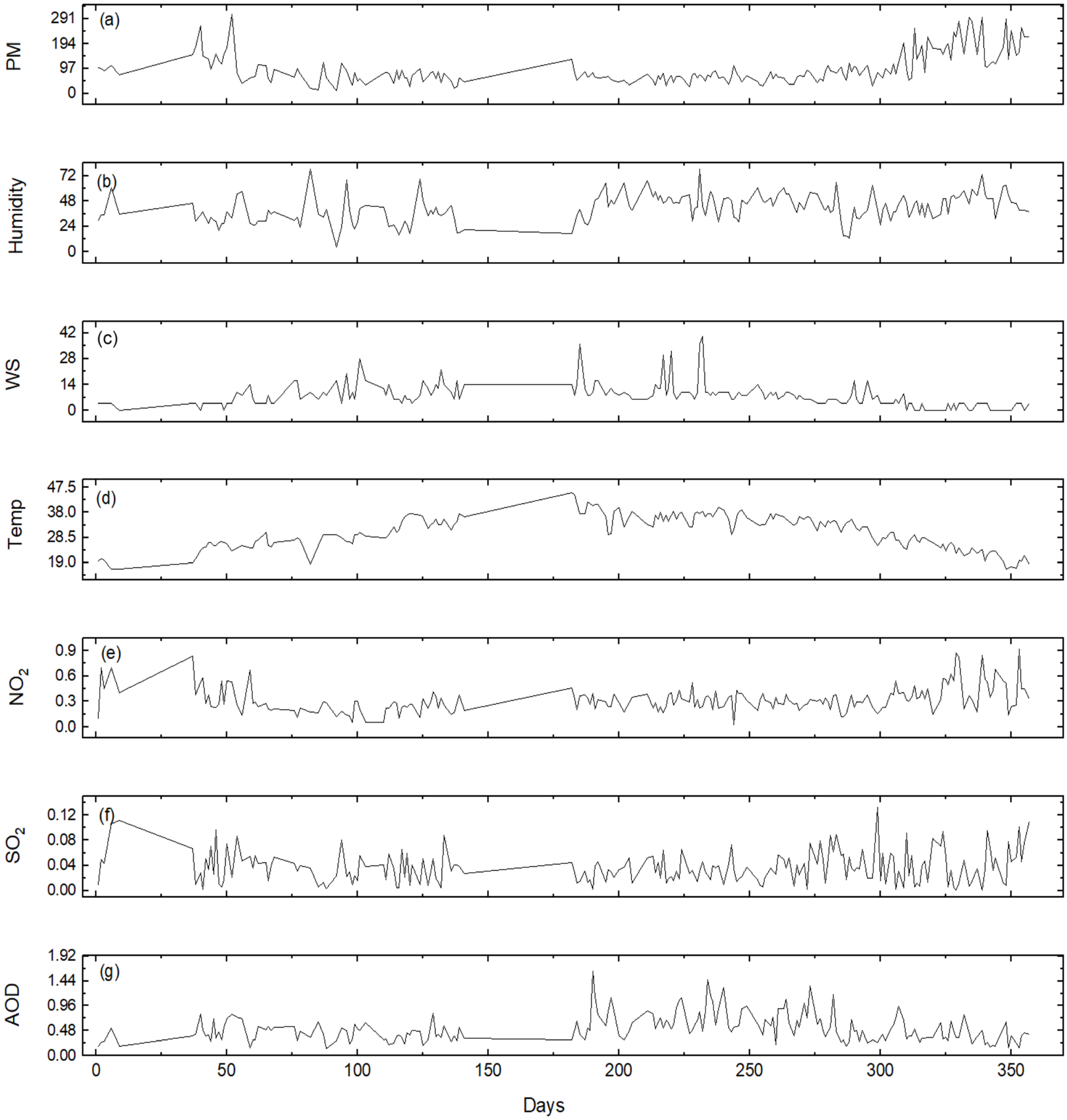

Fig. 2 Time series status of air pollutant variables, (a) PM, (b) humidity, (c) wind speed, (d) temperature, (e) $NO_2$, (f) $SO_2$, and (g) AOD in the study area for the year 2021.

### 2.6 Deep learning

One of machine learning's subclasses is deep learning.[10] It comprises neural networks with multiple (deep) layers. These networks are useful to learn complex patterns of datasets in a model.[28] Deep learning models can automatically detect features without manual input. They can execute any task for manual feature extraction. This property of machine learning makes it preferable over old and ordinary machine learning algorithms, which are dependent on handcrafted features.[10,16] The primary architectures of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) are employed in this work.

### 2.7 Convolutional neural network

The architecture of CNN employs convolutional deep learning to execute the required key features. The model automatically learns these features. The model optimally adjusts these features to obtain the required output.[31] A typical CNN model is a mathematical structure made up of three layers. The convolutional layer, the pooling layer, and the fully connected layer are the three essential parts of the Convolutional Neural Network (CNN) architecture used in this investigation (Fig. 3). Hierarchical patterns are automatically extracted from the input data by the first convolutional and pooling layers, which oversee

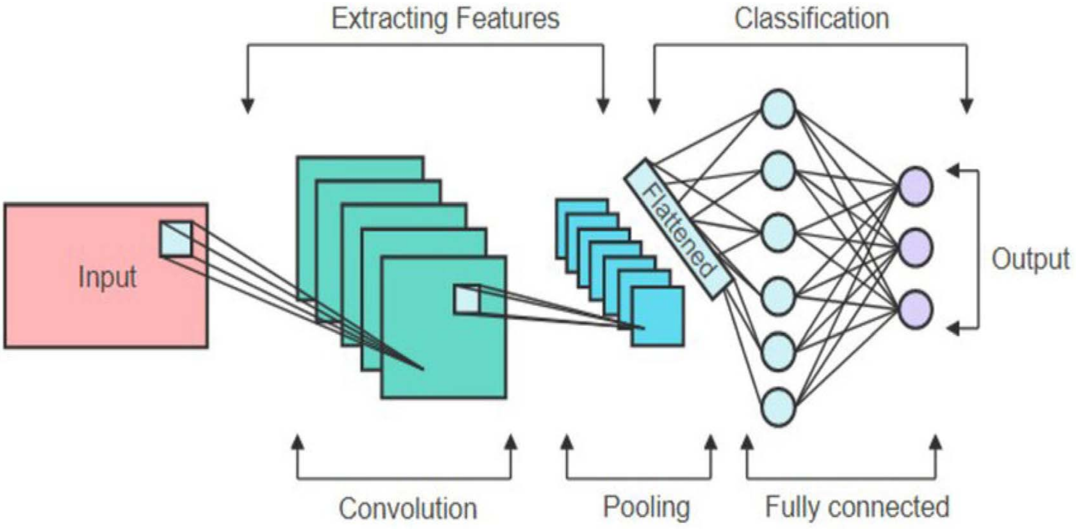

Fig. 3 The schematic of convolutional neural networks.[30]

feature engineering and learning. The final classification or regression output is then generated by the final fully connected layer using these learned features and weighing them.[10,32] The convolutional layer, which carries out two crucial mathematical operations, lies at the heart of its architecture. It starts by detecting spatial features using a convolution process. A non-linear transformation employing a Rectified Linear Unit (ReLU) activation function quickly follows, adding non-linearity and boosting the representational capability of the model.[33,34] The CNN model may effectively extract the spatial features from inputs (pollutants). This may lead to a high level of accuracy. As a result, the complexity of convolutional layers is reduced to

