scientific reports

Check for updates

OPEN

# Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru

Chardin Hoyos Cordova[1], Manuel Niño Lopez Portocarrero[1], Rodrigo Salas[2], Romina Torres[3], Paulo Canas Rodrigues[4] & Javier Linkolk López-Gonzales[1,5]✉

**The prediction of air pollution is of great importance in highly populated areas because it directly impacts both the management of the city's economic activity and the health of its inhabitants. This work evaluates and predicts the Spatio-temporal behavior of air quality in Metropolitan Lima, Peru, using artificial neural networks. The conventional feedforward backpropagation known as Multilayer Perceptron (MLP) and the Recurrent Artificial Neural network known as Long Short-Term Memory networks (LSTM) were implemented for the hourly prediction of $PM_{10}$ based on the past values of this pollutant and three meteorological variables obtained from five monitoring stations. The models were validated using two schemes: The Hold-Out and the Blocked-Nested Cross-Validation (BNCV). The simulation results show that periods of moderate $PM_{10}$ concentration are predicted with high precision. Whereas, for periods of high contamination, the performance of both models, the MLP and LSTM, were diminished. On the other hand, the prediction performance improved slightly when the models were trained and validated with the BNCV scheme. The simulation results showed that the models obtained a good performance for the CDM, CRB, and SMP monitoring stations, characterized by a moderate to low level of contamination. However, the results show the difficulty of predicting this contaminant in those stations that present critical contamination episodes, such as ATE and HCH. In conclusion, the LSTM recurrent artificial neural networks with BNCV adapt more precisely to critical pollution episodes and have better predictability performance for this type of environmental data.**

The World Health Organization (WHO) reported that air pollution causes 4.2 million premature deaths per year in cities and rural areas around the world[1]. The US Environmental Protection Agency[2] mentions that one of the pollutants with the most significant negative impact on public health is particulate material with a diameter of less than ten $\mu$m ($PM_{10}$) because it can easily access the respiratory tract causing severe damage to health. For their part, Valdivia and Pacsi[3] report that Metropolitan Lima (LIM) is vulnerable to high concentrations of $PM_{10}$, due to its accelerated industrial and economic growth, in addition to its large population, as it is home to 29% of the total Peruvian population[4].

To mitigate the damage caused by $PM_{10}$ to public health, the WHO established concentration thresholds suitable to achieve a minimum adverse effect on health[5]. In various countries, several laws were issued to regulate $PM_{10}$ concentrations and air quality in general[6], as established in Peru by the Ministry of the Environment[7] and in, e.g., the United States by the Environmental Protection Agency (EPA)[8].

In recent years, various forecasting methodologies have been adapted and developed to understand how pollutants behave in the air at the molecular level, simulating diffusion and dispersion patterns based on the size and type of the molecule. However, the results of the prediction tend to achieve a somehow low precision[9,10]. Examples of such models are the Community Multiscale Air Quality model and the Weather Research and Forecasting model coupled with Chemistry developed in Chen et al.[11] and Saide et al.[12], respectively, which are used to forecast air quality in urban areas. On the other hand, some methods tend to be more appropriate to model and forecast air quality because they use statistical modeling techniques, such as Artificial Neural Networks

[1]E.P. Ingeniería Ambiental, Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión, Lima, Peru. [2]Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso, Valparaíso, Chile. [3]Engineering Faculty, Universidad Andres Bello, Viña del Mar, Chile. [4]Department of Statistics, Federal University of Bahia, Salvador, Brazil. [5]Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile. ✉email: javierlinkolk@gmail.com

(ANNs). These models have been widely used to forecast time series and applied to environmental data such as particulate matter in different countries[13,14].

Several studies have been focusing on applying recurrent neural networks to forecast air quality in large cities. For instance, Guarnaccia et al.[15] reported that predicting air quality with high accuracy can be problematic. This issue is becoming increasingly important because it is a tool capable of providing complete information for helping to prevent critical pollution episodes and reduce human exposure to these contaminants[13,16,17]. However, there is a limited number of studies in the context of Lima, Peru, which is one of the cities with the highest pollution levels in South America[18–20]. For instance, Herrera and Trinidad[21] used neural networks to predict $PM_{10}$ in the Carabayllo district - Lima, with a good forecasting performance. Salas et al.[22] developed a NARX model using artificial neural networks to predict the $PM_{10}$ pollutant in Santiago, Chile. Athira et al.[23] aimed at forecasting $PM_{10}$ three days ahead and at comparing the performance of the standard LSTM, GRU, and RNN models, concluding that all three models showed good performance for out-of-sample forecasting.

Lima is considered to be one of the most polluted cities in Latin America in terms of $PM_{10}$. In this sense, the need for sophisticated environmental management instruments arises, aiming at making predictions with greater precision using cutting-edge methodologies, such as deep learning algorithms, which support decision-making to establish mitigation and prevention policies. In addition, it allows the population to avoid being exposed to high concentrations of $PM_{10}$. For this reason, this study aims to assess the air quality of Lima, to understand its behavior, and the possible causes and factors that favor pollution. Subsequently, we applied the Multilayer Perceptron (MLP) and the Long Short-Term Memory (LSTM) models to forecast $PM_{10}$ concentrations, where the models were evaluated under two validation schemes: the Hold-out (HO) and the Block Nested Cross-Validation (BNCV). Our contributions are summarized below:

- In this study, we have implemented artificial neural networks to model time series data collected from five meteorological and air quality monitoring stations from Lima, Peru. The monitoring stations are ATE, Campo de Marte (CDM), Carabayllo (CRB), Huachipa (HCH) and San Martin de Porres (SMP). We have investigated the geographical and meteorological divergence of the forecast results from the five air quality monitoring areas in LIM using data collected from two years.
- The proposed time series forecasting model based on the MLP and LSTM neural networks efficiently predicted one-hour-ahead $PM_{10}$ concentrations. The prediction performances between the five stations were compared. According to the literature review, this study is the first to use deep learning algorithms to predict air quality ($PM_{10}$) in LIM.
- We have focused the study in LIM because its air pollution has worsened in recent years. The main reason for this change is that population growth has been unsustainable, and high industrial activity and the accelerated growth of the automobile fleet have increased. These factors make it challenging to predict $PM_{10}$ pollution concentrations.

The remainder of the paper is structured as follows: Section "Materials and methods" presents the developed methodology based on an exploratory study described in two phases. In Section 3, we present the main results and their discussion. Finally, in Section 4, we provide the main conclusions and give some future works.

## Materials and methods

In this work, we follow the Knowledge Discovery from Databases (KDD) methodology to obtain relevant information for air quality management decision-making. The main goal of the KDD is to extract implicit, previously unknown, and potentially helpful information[24] from raw data stored in databases. Therefore, the resulting models can predict, e.g., one-hour ahead, the air quality and support the city's management decision-making (see Fig. 1).

The KDD methodology has the following stages: (a) Phenomena Understanding; (b) Data Understanding; (c) Data Preparation; (d) Modeling; (e) Evaluation; and, (d) Selection/Interpretation. In the following subsections, we explain each stage of the process.

**Phenomena Understanding.** In this first stage, we contextualize the contamination phenomenon concerning the $PM_{10}$ concentrations in the five Lima monitoring stations. The main focus is to predict air pollution to support decision-making related to establishing pollution mitigation policies. For this, we use both MLP and LSTM as computational statistical methods for $PM_{10}$ prediction.

Lima is the capital of the Republic of Peru. It is located in the center of the western side of the South American continent in the 77° W and 12° S and, together with its neighbor, the constitutional province of Callao, form a populated and extensive metropolis with 10,628,470 inhabitants and an area of 2819.3 $km^2$[25,26].

The average relative humidity (temperature) in the summer (December–March) ranges from 65–68% (24 °C–26 °C) in the mornings, while at night the values fluctuate between 87–90% (18 °C–20 °C). In the winter (June–September), the average daytime relative humidity (temperature) ranges between 85–87% (18 °C–19 °C) and at night it fluctuates between 90–92% (18 °C–19 °C). The average annual precipitation is 10 mm. On the other hand, the average altitudes reached by the thermal inversion in summer and winter are approximately 500 and 1500 m above sea level, respectively[27,28].

**Data understanding.** Lima has ten air quality monitoring stations located in the constitutional province of Callao and the north, south, east, and center of Lima. The data used comprise hourly observations from January 1st, 2017, to December 31st, 2018, and includes three meteorological variables and the concentration of

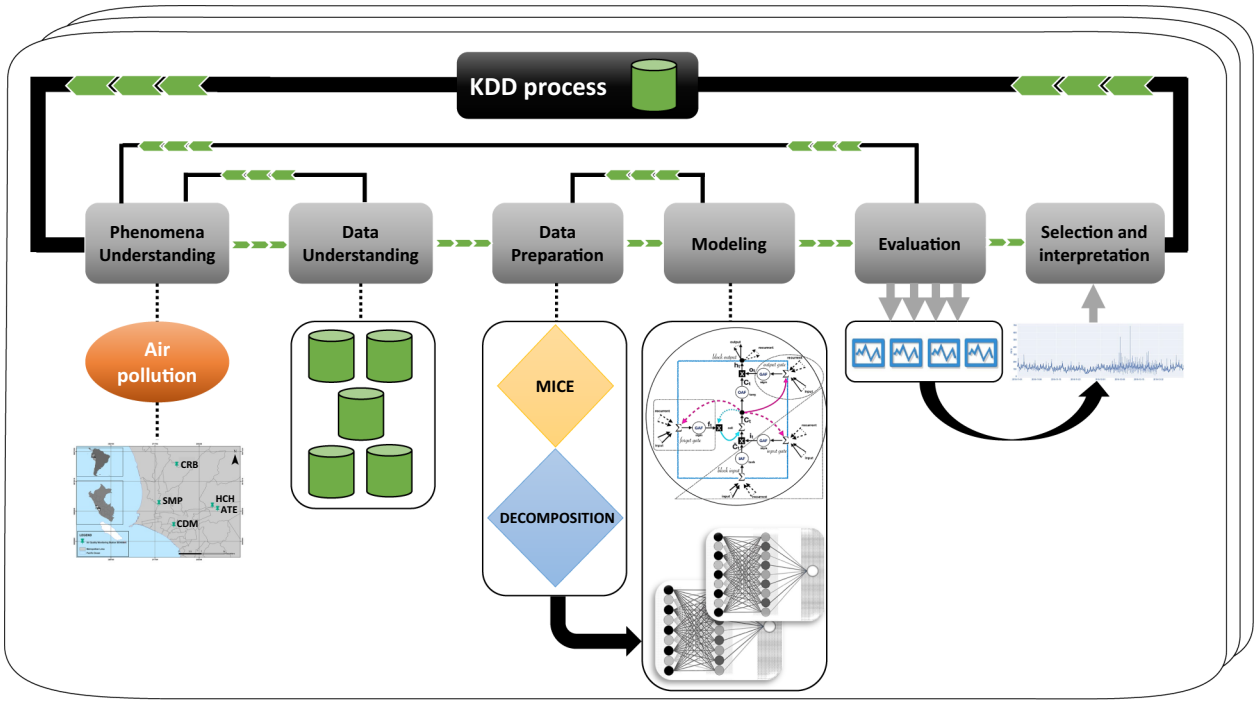



**Figure 1.** Knowledge Discovery from Databases (KDD) methodology used for Air Quality Assessment and Pollution Forecasting.