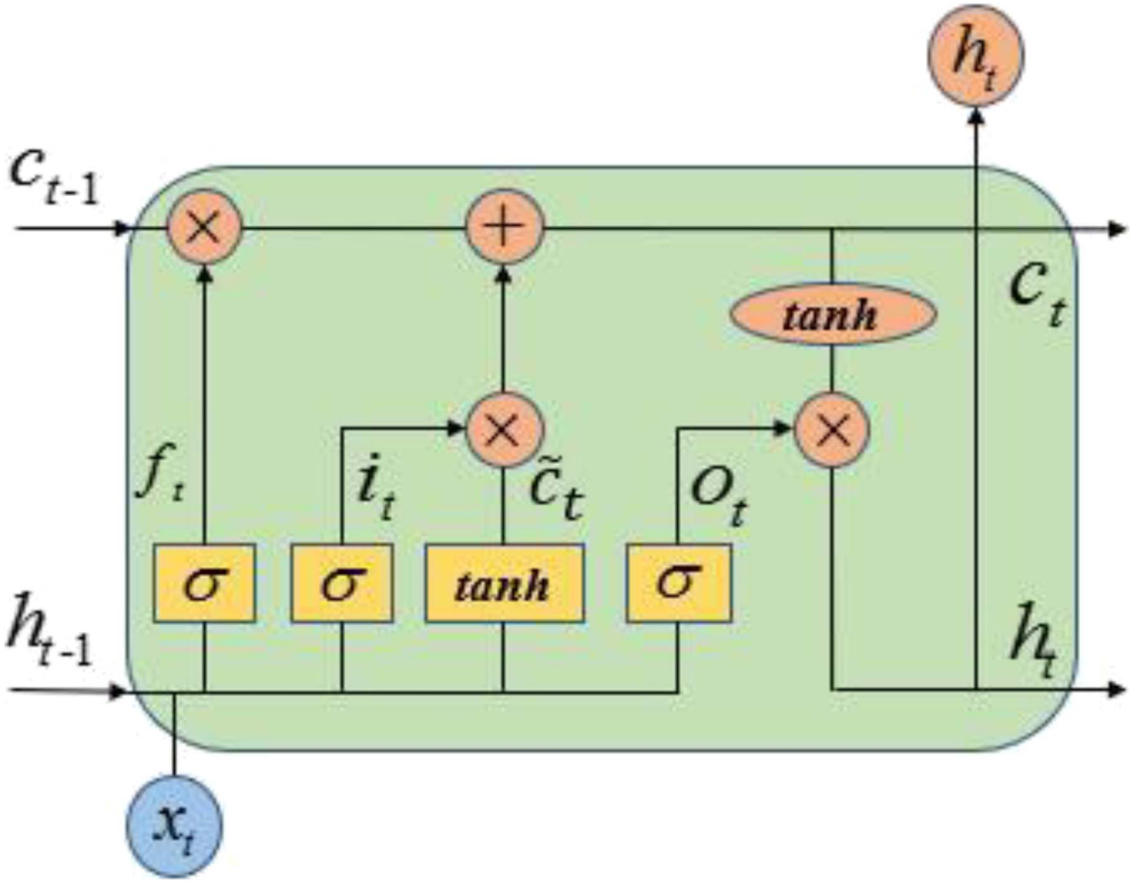

Fig. 4 The architecture of long short-term memory.[31]

a desirable level. It is also used here for time series prediction. This model will aid in increasing $PM_{2.5}$ estimation accuracy. CNN is utilized in this work to extract the time series data's features based on $PM_{2.5}$ prediction.[35] $PM_{2.5}$ as an output is attained by a sliding filter (convolution kernel), or weight matrix, over a specific portion of $PM_{2.5}$ as an input. The input features of the CNN model consist exclusively of $PM_{2.5}$ values and do not include AOD, temperature, wind speed, or others. It contains a sliding window of 24 hours from a univariate $PM_{2.5}$ time series data. Furthermore, at each point, it is computed as the dot product between a convolution of input data and a selected filter. In this way, the model allows an environment to learn filters that may further enable it to recognize patterns in the data input.

### 2.8 Long- and short-term model

Many researchers have used an RNN model to tackle the problems of the shrinking and exploding RNN model. It prevents instability during the training process. For example, Li *et al.*[31] describe three gate structures of the LSTM model. As shown in Fig. 4, it is intended to include three gate structures: input, forget, and output. The fourth key element is also called the memory cell in the LSTM model. It is also called the cell state. It drives an unbroken flow across multiple LSTM blocks. This connection features minimal linear interactions.[19,36] The term $x_t$ acts as the input at the current time of the LSTM model; likewise, $(h_{t-1})$ is the state of the hidden layer at the previous time. Similarly, $(c_{t-1})$ is the cell state, which acts as the third input of a given LSTM block. It is also of great importance to mention here that each output of a gate is a vector and is of the same size as the hidden vector $(h_t)$. The working mechanisms behind these features are outlined in the following equations:[31]

$$f_t = \delta(w_{\mathrm{xf}}x_t + w_{\mathrm{hi}}h_{t-1} + b_{\mathrm{f}}) \tag{2}$$

$$i_t = \delta(w_{\mathrm{xi}}x_t + w_{\mathrm{hi}}h_{t-1} + b_{\mathrm{i}}) \tag{3}$$

$$o_t = \delta(w_{\mathrm{xo}}x_t + w_{\mathrm{ho}}h_{t-1} + b_{\mathrm{o}}) \tag{4}$$

$$\mathrm{pc} = \tanh(w_{\mathrm{xc}}x_t + w_{\mathrm{hc}}h_{t-1} + b_{\mathrm{c}}) \tag{5}$$

$$c_t = f_{\mathrm{i}}c_{t-1} + i_{\mathrm{t}}\mathrm{pc} \tag{6}$$

$$h_t = o_{\mathrm{t}}\tanh(c_t) \tag{7}$$

Eqn (2)–(4) describes the forget, input, and output gates, respectively. In each equation, the sigmoid activation function $(\delta)$ is used for non-linearity.

Eqn (5) and (7) are used to derive the potential values in the memory cell and the hidden state. Likewise, the activation function is represented by tanh. The values of the hidden state and the current cell memory $(c_t)$ are shown in eqn (6). These values are computed over element-wise multiplication and are disclosed outside the block.[10,31] It is used against the forget gate's output instead of the repeated matrix to prevent RNN issues. This stage of the said model gives LSTM preference over RNN. In this study, the LSTM model is constructed and optimized. The LSTM model also uses univariate input data comprising $PM_{2.5}$ values, but in this case, the sequence length is shorter (5 hours) as compared to the CNN. The LSTM architecture includes a single layer with 150 units and ReLU activation. A dense output layer having one unit follows it. The entire $PM_{2.5}$ data set is split into a test dataset of 15% and a training dataset of 85%. Furthermore, the model is made to fit all training datasets. The estimation stage is based on test datasets at each step of the model. In this study, the model runs for 100 epochs. At the end, the model is evaluated using statistical measurement parameters, given in eqn (2)–(7). The model is developed on the Python platform.

### 2.9 Evaluation of statistical indices

Several statistical indices are used to assess the effectiveness of the model; some of them are well-known, such as the root mean square error (RMSE), correlation coefficient ($R$), and coefficient of determination ($R^2$). Eqn (8)–(10) provide these indicators.[11,37] The model's correctness is evaluated in comparison using the values of these indices. RMSE calculates the deviation from the model's estimated values in a variety of performance criteria.[11] The model's influence will be better if these indices have lower values. On the other hand, $R^2$ shows how much the model's estimated and observed values differ from one another. Better model performance is indicated by higher $R^2$ values.[37–39] The overall coefficient of determination is also computed in order to assess the importance and strength of the estimation of various models.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i - E_i)^2}{n}} \tag{8}$$

$$\mathrm{MSE} = \frac{I}{n}\sum_{i=1}^{n}(O_i - E_i)^2 \tag{9}$$

$$R = \frac{\sum(x - \breve{x})(x - \breve{x})}{\sqrt{\sum(x_i - \breve{x})^2 - (y_i - \breve{y})^2}} \tag{10}$$

where $O_i$, and $E_i$ represent observed and estimated values, respectively. Similarly, $x_i$ and $\check{x}$ represent the value and mean of the $x$ variable, respectively, in eqn (10). Likewise, $y_i$ represents the value of the $y$ variable, and $\check{y}$ represents the mean value of the $y$ variable.