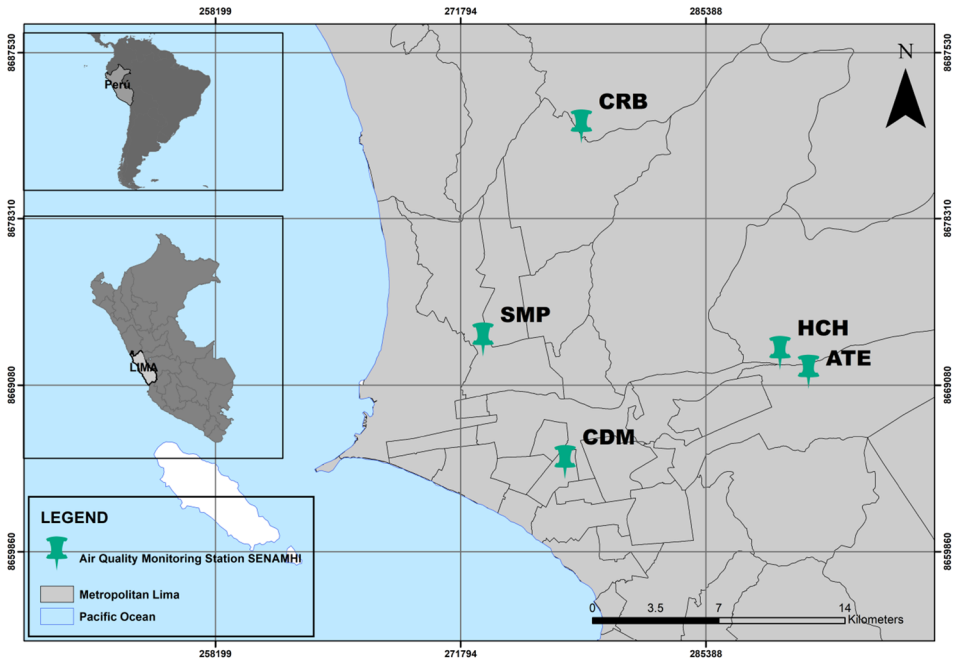



**Figure 2.** Map with the study area and the locations of the Lima air quality monitoring stations: ATE, Campo de Marte (CDM), Carabayllo (CRB), Huachipa (HCH) and San Martin de Porres (SMP).

particulate matter $PM_{10}$. Where the latter is considered to be an agent that, when released into the environment, causes damage to ecosystems and living beings[29,30]. For this study, the hourly data, recorded at five air quality monitoring stations (see Fig. 2), which are managed by the National Service of Meteorology and Hydrology of Peru (SENAMHI), was considered. Table 1 shows the considered variables and their units of measurement.

When considering environmental data, such as $PM_{10}$ concentrations, from different locations, preliminary spatio-temporal visualization studies are of great use to better understand the behavior of the meteorological variables, the topography of the area, and the pollutants[31].

| Variable | Unit of measurement |
|---|---|
| $PM_{10}$ | $\mu g/m^3$ |
| Temperature | °C |
| Relative humidity | % |
| Wind speed | m/s |
| Wind direction | Degrees (°) |

**Table 1.** Pollutant and weather variables used in this study, and their units of measurement.

**Data preparation.** This stage is very relevant because it precedes the modeling stage. The preparation of the data had various stages. First, we address the problem of missing data. The treatment was performed with the MICE library. This library performs multiple imputations using the Fully Conditional Specification[32] and requires a specification of a separate univariate imputation method for each incomplete variable. In this context, predictive mean matching, a versatile semiparametric method focusing on continuous data, was used, which allows the imputed values to match one of the observed values for each variable. The data imputation was performed for each of the five stations with a percentage of missing data below 25%.

The data from the monitoring stations consist of a sequence of observed values $\{x_t\}$ recorded at specific times $t$. In this case, the time series is collected at hourly intervals. After the data imputation, we proceed to normalize all the observations in the range [0,1] as follows:

$$X_t = \frac{x_t - \min\{x_t\}}{\max\{x_t\} - \min\{x_t\}} \tag{1}$$

Moreover, the time series is decomposed into the trend, seasonality, and the irregular components following an additive model (the cyclic component is omitted in this work):

$$X_t = Trend_t + Cyclic_t + Seasonal_t + Irregular_t \tag{2}$$

The trend component $Trend_t$ at time $t$ reflects the long-term progression of the series that could be linear or non-linear. The seasonal component $Seasonal_t$ at time $t$, reflects the seasonal variation. The irregular component $Irregular_t$ (or "noise") at time $t$ describes the random and irregular influences. In some cases, the time series has a cyclic component $Cyclic_t$ that reflects the repeated but non-periodic fluctuations. The main idea of applying this decomposition is to obtain the deterministic and the random components, where a forecasting model is obtained using the deterministic part[33,34]. In this article, we have used the method implemented in Statmodels for Python[35], where a centered moving average filter is applied to the time series.

**Modeling using artificial neural networks.** Artificial Neural Networks have received a great deal of attention in engineering and science. Inspired by the study of brain architecture, ANNs represent a class of non-linear models capable of learning from data[36]. The essential features of an ANN are the basic processing elements referred to as neurons or nodes, the network architecture describing the connections between nodes, and the training algorithm used to estimate values of the network parameters.

Researchers see ANNs as either highly parameterized models, or semiparametric structures[36]. ANNs can be considered as hypotheses of the parametric form $h(\cdot; \mathbf{w})$, where the hypothesis $h$ is indexed by the vector of parameters $\mathbf{w}$. The learning process consists of estimating the value of the vector of parameters $\mathbf{w}$ to adapt the learner $h$ to perform a particular task.

Machine Learning and Deep learning methods have been successfully applied for time series forecasting[37–42]. For instance, recurrent artificial neural networks (RNNs) are dynamic models frequently used for processing sequences of real data step by step, predicting what comes next. They are applied in many domains, such as the prediction of pollutants[43]. It is known that when there are long-term dependencies in the data, RNNs are challenging to train, which leads to the development of models such as the LSTM that have been successfully applied in time series forecasting[44].

The Multilayer Perceptron model consists of a set of elementary processing elements called neurons[36,45–48]. These units are organized in architecture with three layers: input, hidden, and output. The neurons corresponding to one layer are linked to the neurons of the subsequent layer. Figure 3 illustrates the architecture of this artificial neural network with one hidden layer. The non-linear function $\mathbf{g}(\mathbf{x}, \mathbf{w})$ represents the output of the model, where $\mathbf{x}$ is the input signal and $\mathbf{w}$ being its parameter vector. For a three-layer FANN (one hidden layer), the $k$-th output computation is given by the following equation

$$g_k(\mathbf{x}, \mathbf{w}) = f_2\left(\sum_{j=1}^{\lambda} w_{kj}^{[2]} f_1\left(\sum_{i=1}^{d} w_{ji}^{[1]} x_i + w_{j0}^{[1]}\right) + w_{k0}^{[2]}\right) \tag{3}$$

where $\lambda$ is the number of hidden neurons. An important factor in the specification of neural models is the activation function's choice. These can be any non-linear functions as long as they are continuous, bounded, and differentiable. The transfer function of the hidden neurons $f_1(\cdot)$ should be nonlinear while for the output neurons the function $f_2(\cdot)$ could be a linear function or nonlinear functions. One of the most used functions is the sigmoid:

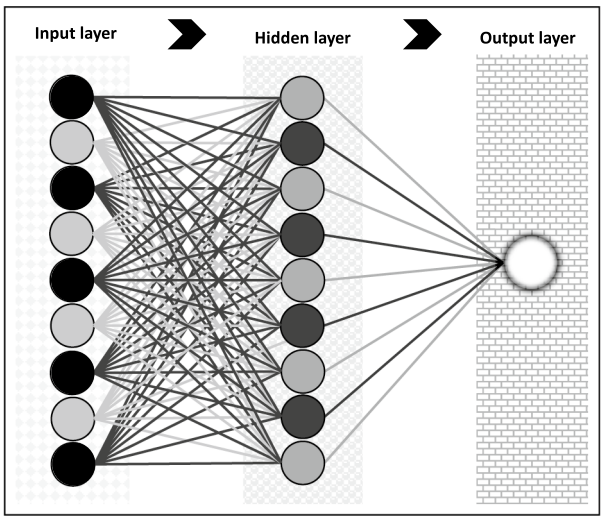



**Figure 3.** Schematic of the architecture of the Multilayer Perceptron. The figure shows three layers of neurons: input, hidden and output layers.

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{4}$$

The MLP operates as follows. The input layer neurons receive the input signal; these neurons propagate the signal to the first hidden layer and do not make any processing. The first hidden layer processes the signal and transfers it to the subsequent layer; the second hidden layer propagates the signal to the third, and so on. When the signal is received and processed by the output layer, it generates the response.

The Long Short-Term Memory networks model is a type of RNN, having as its primary strength the ability to learn long-term dependencies and being a solution for long time series intervals[20,49]. In such a model, memory blocks replace the neurons in the hidden layer of the standard RNN[50]. The memory block consists of three gates that control the system's state: Input, forget, and output gates. First, the input gate determines how much information will be added to the cell. Second, the forget gate controls the information lost in the cells. Lastly, the output gate performs the function of determining the final output value based on the input and memory of the cell[51,52].

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{5}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{6}$$

$$\tilde{C}_t = \tanh\left(W_{\tilde{C}} \cdot [h_{t-1}, x_t] + b_{\tilde{C}}\right) \tag{7}$$

$$C_t = \left(f_t \cdot C_{t-1}\right) + \left(i_t \cdot \tilde{C}_t\right) \tag{8}$$

$$o_t = \sigma\left(W_o [h_{t-1}, x_t] + b_o\right) \tag{9}$$

$$h_t = o_t \cdot tanh(C_t) \tag{10}$$

Figure 4 shows the LSTM model block, with the output and input blocks, which consists of three gates. At each step, an LSTM maintains a hidden vector $h$ and a memory vector $o$ responsible for controlling status updates and outputs.

The first step is to decide what information will not be considered in the status cell. This decision is made by the forget gate, which uses a hyperbolic tangent activation function (IAF). $f_t$ represents the output of the forget gate, which can be calculated using equation (5). This gate considers the concatenation of the vectors $h_{t-1}$ and $x_t$. It generates a number between 0 and 1 for each number in the state cell $C_{t-1}$, where $W_f$ and $b_f$ are the weight matrices and the bias vector parameters, respectively. Both must be learned during training and are stored in the vector $f_t$. If one of the values of this vector is equal to or close to zero, then the LSTM will eliminate that information. On the other hand, if it reaches values equal to or close to 1, this information will be maintained and reach the status cell.

The next step is to decide what new information to store in the status cell. This is done by the input gate, linked to a sigmoid activation function (GAF), and with an output for that gate ($i_t$), all this is calculated by the equation (6, 7). In addition, for the input block, the hyperbolic tangent activation function (IAF) is used. First, the vectors $h_{t-1}$ and $x_t$ are concatenated. Being $W_i$ and $b_i$, the weight matrices and the bias vector parameters, respectively,

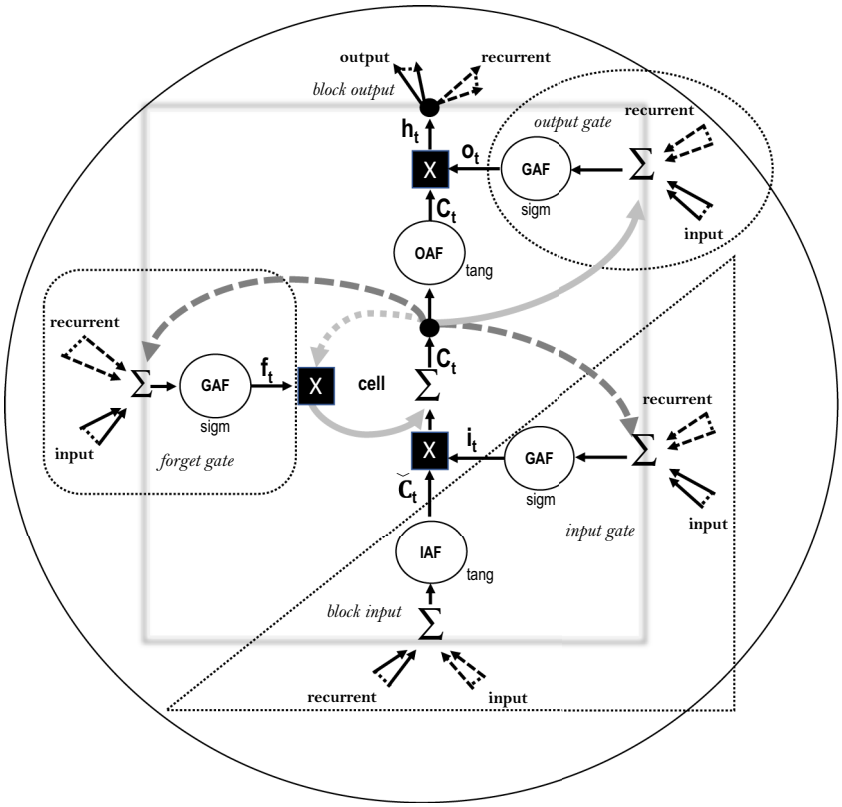



**Figure 4.** Model of one block of the LSTM. The block is composed of the input gate, forget gate and output gate.

must be learned during training; all this is stored in the vector $i_t$ called the input gate, which decides which values to update. Then a hyperbolic tangent function creates a vector of new candidate values, $\tilde{C}_t$, involving the vectors $h_{t-1}$ and $x_t$. In the next step, these values are filtered by multiplying point by point both vectors to create a status cell update. The previous cell, $C_{t-1}$ is updated to the new state of cell $C_t$ (equation 8).

In addition, the output gate, also linked with the GAF activation function and with an output of the output gate ($o_t$), for its calculation uses the equation (equation 9). Finally, $h_t$, expresses the new output of the model (equation 10). The current cell state is represented by $C_t$, while $W$ is the weight vector o parameters of the model, and $b$ is the bias of the model.

**Model evaluation.** To evaluate the forecast ability of the models, the performance metrics given below were used (see[53,54]). In what follows, we will consider: $y_i$, $i = 1, \ldots, n$, are the target values; $\hat{y}_i$, $i = 1, \ldots, n$, are the model's predictions; $\bar{y}_i$ is the mean of the target values; and $n$ is the number of samples.

1. Mean Absolute Error: The average absolute difference between the target and the predicted values.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{11}$$

2. Root Mean Squared Error: The squared root of the average of the squared errors.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{12}$$

3. Symmetric Mean Absolute Percentage Error: A measure of accuracy based on a percentage of relative errors.

$$\mathrm{sMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|\hat{y}_i| + |y_i|} \tag{13}$$

4. Spearman's rank correlation coefficient: A nonparametric correlation measure between the target and the prediction. Spearman's correlation assesses monotonic relationships by using the rank of the variables.

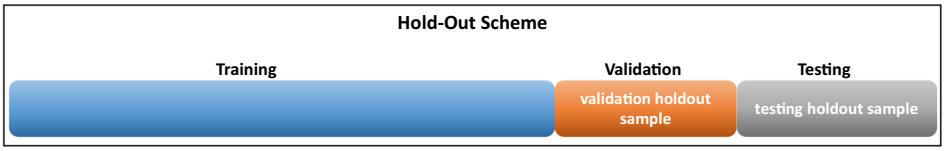



**Figure 5.** Hold-Out Scheme used for the validation of the models. The dataset is split into three sets: training, validation, and testing. The train set is the basis for training the model, and the test set is used to see how well the model performs in untrained $PM_{10}$ concentrations.

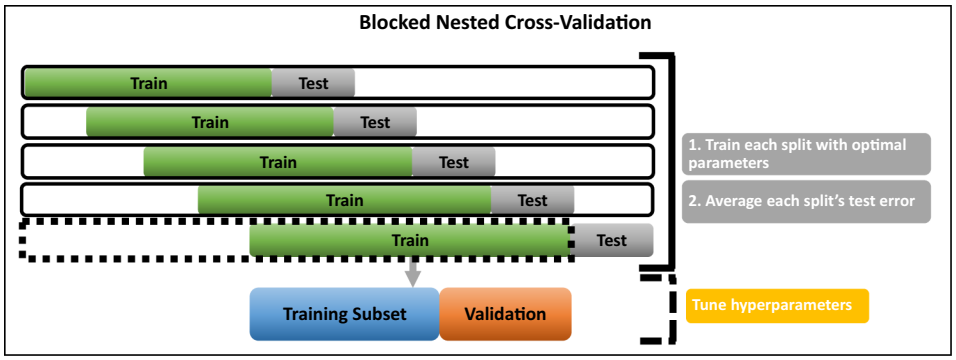



**Figure 6.** Blocked Nested Cross-Validation Scheme used for the validation of the models. The dataset is separated into three sets using a time-window of fixed size: training, validation, and testing. The last day is used for testing.

| SM | Minimun | Maximun | 1st Qu. | 3rd Qu. | Median | Mean ± DS | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| CRB | 5.44 | 488.02 | 31.49 | 58.45 | 198.31 | 48.69 ± 28.39 | 806.03 | 3.24 | 22.27 |
| SMP | 7.77 | 426.80 | 61.95 | 105.10 | 142.50 | 86.05 ± 35.73 | 1276.41 | 1.00 | 2.86 |
| CDM | 6.08 | 463.60 | 35.84 | 63.45 | 145.50 | 52.30 ± 24.61 | 605.54 | 2.30 | 18.25 |
| ATE | 6.41 | 931.00 | 82.90 | 148.00 | 421.90 | 121.56 ± 60.30 | 3635.75 | 2.08 | 11.07 |
| HCH | 5.21 | 974.00 | 62.10 | 176.50 | 138.40 | 130.03 ± 91.68 | 8404.34 | 1.53 | 4.89 |

**Table 2.** Descriptive statistics for the five $PM_{10}$ monitoring stations.

$$S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{14}$$

where $d_i = rg(y_i) - rg(\hat{y}_i)$ is the difference between the ranks of the targets $rg(y_i)$ and the predictions $rg(\hat{y}_i)$.

**Model selection and interpretation.** The model selection and interpretation is the final step in the KDD process and requires that the knowledge extracted from the previous step be applied to the specific domain of the $PM_{10}$ prediction in a visualized format. At this stage, in addition to selecting the model with the best precision in the prediction, it also drives the decision-making process based on the air quality assessment in Lima.

We have used two schemes for the validation: Hold-Out (HO) and Blocked Nested Cross-Validation (BNCV). On the one hand, HO has the conventional separation of the dataset in training, validation, and testing subsets (see Fig. 5). On the other hand, the BNCV is a fixed-size window that slides, and the model is retrained with all the data up to the current day (see Fig. 6).

## Results and discussion

**Air quality assessment in Metropolitan Lima-Peru.** In this section, we report the results of the statistical analysis of air pollution in LIM.