### 2.10 Air quality test

The different temporal values and their associated variations are used to assign $PM_{2.5}$ levels in different categories. They are categorized as Hazardous, Unhealthy, Very Unhealthy, Unhealthy for Sensitive Groups (USG), Good, and Moderate. These classifications are employed in calculating the air quality index (AQI), given in eqn (11).[8]

$$\mathrm{AQI} = \frac{I_{\mathrm{HI}} - I_{\mathrm{LO}}}{\mathrm{BP}_{\mathrm{HI}} - \mathrm{BP}_{\mathrm{LO}}}(C_{\mathrm{P}} - \mathrm{BP}_{\mathrm{LO}}) + I_{\mathrm{LO}} \tag{11}$$

where $I_{HI}$ represents the AQI value in relation to $BP_{HI}$, $I_{LO}$ represents the AQI value against $BP_{LO}$, $BP_{HI}$ represents the concentration breakpoint ($\geq C_P$), and $C_P$ represents the truncated concentration of a pollutant (P). These datasets are validated through standard consistency checks. In this study, daily, monthly, and seasonal temporal variation of $PM_{2.5}$ levels at Peshawar City were analyzed to identify the most polluted episodes.

## 3 Results and discussion

In this research work, the estimations of $PM_{2.5}$ concentrations are performed using atmospheric parameters of AOD, temperature, wind speed, $SO_2$, and $NO_2$ averaged diurnally for the year 2021. In this work, MLR, CNN, and LSTM models were used to estimate $PM_{2.5}$. The metrics $R$, $R^2$, and RMSE were used to assess the performance of the suggested models. These measures shed light on how well CNN and LSTM's deep learning and MLR models work. The study's conclusions include a thorough comparison of the previously described models. The chosen datasets were divided into 80% for training and 20% for testing to facilitate machine learning (*e.g.*, CNN and LSTM). With each particular input variable, the data will be fed into a variety of models that have been chosen for each method.

Table 1 The statistical results of the multiple linear regression model

| Model inputs | Notation | RMSE | $R$ | $R^2$ |
|---|---|---|---|---|
| Pollutants<br>Meteorological parameters | MLR | 44.35 | 0.68 | 0.46 |

### 3.1 $PM_{2.5}$ estimation through MLR

Using the diurnal average values of AOD, temperature, wind speed, $SO_2$, and $NO_2$ for the year 2021 as input variables, the MLR model focuses on estimating $PM_{2.5}$ concentrations. The MLR model used AOD, trace gases of $NO_2$, $SO_2$, and meteorological values of temperature, wind speed, and humidity. Several different MLR models were employed, but the best-fit model with the highest $R$-value of 0.68 was selected, as given by eqn (12).

$$PM_{2.5} = 193.49 + 15.54(\text{AOD}) - 4.21(\text{temp.}) - 1.65(\text{WS}) + 0.08(\text{H}) - 267.28(SO_2) - 0.00(NO_2) \quad (12)$$

The MLR model's results are shown in Table 1. The findings indicate that, of the three models chosen, the MLR model that used aerosol optical characteristics, trace gases, and meteorological data as input had the least striking outcomes. With an RMSE of 44.35 μg $m^{-3}$, the MLR's inaccuracy was also greater than that of the other two CNN and LSTM models. However, the MLR model performed better than LSTM due to its utilization of meteorological characteristics as input with a lower error. Likewise, the values of $R$ and $R^2$ were found to be lower than those of CNN and LSTM. Additionally, the $R$ value (0.68)

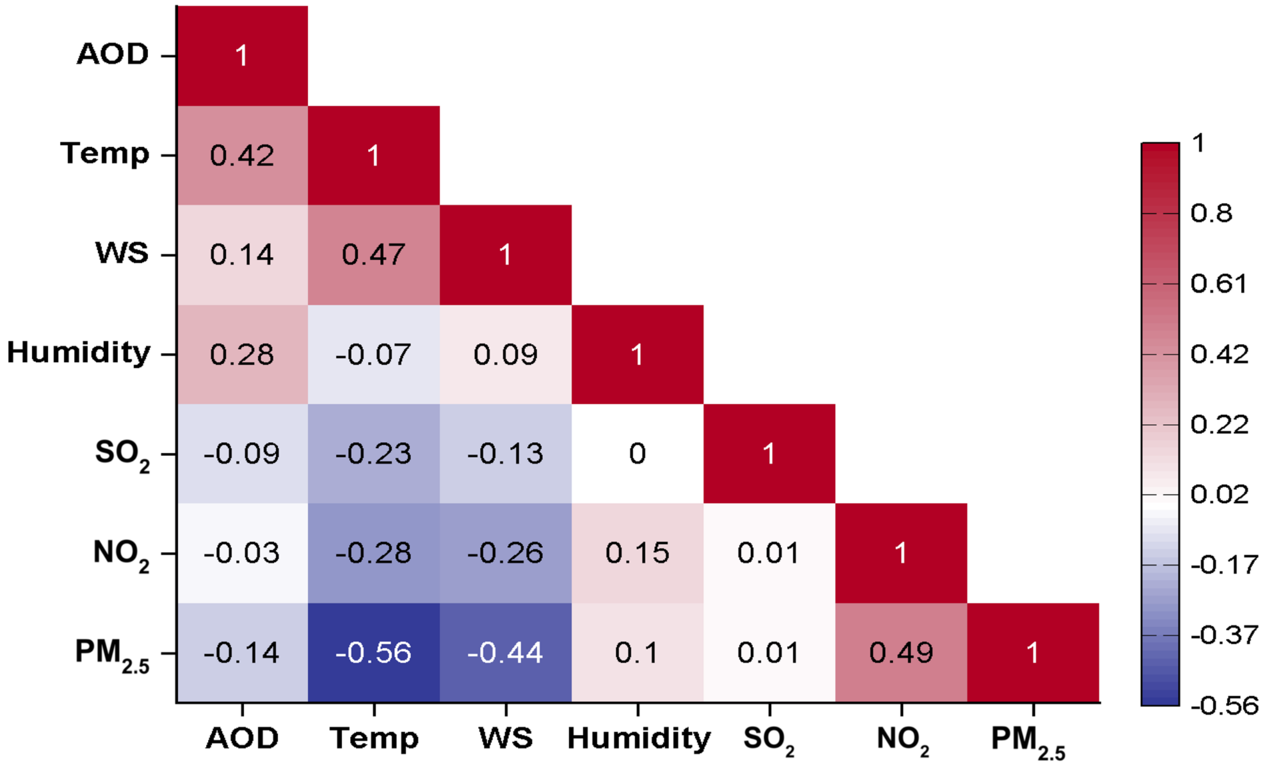

Fig. 5 Heat map of the selected air pollutant variables.

Table 2 Statistical results of MLR analysis

| | Coefficients | *p*-Value | Standard error |
|---|---|---|---|
| Intercept | 193.49 | 0.00 | 24.37 |
| AOD | 15.54 | 0.28 | 14.28 |
| Temperature | −4.21 | 0.00 | 0.67 |
| Wind speed | −1.65 | 0.01 | 0.58 |
| Humidity | 0.08 | 0.75 | 0.27 |
| $SO_2$ | −267.28 | 0.04 | 127.27 |
| $NO_2$ | 0.00 | 0.00 | 0.00 |

between estimated and observed values was calculated, reflecting a positive and significant relationship. The correlation heat map, based on Pearson's correlation, illustrates different degrees of association between $PM_{2.5}$ and input variables, as shown in Fig. 5. The figure shows that $PM_{2.5}$ has a weak negative relationship with AOD, but a positive relationship with humidity and $SO_2$, with coefficient values of −0.14, 0.1, and 0.01 ($p < 0.05$), respectively. Likewise, a negative strong and moderate relationship of $PM_{2.5}$ with temperature (−0.56, $p < 0.05$), and wind speed (−0.44, $p < 0.05$) was observed. The results in Fig. 5 clearly illustrate that AOD, temperature, and wind speed are negatively correlated with $PM_{2.5}$ concentration. Furthermore, a positive and moderate relationship of 0.49 (with $p < 0.05$) between $PM_{2.5}$ and $NO_2$ can also be observed in the figure. Apparently, the main affecting factors of $PM_{2.5}$ are temperature and $NO_2$. The detailed statistical outcomes of the MLR model are given in Table 2.

Fig. 6 displays the time series of $PM_{2.5}$ concentrations during 2021. The combined calculated and observed daily average $PM_{2.5}$ values are displayed in Fig. 6(a) and (b), respectively, and Fig. 6(c) compares them. Likewise, the estimated values and the corresponding standard deviation values are displayed in Fig. 6(d). Geography, land use and land cover changes, and local meteorological conditions may also have an impact on this fluctuation pattern.[40] Similar findings were confirmed by Liu *et al.*,[41] Zhao *et al.*,[42] Xin *et al.*,[43] and Ahmad *et al.*[11] Similar

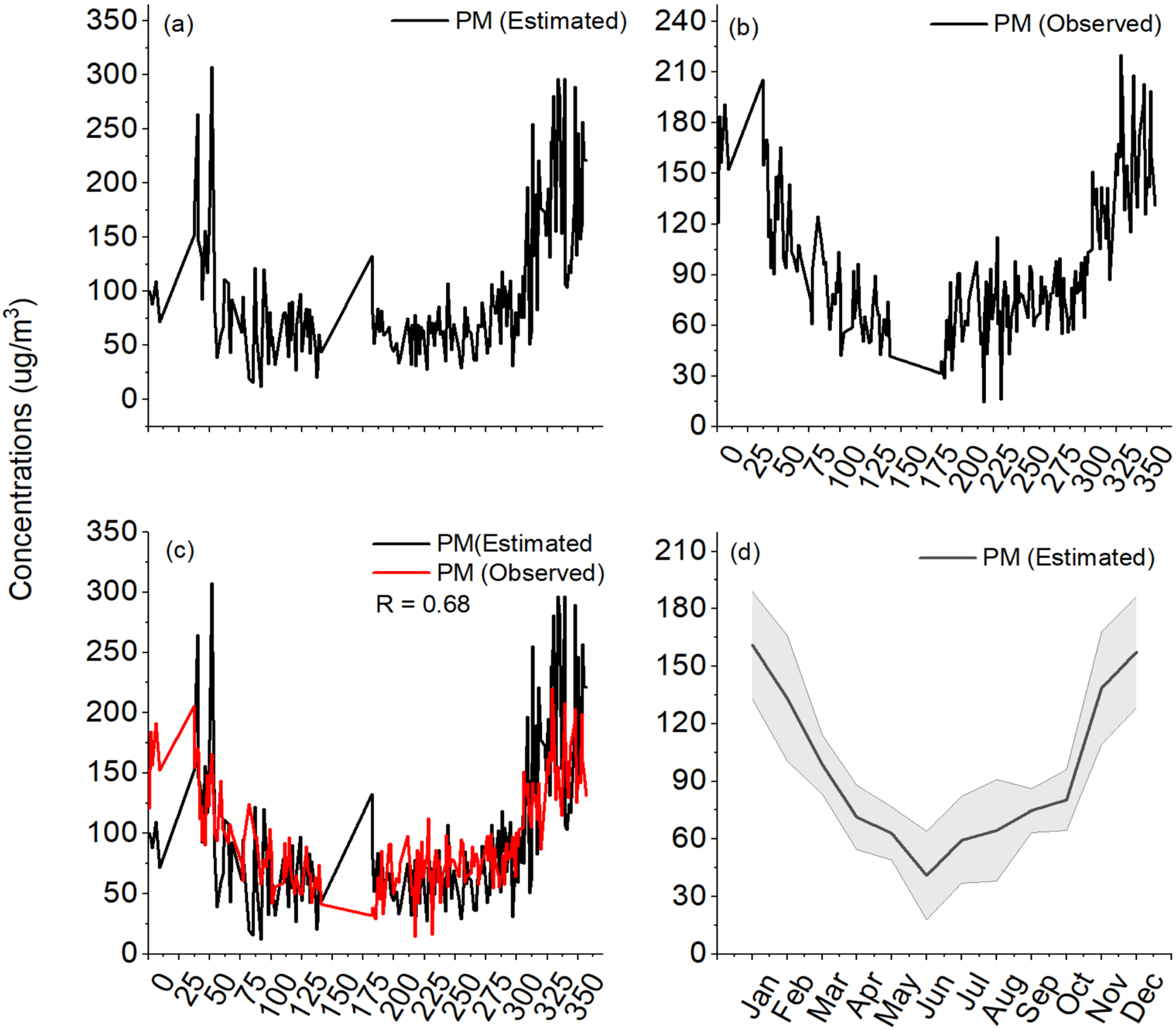

Fig. 6 Time series variation of $PM_{2.5}$ (a) estimated, (b) observed, (c) correlation with observed $PM_{2.5}$, and (d) estimated PM with standard deviation (shaded portion of the graph).

Table 3 Comparison of the estimation performance of different machine learning models

| Model inputs | Notation | Train/test data split | $R$ | $R^2$ | RMSE |
|---|---|---|---|---|---|
| $PM_{2.5}$ | CNN | 80–20 | 0.89 | 0.79 | 34.89 |
| $PM_{2.5}$ | LSTM | | 0.87 | 0.75 | 51.93 |

approaches were used by Nguyen *et al.*[44] in their investigation to estimate ground-based PM concentration in Vietnam using the same regression techniques. They discovered that the concentration of AOD on the ground and satellites had a strong association ($R = 0.69$). To predict $PM_{2.5}$ concentrations, satellite AOD data alone are insufficient.

Therefore, meteorological and trace gas data have proven their importance in this regard.[45] The findings of MLR conclude that overall agreement between the measured and the estimated values is reasonable; however, the estimated values recurrently underestimated the $PM_{2.5}$ level. Particulate matter is thought to be one of the major air pollutants with the worst effects on human health among all of them. According to the findings, there is a negative link between weather variables (temperature and wind speed) and $PM_{2.5}$ concentrations, whereas a positive correlation between $PM_{2.5}$ and $SO_2$ and $NO_2$ indicates that air pollution levels primarily regulate $PM_{2.5}$ concentrations.[13] According to Mirzaei and Zandkarimi,[46] this could indicate that strong wind speeds are capable of restoring air purity and lowering $PM_{2.5}$ concentrations above lower atmospheric boundary levels. Furthermore, it is found that the temperature inversion during winter may contribute prominently due to the restriction of suspended particles in the lower atmospheric layer.

### 3.2 $PM_{2.5}$ estimation and accuracy assessment through CNN and LSTM

For the comparison of different machine learning models, this study chooses two neural networks, CNN and LSTM. This study estimated the daily average $PM_{2.5}$ concentrations in Peshawar between January 1st, 2021, and December 31st, 2021, using temporally aware CNN and LSTM models. Table 3 provides the estimation performance of various evaluation indicators. The table makes it evident that CNN performs better than LSTM. Specifically, the projected values of the suggested CNN model closely resemble the $PM_{2.5}$ levels that were found in Peshawar. With the lowest RMSE and highest $R^2$ values, the CNN model provides the most accurate daily $PM_{2.5}$ estimations, as seen in Table 3. This demonstrates that the suggested CNN architecture has a higher accuracy than the LSTM architecture, indicating that it is better able to identify changes and patterns throughout the estimation process. Furthermore, the CNN and LSTM performance metrics demonstrate that the CNN deep neural network model's estimation accuracy is superior to that of the LSTM models. Since the distribution of $PM_{2.5}$ measuring stations in Peshawar is not uniform, it shows a spatial trend of