*Statistical analysis of the concentration of* $PM_{10}$. Table 2 shows the descriptive analysis of the data from the five monitoring stations focused in the $PM_{10}$, between 01-01-2017 and 31-12- 2018. Additionally, the histogram (see Fig. 7) is reported to show the behavior of the pollutant in every season. In the probability distribution, it is observed that they are skewed to the right, which indicates the existence of critical episodes of contamination, being the HCH station the one with the highest incidence, with an average of $130.03 \pm 91.68\ \mu g/m^3$. This

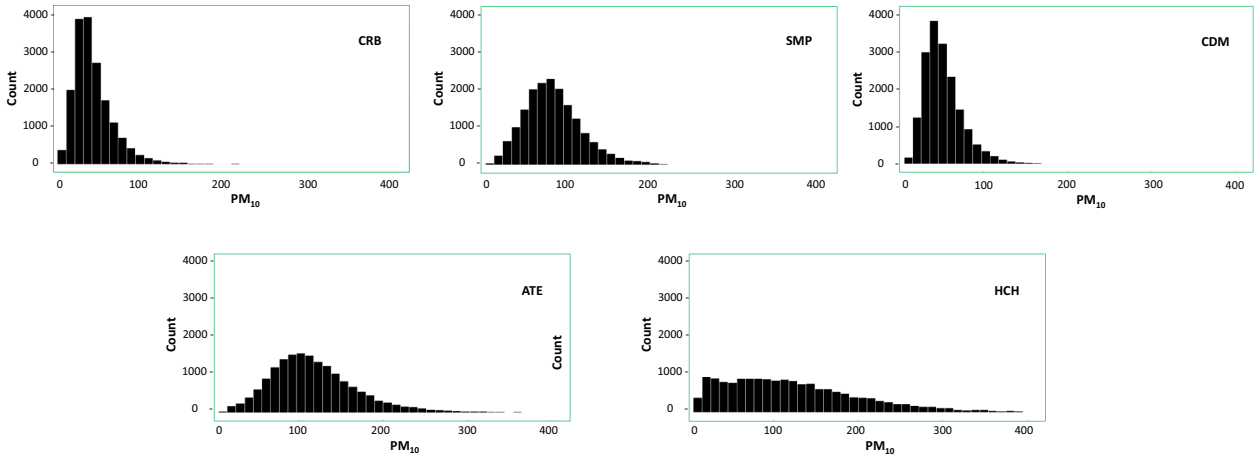



**Figure 7.** $PM_{10}$ Histograms for each of the five monitoring stations, respectively CRB, SMP, CDM, ATE, and HCH.

value exceeds that standardized by the Peruvian norm[7], and shows relevant fluctuations and high dispersion of pollutants (8404.34 $\mu g/m^3$) that cause a high standard deviation. The stations HCH and ATE register higher concentration levels. The order of the stations from the lowest to the highest levels of the mean of $PM_{10}$ is as follows: **CRB**; **CDM**; **SMP**; **ATE**; **HCH**. Similar behaviour was found in other studies[31,55]. Encalada et al.[31] carried out a study of visualization of $PM_{10}$ concentrations in Lima using the same data, where similar behavior patterns of $PM_{10}$ concentrations are shown in the five stations. In addition, all the stations surpass the $PM_{10}$ limits established by the WHO. Moreover, four of the five stations (except CRB) exceed the utmost limits of the annual arithmetic mean of $PM_{10}$ proposed in the Quality Standards Environmental (ECA) in Peru.

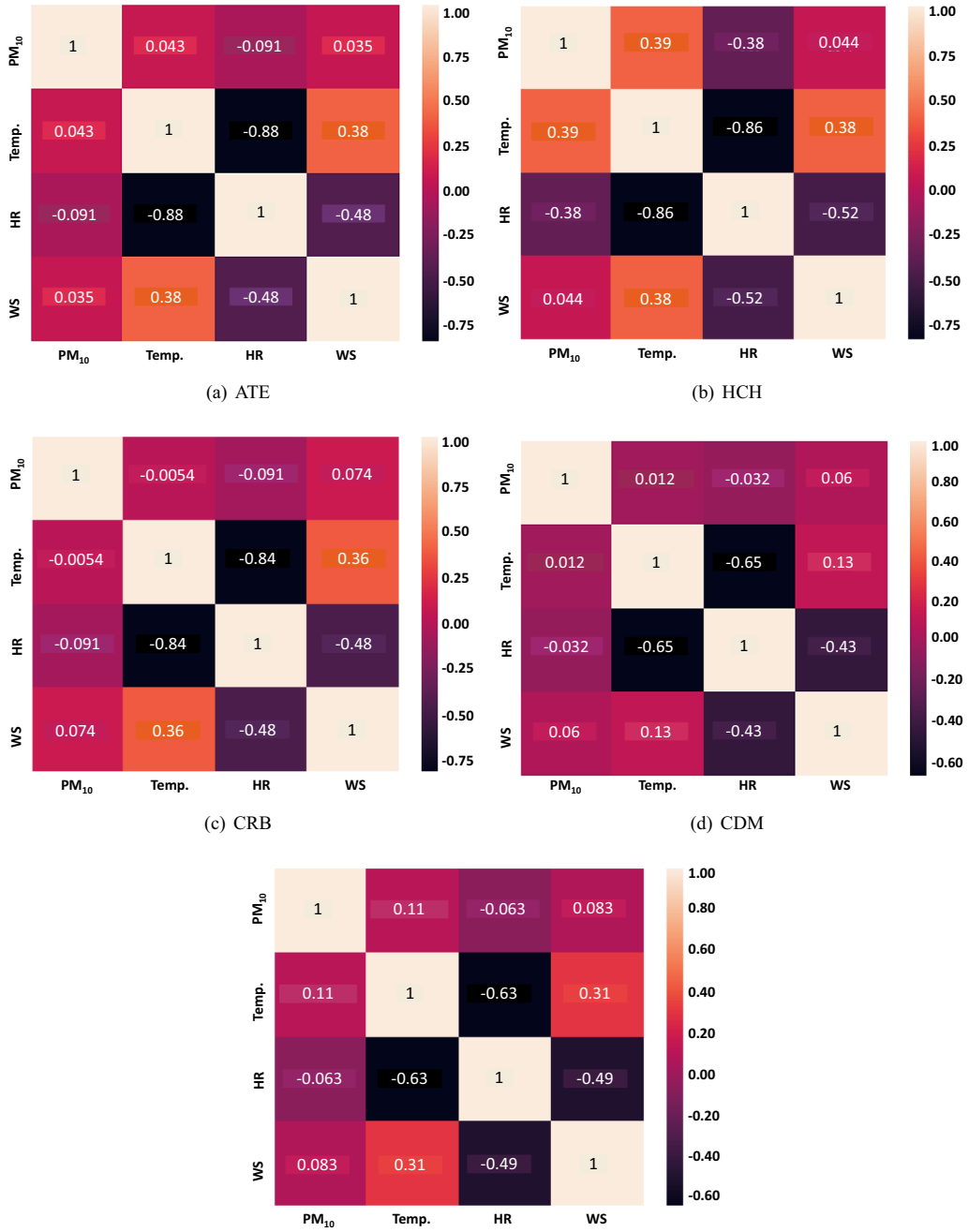
*Analysis of the correlations with the meteorological variables.* A significant correlation between $PM_{10}$ and the meteorological variables was observed in the station HCH, which is the area with the highest $PM_{10}$ concentration. Factors such as dust, population / area ratio and weather conditions have a predominant effect on $PM_{10}$ concentration[56]. Figure 8 shows that there is a moderate positive correlation (0.39) between temperature and $PM_{10}$ and a moderate negative correlation (-0.38) between relative humidity and $PM_{10}$. This is due to the meteorological patterns that occur in the study area. According to Silva et al.[57] between the years 1992 and 2014, the base of thermal inversions in Lima ranged between 0.6 and 0.9 kilometers from June to November and between 0.1 and 0.6 kilometers from December to May, having a minimum average of 0.13 kilometers in March, which coincides with the season that presents critical episodes of $PM_{10}$ concentrations.

The thermal inversion in the summer months reduces the dispersion of atmospheric pollutants because the density of the stratiform clouds decreases. Consequently, solar radiation leads to an increase in temperature and to a reduction in relative humidity. The latter results in a turbulent process causing the resuspension of coarse particles as $PM_{10}$[25]. High temperatures increase the photochemical activity that causes the decomposition of matter and, consequently, the increase of $PM_{10}$[58–60]. On the other hand, stratiform cloudiness increases in winter, as does relative humidity, that accompanied by drizzles in that season, help to significantly decrease the temperature and $PM_{10}$ concentrations due to wet deposition typical of the season[28]. The above explains the high negative correlation observed between temperature and relative humidity in the five monitoring stations (see Fig. 8), which is a normal phenomenon because the relative humidity directly depends on temperature and pressure to determine the capacity of the air in the intake of water vapor[61]. For this reason, the higher the temperature, the lower the relative humidity, as shown in Fig. 9.

*Influence of wind direction and speed on* $PM_{10}$ *concentrations.* The stations located in the highest area (eastern part) of the city have the highest concentration of $PM_{10}$. Contrary to the above, the stations located in the lowest area have a lower concentration of $PM_{10}$. This trend is due to the entry direction of persistent local winds from the coast to the south-southwest, which causes that pollutants such as $PM_{10}$ be transferred to the northeast and east areas of the city, making them in critical places of contamination by particulate matter[28,31].

Although there is no significant correlation between wind speed and $PM_{10}$, this parameter has meteorological influence on the dispersion, resuspension, and horizontal transport of pollutants, provided that there are strong air currents (winds)[61–63], which is not the case of the present study because the highest frequencies of wind speeds are between 0 – 3.10 m/s[31].

The wind speed has a meteorological influence on the dispersion, suspension, and horizontal transport of pollutants provided that there are strong air currents (winds)[61–63]. However, this is not the case of the present study because the highest frequencies of wind speeds are between 0 and 3.10 m/s[31], meaning that there is no significant correlation between wind speed and $PM_{10}$.