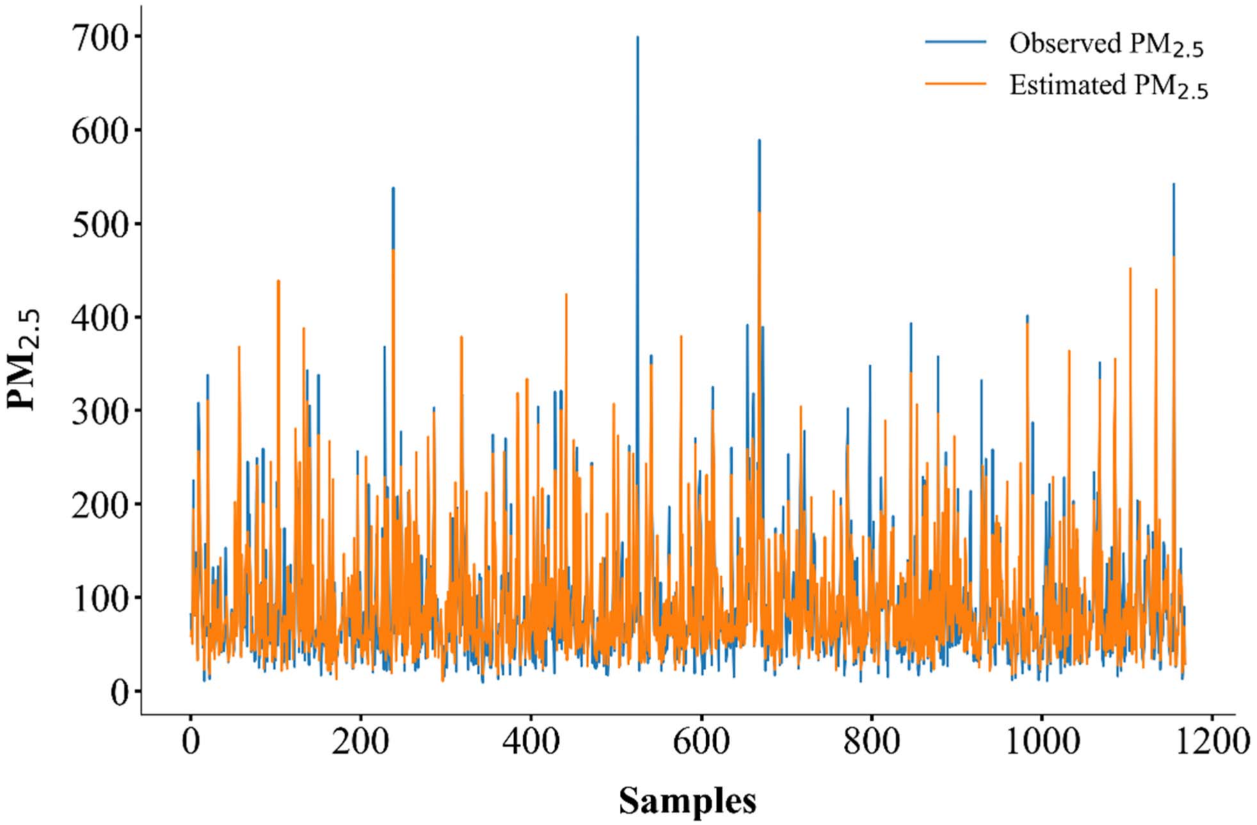

Fig. 7 Comparison of observed and estimated values of $PM_{2.5}$ using the CNN model.

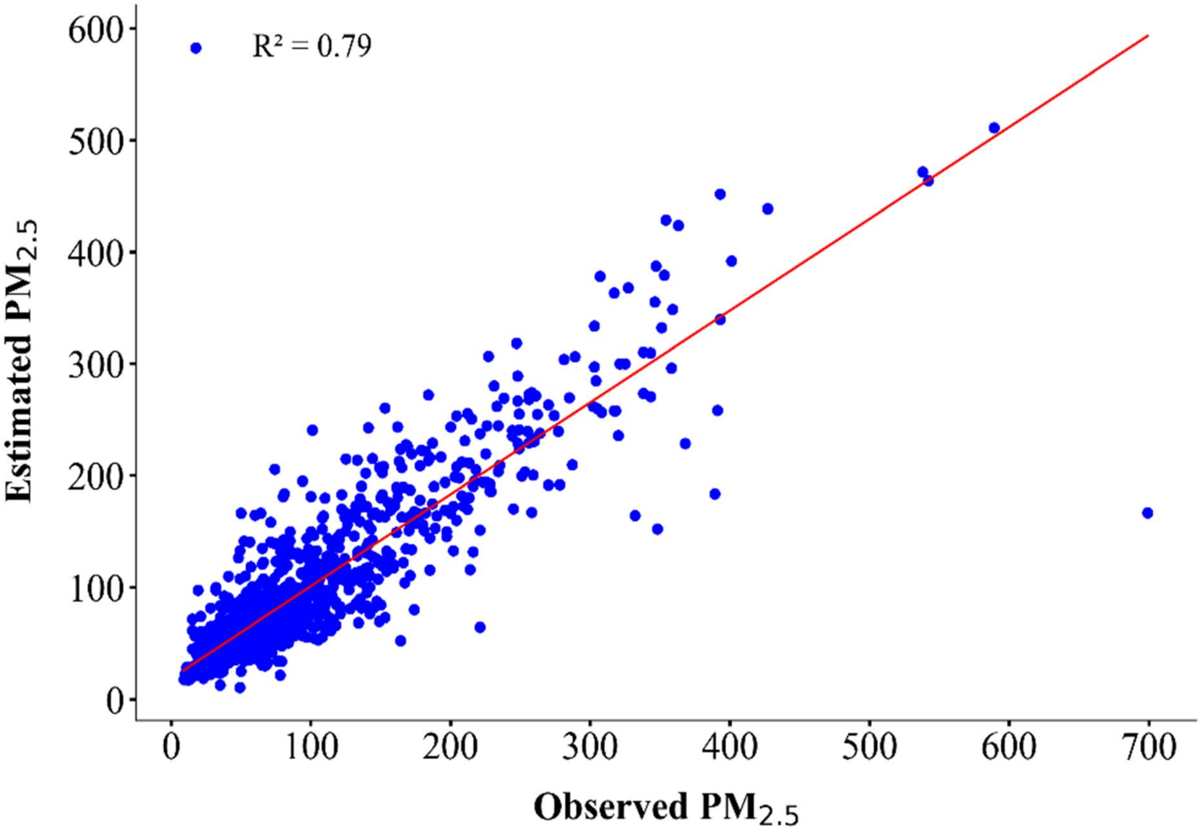

Fig. 8 Correlation between estimated and observed values of $PM_{2.5}$ using the CNN model.

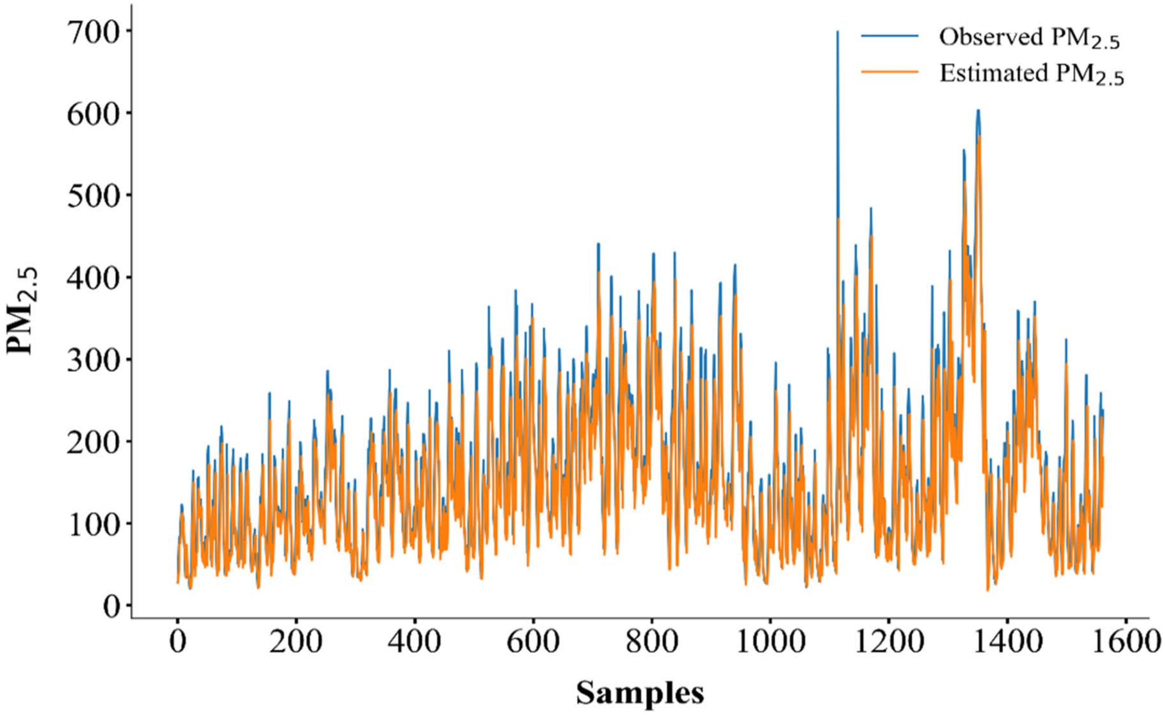

Fig. 9 Comparison of observed and estimated values of $PM_{2.5}$ using the LSTM model.