(a) ATE (b) HCH

(c) CRB (d) CDM

(e) SMP

**Figure 8.** Correlation matrices between the meteorological variables and the $PM_{10}$ for each monitoring station.

*Critical episodes of* $PM_{10}$ *contamination at the HCH station.* The station with the highest average $PM_{10}$ concentration between 2017 and 2018 is HCH (see Table 2). This area has the characteristic of high vehicular traffic compared to the rest of the stations considered. The Ramiro Prialé highway that crosses HCH and is the most used to access the central road connects the center and the east of the Peruvian territory, turning it into high traffic congestion. Moreover, 2,462,321 vehicles were circulating in Lima[64] in 2017, and according to the National Institute of Statistics and Informatics (INEI), the vehicle fleet in Peru grew by 4.4% between 2017 and 2018[65]. The aforementioned explains the influence of high traffic vehicles in critical pollution episodes in HCH, which according to what is referred by Srishti et al.[66], the traffic caused from vehicles contributes to about 21% of $PM_{10}$ of the pollution. In addition, it is associated with the wear of tires and brakes[64].

Another particular feature of HCH compared to the other stations is the dilapidated, unpaved roads and the frequent inadequate disposal of land clearing on public roads by the population. These conditions generate a significant increase in dust, the main component of particulate matter, contributing to 54% of air pollution.

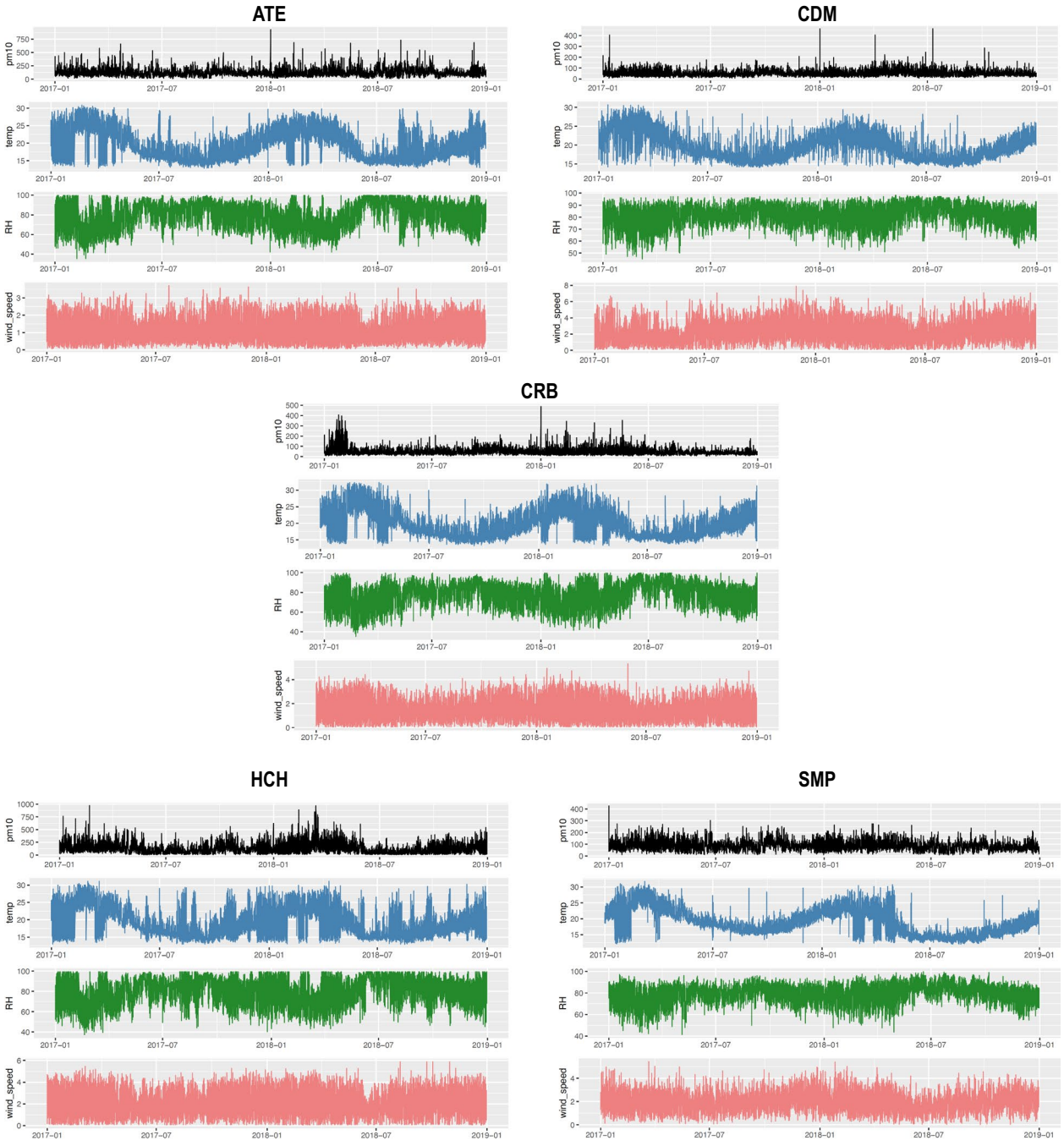



**Figure 9.** Time series of all variables, $PM_{10}$, temperature, relative humidity and wind speed, in each monitoring station, ATE, CDM, CRB, HCH and ATE, respectively.

The soil dust has a more significant impact in seasons or areas with little rainfall[66–68]. Furthermore, Lima is considered a city where it seldom rains and that only slight drizzles or wet haze breakouts from cloud-type clouds nimbostratus[69].

In the surrounding area of HCH, there is also high industrial activity. Industrialization is directly associated with the increased generation of $PM_{10}$[69]. Concepción and Rodríguez[70] note that both the industrial activity and the vehicle fleet are the leading causes of the generation of high concentrations of $PM_{10}$ in Lima, where the primary industries are brick kilns and non-metallic ore extraction. Moreover, it was evidenced that the HCH brick industries do not have the appropriate technology to mitigate air pollution and that in all their processes, high emission of particulate matter, from the movement of land to the burning of tires, plastics, or firewood in the ovens[71]. Added to all this, it is the lack of green areas in HCH, which facilitates the resuspension of $PM_{10}$.

*Exploratory analysis on a daily and monthly scale.* The predominant time scale in the concentration of $PM_{10}$ was evaluated in two episodes (see Fig. 10). That between 07:00 and 11:00 in the morning, followed by the one

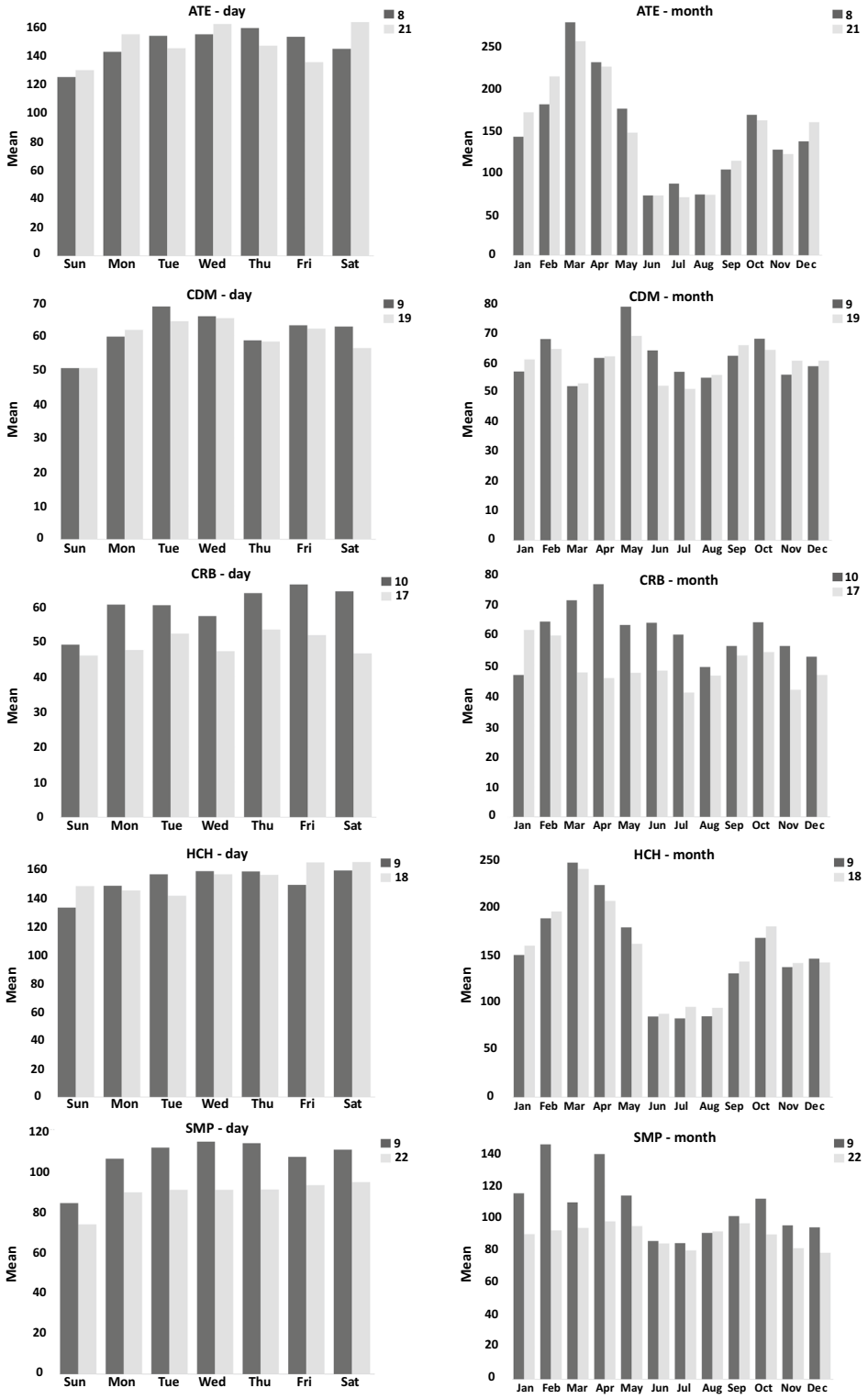



**Figure 10.** Bar plot per day and month for each monitoring station, ATE, CDM, CRB, HCH, and ATE, respectively. The average hourly pollution per day of the week and month of the year is reported for all monitoring stations.

between 17:00 and 22:00 at night. Similar results were found by Sánchez et al.[27], where the air quality of Lima was evaluated in 2015. From the above, it can be inferred that the levels of environmental pollution referring to $PM_{10}$, find the highest peaks in the evening (153.9991 and 151.9256 $\mu g/m^3$), while the lowest peaks are between

03:00 and 04:00 a.m. each day, which coincides with the results reported for the station HCH. As mentioned by Valdivia et al.[3], this is related to the reduction in emissions from mobile sources that are own of the dawn.

The behavior of concentration levels of contamination varies depending on the month. In each monitoring station, we can see two main peaks (see Fig. 10). The first corresponds to February, March, and April, which report the highest contamination in the first semester of the year. In this period, it is the beginning of classes for schoolchildren that intensifies vehicle activity. The end of the summer and the beginning of the autumn are the period associated with the time at which the thermal inversion occurs, which favors the generation of high peaks of $PM_{10}$ contamination[57]. The second peak involves the winter season and the beginning of spring, highlighting mainly October as part of the second semester of the year. Similar results were found by Encalada et al.[31].

In these time windows, the stations with the highest critical episodes were HCH and ATE, while CRB had the lowest $PM_{10}$ concentrations. In addition, from the emissions of high traffic vehicular and fixed sources of pollution, the meteorological and topographic conditions of the study area cause the high emission of $PM_{10}$ in the air, exceeding the proposed standards in all cases by WHO.

**Air pollution forecasting results.** In this study, we focus on the one-hour ahead prediction of the $PM_{10}$ concentration based on both the past values of the pollutant concentration and the current weather variables. For this, the MLP and LSTM were used with a particular architecture. Based on the autocorrelation function (ACF) and the partial autocorrelation function (PACF), relevant lags were detected that are used in the model. The configuration of the network is associated with the information provided by the ACF and PACF, where the lags $t-1, t-2, t-3, t-23$, and $t-24$ of the $PM_{10}$ time series are defined as relevant. In addition, temperature, relative humidity, and wind speed are used with $t-4$ (4 hours ago). In summary, the non-linear autoregressive model with exogenous variables identified has the following structure:

$$X_t = g_{ANN}(X_{t-1}, X_{t-2}, X_{t-3}, X_{t-23}, X_{t-24}, Temperature_{t-4}, Humidity_{t-4}, Wind_{t-4}) + \varepsilon_t \tag{15}$$

where $\{X_t, t \in \mathbb{N}\}$ is the $PM_{10}$ time series. The weather exogenous variables are $\{Temperature_t, t \in \mathbb{N}\}$, $\{Humidity_t, t \in \mathbb{N}\}$ and $\{Wind_t, t \in \mathbb{N}\}$ for temperature, humidity and wind speed respectively. Moreover, $\varepsilon_t$ is the random noise. The non-linear function $g_{ANN}(\cdot)$ stands for either the MLP or the LSTM neural networks.

The purpose of incorporating exogenous variables in this study is to improve the precision of the forecast. The exogenous variables are crucial to improve the efficiency of predictions by identifying the important meteorological covariates that affect $PM_{10}$, such as temperature, relative humidity, and wind speed[72].

In this work, we have implemented a three-layer MLP with 8 input nodes, 16 hidden nodes, and 1 output node. The activation function for the hidden and output nodes is the sigmoid function $f(z) = (1 + e^{-z})^{-1}$. On the other hand, the LSTM was implemented with 16 parallel blocks, and the output of each block is aggregated with a single neuron with a sigmoid activation function. To train both ANN models, we have selected the mean absolute error for the loss function as a robust function due to outliers. The *nadam* optimizer was used for the backpropagation algorithm. A 25% dropout strategy with a 10% of validation data was applied to avoid overfitting. A maximum of 500 epochs and batch sizes of 1024 was used to fit the models' weights.

Two alternatives were considered to obtain out-of-sample forecasts (see Fig. 11). On the one hand, the ANN models were adjusted with the training set only once for the Hold-Out scheme, and the resulting model was used to forecast one-hour ahead for the last 60 days of data. On the other hand, the ANN modes were trained several times with a fixed sliding window for the Blocked Nested Cross-Validation, where the model was updated for each subsequent day belonging to the test set, and the following days (24 samples) were used for the test set.

Table 3 shows the performance results obtained by the MLP and LSTM models evaluated in the test set using the Hold-Out and the Blocked Nested Cross-Validation Schemes. Figure 11 shows the graphs obtained by the predictions of the LSTM neural network for the five monitoring stations. Artificial neural networks show good prediction performance according to the Spearman score (over 0.60) for all the stations, except for ATE that reaches a score near 0.52. ATE and HCH monitoring stations are located in industrial areas with heavy traffic stations. The ATE and HCH monitoring stations have the highest levels of contamination and a more significant presence of outliers, which is reflected in the error metrics with values greater than twice that of the other stations. Notice that RMSE shows a higher value due to the presence of extreme values in the $PM_{10}$ levels, being MAE less affected by this type of value. On the other hand, the models evaluated by applying the BNCV scheme show slightly better performance than their HO counterparts. However, the BNCV scheme keeps the models updated with the latest records through an incremental training process with the new data.

The models' performances were strongly affected by a period of excessive contamination with critical episodes that appeared between December 3rd, 2018, and December 21st, 2018 (just before the Christmas festivities).

The time series of the pollutant was decomposed into trend, seasonality and irregular components using the decomposition method described in equation 2. The irregular component was subtracted from the original time series, and filtered time series is obtained:

$$\tilde{X}_t = Trend_t + Seasonal_t \tag{16}$$

Table 4 shows the performance results obtained by the MLP and LSTM models evaluated in the test set using the Hold-Out and the Blocked Nested Cross-Validation Schemes applied to the filtered time series. Under this situation, both the MLP and the LSTM performed very well in predicting the regular component of the $PM_{10}$ contamination levels at all monitoring stations. A remarkable point is an outstanding performance obtained by the artificial neural network models, which shows that the irregular component is hard to predict. Figure 12 shows the graphs obtained by the predictions of the LSTM neural network for the five monitoring stations.

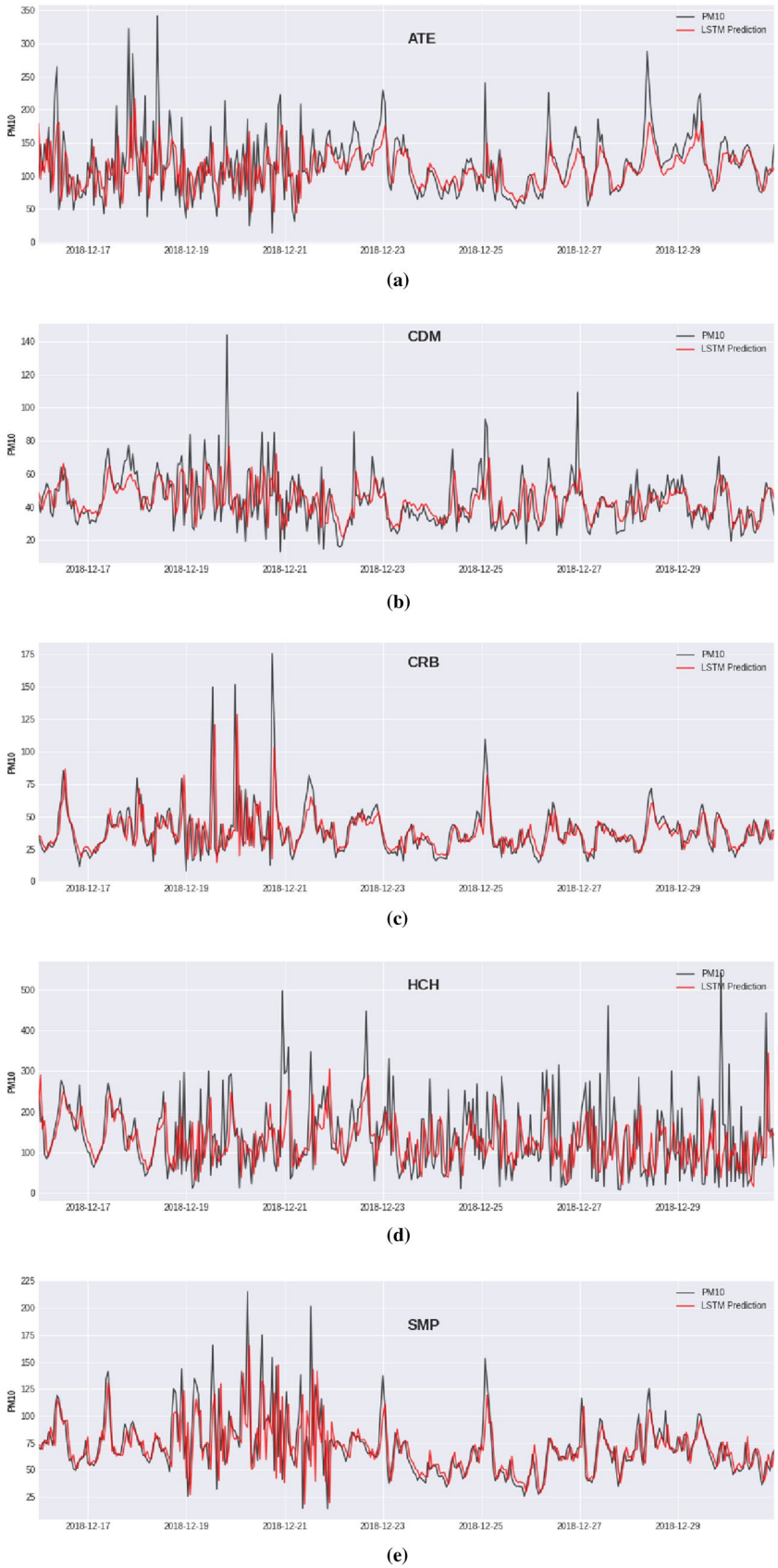



**Figure 11.** Plots for one-hour ahead predictions for the last 15 days of the $PM_{10}$ concentration level using LSTM with the BNCV scheme. Predictions for the following monitoring stations: (**a**) ATE, (**b**) CDM, (**c**) CRB, (**d**) HCH, (**e**) SMP.