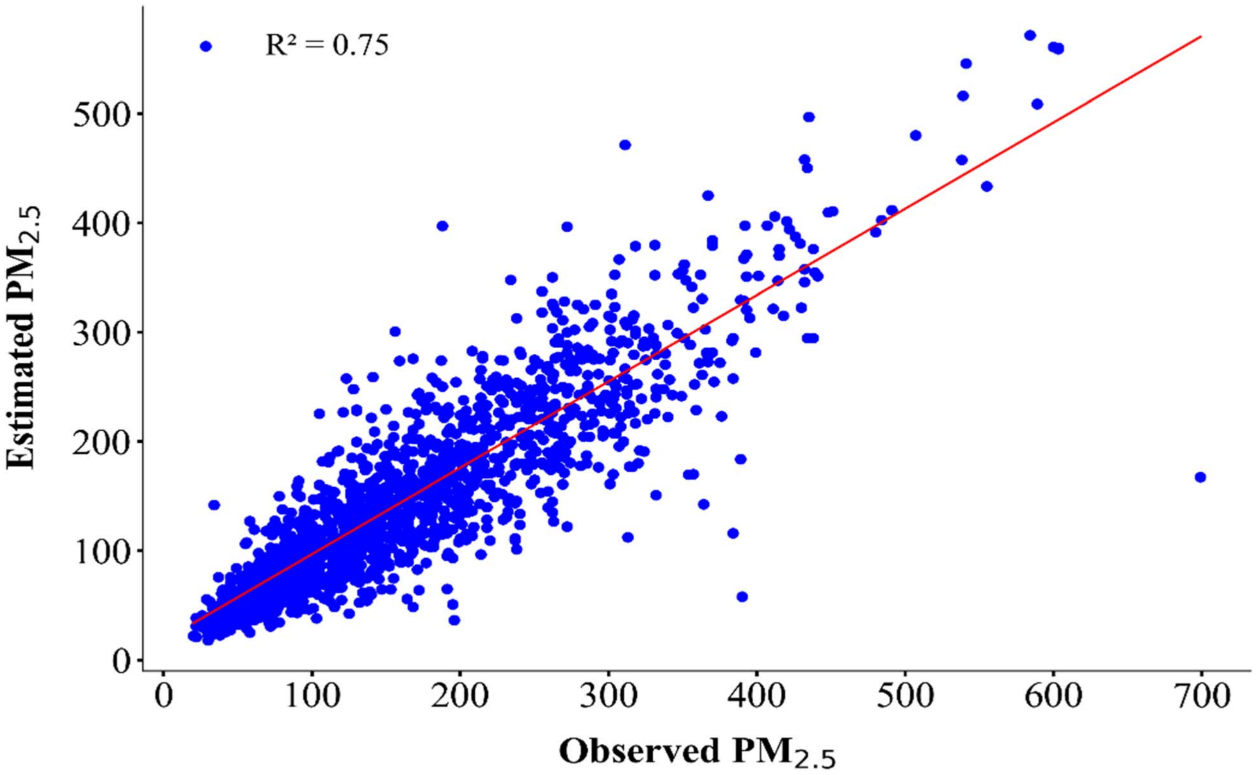

Fig. 10 Correlation between estimated and observed values of $PM_{2.5}$ using the LSTM model.

Table 4 Performance of the MLR, CNN, and LSTM models. The bold values indicate the best values and statistically significant results

| Model | $R$ | $R^2$ | RMSE | MSE |
|---|---|---|---|---|
| MLR | 0.68 | 0.46 | 44.35 | 1967.66 |
| CNN | **0.89** | **0.79** | **34.89** | **1217.69** |
| LSTM | 0.87 | 0.75 | 51.93 | 2696.73 |

higher concentration. Peshawar, as an urban area, has a high population density. The emission sources of $PM_{2.5}$ and regional morphology may disturb the model's performance. All these factors have strongly affected the accuracy of spatial information.[8,51] Fig. 7 displays the estimated CNN performance, whereas Fig. 9 displays the LSTM model. Over the whole research region, the projected values closely match the observed values. For higher observed values, the higher accuracy of the model illustrates that it can deal with non-linear data well. Likewise, it indicates that the model avoids overfitting as well.[37,47] Furthermore, it demonstrates that the CNN model can enhance LSTM estimation and execute features more efficiently. Therefore, we can say that the CNN model has a comparatively good estimation compared to the LSTM model in Peshawar City.

In addition, the $PM_{2.5}$ concentration estimated using the CNN model illustrates a strong relationship with observed values of $R = 0.89$, and $R^2 = 0.79$, as shown in Fig. 8.

The results given in Table 3 reveal that the CNN model outperformed the others by executing the lowest RMSE of 34.89 μg $m^{-3}$. The temporal assessment of LSTM reveals that estimated values are similar to the observed values over the whole study region, as shown in Fig. 10. Furthermore, their correlation yields an $R$ value of 0.87 and an $R^2$ value of 0.74. These results show that the significant agreement between observed and estimated values is enough for the accuracy of the model, which illustrates that it can deal better with non-linear data. As shown in Table 4, the estimation accuracy of the CNN model is higher

Table 5 Distribution of air quality levels (SG stands for sensitive groups)[29]

| AQI range | Category | $PM_{2.5}$ (μg $m^{-3}$) | Health impacts of air quality |
|---|---|---|---|
| 0–50 | Good | 0–12 | Satisfactory |
| 51–100 | Moderate | 12.1–35.4 | Accepted, but little concern for SG |
| 101–150 | USG | 35.5–55.4 | Health effects by SG |
| 151–200 | Unhealthy | 55.5–150.4 | Health effects for everyone |
| 201–300 | Very unhealthy | 150.5–250.4 | Health alerts |
| 301–500 | Hazardous | ≥250 | Emergency conditions |