| Metrics | ATE | | CDM | | CRB | | HCH | | SMP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MLP | LSTM | MLP | LSTM | MLP | LSTM | MLP | LSTM | MLP | LSTM |
| Hold-Out scheme | | | | | | | | | | |
| MAE | 27.458 | 27.637 | 9.639 | 9.609 | 6.577 | 6.548 | 42.740 | 41.514 | 10.441 | 10.105 |
| RMSE | 45.752 | 46.509 | 13.771 | 13.743 | 10.573 | 10.682 | 64.297 | 62.903 | 15.959 | 15.520 |
| sMAPE | 24.059 | 24.071 | 19.344 | 19.328 | 17.283 | 17.208 | 33.846 | 32.829 | 14.331 | 13.935 |
| Spearman r | 0.517 | 0.514 | 0.658 | 0.660 | 0.756 | 0.755 | 0.649 | 0.663 | 0.815 | 0.823 |
| Blocked Nested Cross-Validation scheme | | | | | | | | | | |
| MAE | 26.845 | 27.066 | 9.689 | 9.562 | 6.644 | 6.339 | 44.586 | 43.191 | 10.155 | 9.696 |
| RMSE | 44.718 | 45.923 | 13.885 | 13.808 | 10.840 | 10.722 | 64.785 | 63.690 | 16.162 | 15.752 |
| sMAPE | 23.590 | 23.607 | 19.499 | 19.240 | 17.280 | 16.639 | 35.54 | 34.569 | 14.012 | 13.467 |
| Spearman r | 0.523 | 0.520 | 0.654 | 0.657 | 0.756 | 0.766 | 0.632 | 0.648 | 0.815 | 0.817 |

**Table 3.** Performance results for the MLP and LSTM models were evaluated using The Hold-Out and the Blocked Nested Cross-Validation schemes. The summary of the results corresponds to one-hour ahead predictions of the concentration levels of the pollutant $PM_{10}$ evaluated in the last 60 days of the data set.

| Metrics | ATE | | CDM | | CRB | | HCH | | SMP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MLP | LSTM | MLP | LSTM | MLP | LSTM | MLP | LSTM | MLP | LSTM |
| Hold-Out scheme | | | | | | | | | | |
| MAE | 4.203 | 2.659 | 1.737 | 1.336 | 1.628 | 1.423 | 6.370 | 4.255 | 2.830 | 1.941 |
| RMSE | 5.724 | 3.706 | 2.235 | 1.732 | 2.192 | 1.844 | 8.324 | 5.837 | 3.602 | 2.299 |
| sMAPE | 3.646 | 2.411 | 3.581 | 2.830 | 4.269 | 3.927 | 4.636 | 3.224 | 4.063 | 2.867 |
| Spearman r | 0.986 | 0.991 | 0.973 | 0.982 | 0.967 | 0.974 | 0.981 | 0.988 | 0.982 | 0.990 |
| Blocked Nested Cross-Validation scheme | | | | | | | | | | |
| MAE | 4.217 | 2.720 | 1.829 | 1.325 | 1.645 | 1.333 | 6.621 | 4.561 | 2.749 | 1.841 |
| RMSE | 5.738 | 3.731 | 2.350 | 1.712 | 2.297 | 1.835 | 8.622 | 6.101 | 3.558 | 2.194 |
| sMAPE | 3.619 | 2.468 | 3.743 | 2.810 | 4.330 | 3.575 | 4.905 | 3.454 | 3.856 | 2.709 |
| Spearman r | 0.984 | 0.991 | 0.973 | 0.982 | 0.963 | 0.973 | 0.977 | 0.987 | 0.980 | 0.991 |

**Table 4.** Performance results for the MLP and LSTM models were evaluated using The Hold-Out and the Blocked Nested Cross-Validation schemes. The summary of the results corresponds to one-hour ahead predictions of the filtered time series of the concentration levels of the pollutant $PM_{10}$ evaluated in the last 60 days of the data set.

**Comparison of the present study with past studies.** This section shows the comparison of the present study with other previous studies on the evaluation and prediction of $PM_{10}$ in Lima, showing the duration of the study and the main findings. It is observed that our results agree with the other studies in that vehicular traffic is the main activity that causes critical episodes of $PM_{10}$, and this is exacerbated in the summer months.

- Silva et al.[28] shows that the highest concentrations of $PM_{10}$ were observed in the eastern part of the city. The main sources of particulate material are the large open areas, vehicular traffic, the commercialization of rubble, bricks, and cement. The highest concentrations of $PM_{10}$ are observed in summer. **Pollutant types**: $PM_{10}$, $PM_{2.5}$. **Duration of study**: 6 years (2010-2015).
- Reátegui-Romero at al.[73] show that, for the monitoring stations in the eastern zone, the highest concentrations of $PM_{10}$ are observed in the northern area of Lima, the Relative Humidity is inversely proportional to the concentrations of $PM_{10}$, higher peaks are observed in the summer month. **Pollutant types**: $PM_{10}$, $PM_{2.5}$. **Duration of study**: 2 months (February and July 2016).
- Sanchez et al.[10] show that there is a higher concentration of $PM_{10}$ in the areas with the greatest impact of vehicular traffic, reaching maximum concentrations of 476,8 $\mu g/m^3$ for Santa Anita station. They used the WRF-Chem model to predict $PM_{10}$ concentrations, obtaining low precision results. **Pollutant types**: $PM_{10}$. **Duration of study**: 33 days (2016).
- In our study, we have specified that the major sources of the pollutant $PM_{10}$ are the vehicle fleet, the industrial park, and overcrowding, reaching maximum peaks of 974 $\mu g/m^3$ at the HCH station. The highest concentrations were observed in the summer months. Artificial neural networks were used, specifically, the LSTM model under two validation schemes to predict $PM_{10}$ concentrations. The results showed good prediction performance for both low concentrations and critical episodes. **Pollutant types**: $PM_{10}$. **Duration of study**: 2 years (2017-2018).

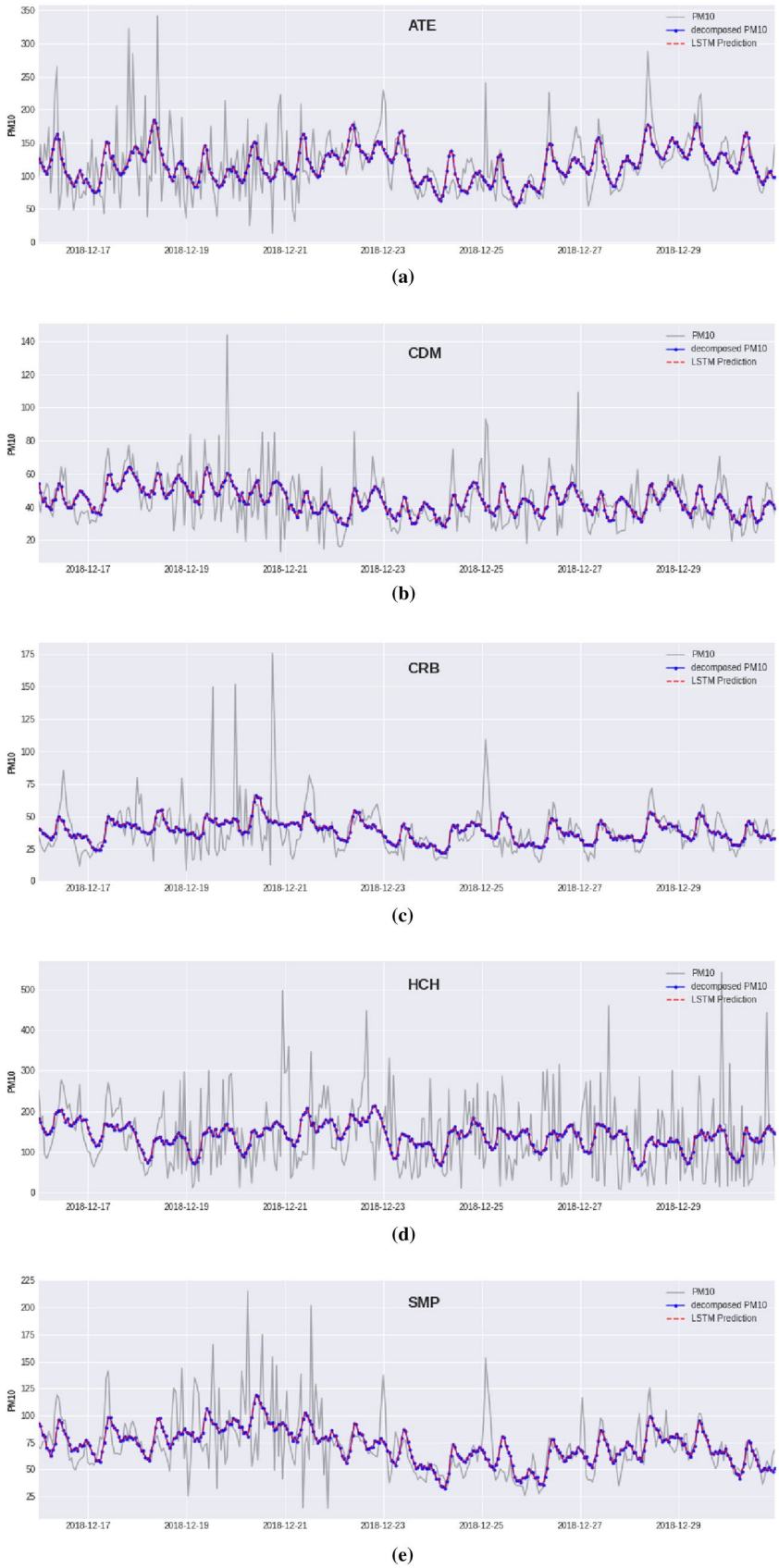



**Figure 12.** Plots for one-hour ahead predictions for the last 15 days of the regular component of the $PM_{10}$ concentration level using LSTM with the BNCV scheme. Predictions for the following monitoring stations: (**a**) ATE, (**b**) CDM, (**c**) CRB, (**d**) HCH, (**e**) SMP.