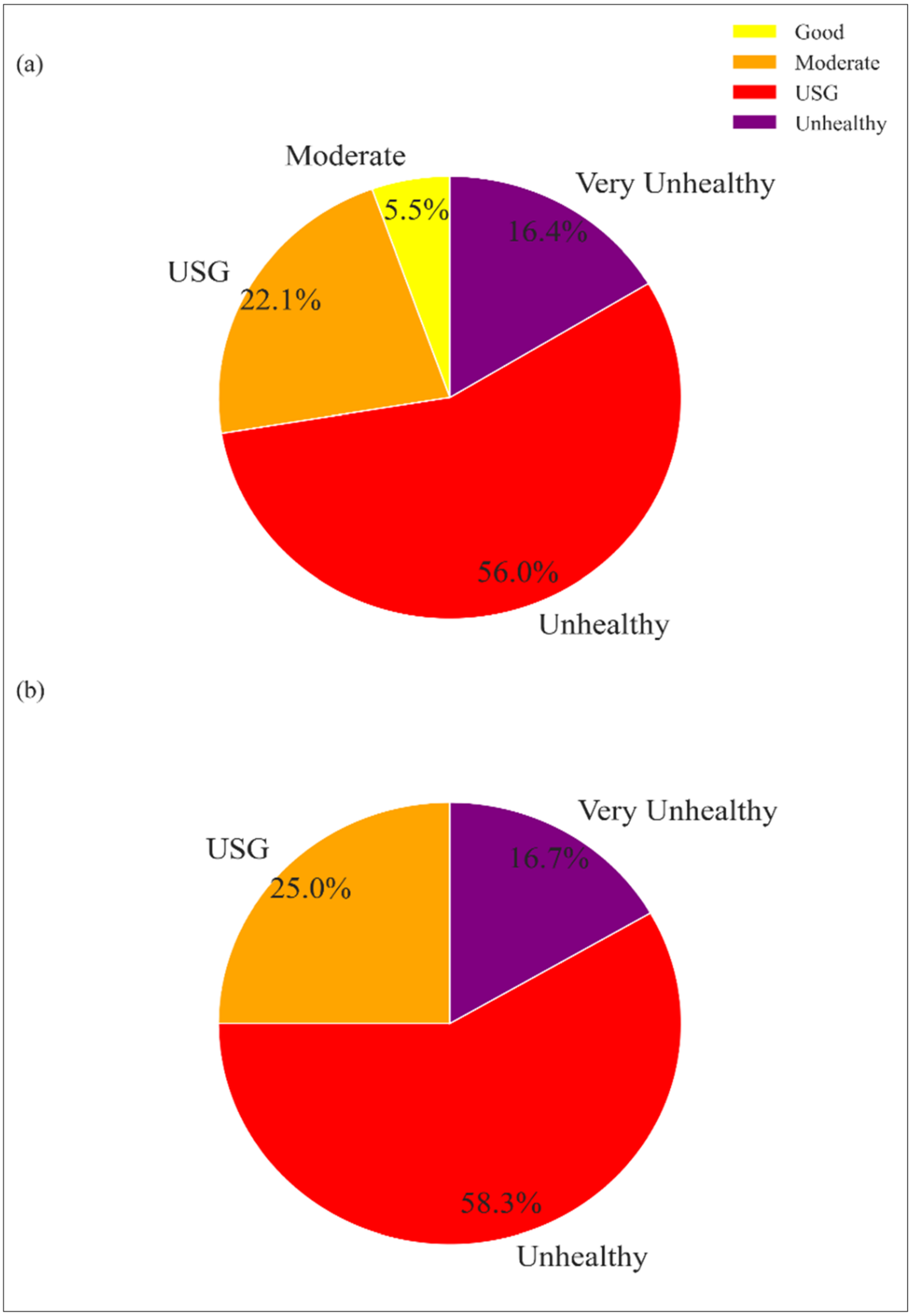

Fig. 11 Daily (a), and monthly (b) air quality status with certain air quality classifications in 2021 for Peshawar City.

than that of the MLR and LSTM models, according to the comparison results of a few chosen models. The CNN model lowers the estimation error and fully accounts for the spatio-temporal diffusion and correlation of the $PM_{2.5}$ concentration. Consequently, it can be said that the CNN model outperforms MLR and LSTM in terms of estimation accuracy. Furthermore, the suggested models can be used to assess current air quality trends based on historical data, making them useful instruments for the study, sustainable planning, and control of air pollution and related air quality. However, different factors (predictors) create issues during the estimation of $PM_{2.5}$ concentration, which makes it complicated.

### 3.3 Air quality assessment and state of $PM_{2.5}$ levels

Air quality is frequently assessed by analyzing AQI values. In this study, AQI is subjected to $PM_{2.5}$ concentrations only. The values of AQI are assigned to different values at different air quality levels from 0 to 500, as given in Table 5. However, these values may be different for different sources. The most well-known and frequently applied of these are the Environmental Protection Agency (EPA) standards of the United States and the World Health Organization (WHO). The WHO has modified the $PM_{2.5}$ air quality criteria to be 5 μg m$^{-3}$ for annual averages and 15 μg m$^{-3}$ for 24-hour (daily) averages. According to these recommendations, the annual average of $PM_{2.5}$ should be less than 5 μg m$^{-3}$, and it should never surpass 15 μg m$^{-3}$ on any one day.

Though these updated guidelines for $PM_{2.5}$ exposure are not easy to achieve in several parts of the world because 5 μg m$^{-3}$ can be easily achieved through a natural background alone.[48] Furthermore, $PM_{2.5}$-based assessment of air quality can be very different across different counties. In urban areas, it is strongly linked with massive traffic and industries that produce large amounts of $PM_{2.5}$. The land use and land cover changes and industrial activities may experience elevated levels of $PM_{2.5}$.[8] In most regions, it is considered one of the main air pollutants, causing adverse effects on air quality.

If this trend continues for a long time, it will be injurious to air quality. This study utilizes both 24-hour (daily) and monthly average concentrations of $PM_{2.5}$ using AQI to investigate the air quality of Peshawar City, as shown in Fig. 11. This study investigated the status of air quality and observed that AQI ranged from 1 to 594, with an average of 94 in 2021. The monthly maximum amount of AQI was observed during the winter season in the entire study period. The highest amount of AQI values falls within the category of healthy and very unhealthy, while the maximum of the monthly mean occurred in January 2021, and categorized as very unhealthy. However, the minimum monthly AQI occurred in April and May, and was categorized as unhealthy for sensitive groups. Raza *et al.*[49] and Razzaq *et al.*[50] corroborated the same results. Based on the findings of this study, it was concluded that for daily $PM_{2.5}$, the study area is facing poor air quality with 56% unhealthy, as shown in Fig. 11(a), and for monthly $PM_{2.5}$, 58.3% unhealthy, as shown in Fig. 11(b), during winter 2021.

## 4 Conclusions

The $PM_{2.5}$ concentration over Peshawar City for the year 2021 was estimated using MLR, CNN, and LSTM models to examine and validate the effectiveness of linear regression and deep learning models. Each model was repeated, and their associated results were used for comparison in terms of $R$, $R^2$, RMSE, and MSE. It was found from the investigations that deep learning techniques, such as CNN and LSTM, performed significantly better than MLR. The estimated values of the two models were close to the observed values. In contrast to CNN and LSTM models, MLR's $R^2$ value was lower, and its estimated value distributions were more widely distributed, according to goodness of fit plots. This demonstrates the shortcomings of the MLR approach on the one hand, while also demonstrating the exceptional performance of deep learning models in modeling long-term dependencies for precise $PM_{2.5}$ concentration estimation on the other. The inability of conventional models to understand the temporal correlation of air contaminants may be the cause. The LSTM model, making it a viable option for predicting the long-term future concentration of $PM_{2.5}$, may precisely determine the spatiotemporal correlation of air pollutants. In summary, the CNN and LSTM models provide a tangible way for government organizations to address the problem of air pollution. Effective methods for reducing the effects of air quality will be framed by this understanding. By combining daily and hourly data, we hope to expand our research in the future to include air pollution and provide a more thorough analysis of pollution levels in various urban regions. In the future, multiscale estimation in the spatial domain may be investigated since the suggested models can improve $PM_{2.5}$ estimation in the temporal domain. Finally, these models can be expanded to estimate other contaminants.

## Author contributions

Conceptualization, formal analysis, resources, software, validation, visualization, and writing – review & editing: Sehrish Khan, Maqbool Ahmad, Bahadar Zeb, Shahla Nazneen, Beenish Ali, Mubarak Ahmad, Khan Alam, and Allah Ditta; data curation and investigation: Sehrish Khan and Maqbool Ahmad; methodology: Khan Alam, Shahla Nazneen, and Maqbool Ahmad; project administration: Khan Alam, Shahla Nazneen, and Allah Ditta; supervision: Khan Alam; writing – original draft: Sehrish Khan and Maqbool Ahmad. All the authors read and approved the final submission.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

Data will be made available on a reasonable request to the corresponding author.

## Acknowledgements

The authors express their gratitude to the US Consulate for providing NowCast $PM_{2.5}$ data and OMI for $NO_2$, $O_3$, and $SO_2$ satellite data.

## Notes and references

1 J. Madrigano, I. Kloog, R. Goldberg, B. A. Coull, M. A. Mittleman and J. Schwartz, *Environ. Health Perspect.*, 2013, **121**(2), 192–196.

2 Z. Nur'atiah, L. W. Ean, A. N. Ahmed, M. A. Malek and M. F. Chow, *Sci. Rep.*, 2022, **12**(1), 17565.

3 S. Thakur, A. Tangri, K. Singh and S. Sharma, *Environ. Monit. Assess.*, 2025, **197**(7), 722.

4 K. Alam, N. Rahman, H. U. Khan, B. S. Haq and S. Rahman, *Aerosol Air Qual. Res.*, 2015, **15**(2), 634–647.

5 B. Zeb, J. Nasir, K. Alam, A. Ditta, M. A. Aman and M. Shafiq, *Environ. Monit. Assess.*, 2025, **197**, 1237.

6 S. K. Sahu, H. Zhang, H. Guo, J. Hu, Q. Ying and S. K. Kota, *Air Qual., Atmos. Health*, 2019, **12**(3), 327–340.

7 T. Shi, M. Liu, Y. Hu, C. Li, C. Zhang and B. Ren, *Int. J. Environ. Res. Public Health*, 2019, **16**(7), 1099.

8 M. Ahmad, K. Hussain, J. Nasir, Z. Huang, K. Alam, S. Liaquat, P. Wang, W. Hussain, L. Mihaylova, A. Ali and S. B. Farhan, *Atmosphere*, 2022, **13**(12), 1994.

9 M. Ahmad, M. Ahmad, K. Alam, B. Zeb, K. Khan and A. Ditta, *Environ. Monit. Assess.*, 2025, **197**, 1055.

10 Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang and Z. Wang, *Curr. Pollut. Rep.*, 2020, **6**, 399–409.

11 M. Ahmad, K. Alam, S. Tariq, S. Anwar, J. Nasir and M. Mansha, *Atmos. Environ.*, 2019, **219**, 117050.

12 N. T. Hung, H. W. Ting and K. H. Chi, *Aerosol Air Qual. Res.*, 2018, **18**, 2591–2599.

13 B. Bera, S. Bhattacharjee, N. Sengupta and S. Saha, *Environ. Challenges*, 2021, **4**, 100155.

14 B. Zeb, K. Alam, R. Khan, A. Ditta, R. Iqbal, M. F. Elsadek, A. Raza and M. S. Elsheikh, *Sci. Rep.*, 2024, **14**(1), 8548.

15 H. Karimian, Q. Li, C. Wu, Y. Qi, Y. Mo, G. Chen, X. Zhang and S. Sachdeva, *Aerosol Air Qual. Res.*, 2019, **19**(6), 1400–1410.

16 S. Anwar, K. Hwang and W. Sung, *ACM J. Emerg. Technol. Comput. Syst.*, 2017, **13**(3), 1–18.

17 Y. Zheng, F. Liu and H. P. Hsieh, *Proceedings of the 19th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2013, pp. 1436–1444.

18 S. Zohreh and M. Jamshid, *Sci. Rep.*, 2025, **15**(1), 21449.

19 R. Feng, H. J. Zheng, H. Gao, A. R. Zhang, C. Huang, J. X. Zhang, K. Luo and J. R. Fan, *J. Cleaner Prod.*, 2019, **231**, 1005–1015.

20 G. Ali, Y. Bao, W. Ullah, S. Ullah, Q. Guan, X. Liu, L. Li, Y. Lei, G. Li and J. Ma, *Atmosphere*, 2020, **11**(3), 306.

21 S. Ahmad, B. Zeb, A. Ditta, K. Alam, I. Ahmad and F. Usman, *Environ. Sci.: Adv.*, 2025, **4**, 1103–1116.

22 R. Esworthy, *Air Quality: EPA's 2013 Changes to the Particulate Matter (PM) Standard*, Library of Congress, Congressional Research Service, Washington, DC, USA, 2013.

23 N. C. Hsu, M. J. Jeong, C. Bettenhausen, A. M. Sayer, R. Hansell, C. S. Seftor, J. Huang and S. C. Tsay, *J. Geophys. Res.: Atmos.*, 2013, **118**(16), 9296–9315.

24 H. Bibi, K. Alam, F. Chishtie, S. Bibi, I. Shahid and T. Blaschke, *Atmos. Environ.*, 2015, **111**, 113–126.

25 R. Boiyo, K. R. Kumar and T. Zhao, *Int. J. Climatol.*, 2018, **38**, 1221–1240.

26 M. Kumar, K. S. Parmar, D. B. Kumar, A. Mhawish, D. M. Broday, R. K. Mall and T. Banerjee, *Atmos. Environ.*, 2018, **180**, 37–50.

27 C. Wu and J. Z. Yu, *Atmos. Meas. Tech.*, 2018, **11**, 1233–1250.

28 M. K. Hossen, Y. T. Peng, A. Shao and M. C. Chen, *Sci. Rep.*, 2025, **15**(1), 24830.

29 N. M. Sobri, W. F. W. Yaacob, N. A. Ismail, M. A. A. Malik, R. A. Rahman, N. A. Baser and S. A. M. Sukhairi, *J. Phys.: Conf. Ser.*, 2021, **2084**(1), 012010.

30 W. N. Ismail, H. A. Alsalamah, M. M. Hassan and E. Mohamed, *Heliyon*, 2023, **9**(2), e13636.

31 S. Li, G. Xie, J. Ren, L. Guo, Y. Yang and X. Xu, *Appl. Sci.*, 2020, **10**(6), 1953.

32 J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, *arXiv*, 2014, preprint, arXiv:1412.6806, DOI: 10.48550/arXiv.1412.6806.

33 C. Zhang, J. Yan, C. Li, X. Rui, L. Liu and R. Bie, in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 297–301.

34 G. Klambauer, T. Unterthiner, A. Mayr and S. Hochreiter, *Advances in Neural Information Processing Systems*, 2017, 30, pp. 971–980.

35 C. J. Huang and P. H. Kuo, *Sensors*, 2018, **18**(7), 2220.

36 J. Ma, Y. Ding, J. C. Cheng, F. Jiang and Z. Xu, *Water Res.*, 2020, **170**, 115350.

37 X. Bai, N. Zhang, X. Cao and W. Chen, *PeerJ*, 2024, **12**, e17811.

38 Y. Liu, L. He, W. Qin, A. Lin and Y. Yang, *Remote Sens.*, 2021, **14**, 7.

39 C. Bowen, M. Liu, S. Li, Z. Jin, Y. Zeng and X. Lin, *Atmos. Pollut. Res.*, 2023, **14**(9), 101833.

40 H. Xu, M. J. Bechle, M. Wang, A. A. Szpiro, S. Vedal, Y. Bai and J. D. Marshall, *Sci. Total Environ.*, 2019, **655**, 423–433.

41 Y. Liu, P. Koutrakis and R. Kahn, *J. Air Waste Manage. Assoc.*, 2007, **57**(11), 1351–1359.

42 X. Zhao, X. Zhang, X. Xu, J. Xu, W. Meng and W. Pu, *Atmos. Environ.*, 2009, **43**(18), 2893–2900.

43 J. Xin, L. Wang and Y. Zhang, *Environ. Sci. Technol.*, 2014, **48**(13), 7436–7444.

44 T. T. Nguyen, H. Q. Bui, H. V. Pham, H. V. Luu, C. D. Man, H. N. Pham, H. T. Le and T. T. Nguyen, *Environ. Res. Lett.*, 2015, **10**(9), 095016.

45 A. B. Chelani, *Atmos. Pollut. Res.*, 2019, **10**(3), 847–857.

46 M. Mirzaei and F. Zandkarimi, *Theor. Appl. Climatol.*, 2019, **137**(3–4), 1457–1465.

47 A. S. Moursi, N. El-Fishawy, S. Djahel and M. A. Shouman, *Complex Intell. Syst.*, 2021, **7**(6), 2923–2947.

48 WHO, World Health Organization, 2021, Licence: CC BY-NC-SA 3.0 IGO, https://apps.who.int/iris/handle/10665/345329.

49 W. Raza, S. Saeed, H. Saulat, H. Gul, M. Sarfraz, C. Sonne, Z. H. Sohn, R. J. Brown and K. H. Kim, *J. Cleaner Prod.*, 2021, **279**, 123676.

50 A. Razzaq, M. M. Zafar, L. T. Zahra, F. Qadir, F. Qiao and X. Jiang, *Environ. Challenges*, 2024, 100999.

51 C. Bowen, M. Liu, S. Li, Z. Jin, Y. Zeng and X. Lin, *Atmos. Pollut. Res.*, 2023, **14**(9), 101833.