## Limitations

This study has some limitations. First, the number of data points represents a relatively short period (two years). A more extended period of hourly data may have allowed a more rigorous statistical analysis and more conclusive results. It is worth mentioning that the data related to $PM_{10}$ in Lima requires greater attention since many stations do not have the pertinent record of this pollutant, added to the scarce existing research related to this topic. Second, the collection of data related to other meteorological variables was also restricted since the monitoring stations do not record correctly for the most part. Third, the study does not consider data related to vehicular traffic or hospital care; the use of both variables may have enriched the research. However, our findings from the $PM_{10}$ analysis are consistent and complementary to a recent study showing the visual and exploratory aspect of the pollutant[31]. In addition, the MLP and LSTM architectures that allowed the analysis of predictions under two validation schemes are the precedent for future work with a predictive approach, being the first study in Lima that addresses the prediction of $PM_{10}$ using neural networks artificial. Likewise, it will be a support in the taking of preventive actions to critical environmental episodes.

## Conclusions

This study addressed the problem of forecasting $PM_{10}$ concentration on an hourly scale based on air quality indicators from five monitoring stations in Lima, Peru. A comparative study was accomplished between the MLP and LSTM neural networks evaluated with the Hold-Out and Blocked Nested Cross-Validation.

The MLP and LSTM can use the data from the previous period to accurately forecast the value of the $PM_{10}$ concentration in a short time ahead. They can learn the $PM_{10}$ concentration trends accurately. However, the performance is diminished when a station is subject to unpredictable external sources of pollution or due to short-term changes in climate and landforms (ATE and HCH). In this sense, the LSTM with the BNCV could better adapt to data from the monitoring stations that present episodes of extreme values. The results show that periods of moderate $PM_{10}$ concentration are predicted with very high precision. While for periods of high contamination, the model's accuracy is diminished, although in any case, it has a reasonable degree of predictability.

Using a high-performance model in air quality forecasting in large cities, such as Lima, can help develop critical health protection and prevention tools. Deep learning neural networks such as the LSTM are crucial in helping design public policies that prioritize improving air quality conditions to develop more sustainable cities.

The different configurations of the LSTM respond to the forecast of $PM_{10}$ events by selecting the relevant meteorological variables. Precisely, the essential property of the LSTM is that through its memory units, they can remember the patterns over time, which is beneficial when forecasting $PM_{10}$. In this sense, LSTM with BNCV could better adapt to data from the monitoring stations that present episodes of extreme values.

The results show that the $PM_{10}$ concentration prediction achieves better results with artificial intelligence methods since they are suitable for this type of approach. However, it is proposed to conduct this type of study with other cross-validation methods and hybrid and ensemble methods, giving greater precision in the prediction. This study will help in decision-making regarding air pollution mitigation and strategies, not only in Lima but also in other cities in the country and abroad. In this sense, this study of $PM_{10}$ could be extrapolated to other pollutants, both at a national and international level. In fact, a recent study[74] showed that genetic programming had higher prediction accuracy than artificial neural networks and was equally competent for peak predictions. Further works are required to explore other methods (hybrid or ensemble) to increase the accuracy of predictions.

As future work, we expect to apply other variants of deep learning models that include incremental learning[75], as well as to introduce self-identification techniques for the model identification[41,76].

Received: 2 September 2021; Accepted: 7 December 2021
Published online: 20 December 2021

## References

1. Organización Mundial de la Salud. Calidad del aire y salud. https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (2018).
2. Agency, U. E. P. Integrated science assessment (isa) for particulate matter (2009). EPA/600/R-08/139F.
3. Valdivia, S. A. P. Análisis temporal y espacial de la calidad del aire determinado por material particulado pm10 y pm2,5 en lima metropolitana. *Anales Científicos* **77**, 273–283. https://doi.org/10.21704/ac.v77i2.699 (2016).
4. Ordóñez-Aquino, C. & Sánchez-Ccoyllo, O. Caracterización química-morfológica del pm2, 5 en lima metropolitana mediante microscopía electrónica de barrido (meb). *Acta Nova* **8**, 397–420 (2018).
5. Organization, W. H. *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide* (World Health Organization, 2006).
6. Vahlsing, C. & Smith, K. R. Global review of national ambient air quality standards for pm 10 and so 2 (24 h). *Air Quality, Atmosphere& Health* **5**, 393–399. https://doi.org/10.1007/s11869-010-0131-2 (2012).
7. SINIA. Reglamento de estandares nacionales de calidad ambiental del aire. Tech. Rep., MINAM (2001).
8. EPA-US. National ambient air quality standards for particulate matter. Tech. Rep. 10, EPA (2013).
9. Chen, Y., Shi, R., Shu, S. & Gao, W. Ensemble and enhanced pm10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmos. Environ.* **74**, 346–359. https://doi.org/10.1016/j.atmosenv.2013.04.002 (2013).
10. Sánchez-Ccoyllo, O. R. *et al.* Modeling study of the particulate matter in lima with the wrf-chem model: Case study of April 2016. *Int. J. Appl. Eng. Res.: IJAER* **13**, 10129 (2018).
11. Chen, J. *et al.* Seasonal modeling of pm2. 5 in California's San Joaquin valley. *Atmos. Environ.***92**, 182–190. https://doi.org/10.1016/j.atmosenv.2014.04.030 (2014).
12. Saide, P. *et al.* Forecasting urban pm10 and pm2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using wrf–chem co tracer model. *Atmos. Environ.* **45**, 2769–2780. https://doi.org/10.1016/j.atmosenv.2011.02.001 (2011).
13. Li, X., Peng, L., Hu, Y., Shao, J. & Chi, T. Deep learning architecture for air quality predictions. *Environ. Sci. Pollut. Res.* **23**, 22408–22417. https://doi.org/10.1007/s11356-016-7812-9 (2016).
14. Li, C., Hsu, N. C. & Tsay, S.-C. A study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmos. Environ.* **45**, 3663–3675. https://doi.org/10.1016/j.atmosenv.2011.04.032 (2011).

15. Guarnaccia, C. *et al.* Arima models application to air pollution data in Monterrey, Mexico. *AIP Conf. Proc.* **1982**, 020041. https://doi.org/10.1063/1.5045447 (2018).
16. Adams, M. D. & Kanaroglou, P. S. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. *J. Environ. Manage.* **168**, 133–141. https://doi.org/10.1016/j.jenvman.2015.12.012 (2016).
17. Croitoru, C. & Nastase, I. A state of the art regarding urban air quality prediction models. *E3S Web Conf.* **32**, 01010. https://doi.org/10.1016/j.jenvman.2015.12.012 (2018).
18. Salini Calderón, G. & Pérez Jara, P. Estudio de series temporales de contaminación ambiental mediante técnicas de redes neuronales artificiales. *Ingeniare. Revista chilena de ingeniería* **14**, 284–290 (2006).
19. Guzmán, A. A. E. *et al.* Artificial neural network modeling of PM10 and PM2.5 in a tropical climate region: San Francisco de Campeche, Mexico. *Quim. Nova* **40**, 1025–1034. https://doi.org/10.21577/0100-4042.20170115 (2017).
20. Kök, İ., Şimşek, M. U. & Özdemir, S. A deep learning model for air quality prediction in smart cities. In *2017 IEEE International Conference on Big Data (Big Data)*, 1983–1990 (IEEE, 2017).
21. Jacinto Herrera, R. T. *Redes neuronales para predicción de contaminación del aire en Carabayllo-Lima*. Master's thesis, Universidad Nacional Federico Villarreal (2019).
22. Salas, R. & Bustos, A. Constructing a narx model for the prediction of the $pm_{10}$ air pollutant concentration. In *Encuentro Chileno de Computación, Jornada Chilena de Ciencias de la Computación, Valdivia, Chile. Nov. 7-12* (2005).
23. Athira, V., Geetha, P., Vinayakumar, R. & Soman, K. Deepairnet: Applying recurrent networks for air quality prediction. *Procedia Comput. Sci.* **132**, 1394–1403. https://doi.org/10.1016/j.procs.2018.05.068 (2018) (**International Conference on Computational Intelligence and Data Science**).
24. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM* **39**, 27–34. https://doi.org/10.1145/240455.240464 (1996).
25. Rojas, C. S. A. *Condições meteorológicas e níveis de poluição na Região Metropolitana de Lima–Perú*. Master's thesis, Universidad de Sao Paulo (2013).
26. INEI. Instituto nacional de estadistica e informatica (2020).
27. Sánchez Ccoyllo, O. & Ordoñez Aquino, C. Evaluación de la calidad del aire en lima metropolitana 2015. Tech. Rep., Dirección de Meteorología y Evaluación Ambiental Atmosférica-SENAMHI (2016). Accessed on 19-07-2021.
28. Silva, J. *et al.* Particulate matter levels in a South American megacity: the metropolitan area of Lima-Callao, Peru. *Environ. Monit. Assess.* **189**, 635. https://doi.org/10.1007/s10661-017-6327-2 (2017).
29. Navares, R. & Aznarte, J. L. Predicting air quality with deep learning lstm: Towards comprehensive models. *Eco. Inform.* **55**, 101019. https://doi.org/10.1016/j.ecoinf.2019.101019 (2020).
30. Rivera Poma, J. M. Desarrollo de un modelo dinámico para determinar la incidencia de los factores contaminantes del aire en la población de lima metropolitana. *Ind. Data* **15**, 054–062. https://doi.org/10.15381/idata.v15i2.6372 (2012).
31. Encalada-Malca, A. A., Cochachi-Bustamante, J. D., Rodrigues, P. C., Salas, R. & López-Gonzales, J. L. A spatio-temporal visualization approach of pm10 concentration data in metropolitan lima. *Atmosphere* **12**, 609. https://doi.org/10.3390/atmos12050609 (2021).
32. Royston, P. & White, I. R. Multiple imputation by chained equations (mice): Implementation in stata. *J. Stat. Softw.* **45**, 1–20. https://doi.org/10.18637/jss.v045.i04 (2011).
33. Cleveland, W. P. & Tiao, G. C. Decomposition of seasonal time series: a model for the census x–11 program. *J. Am. Stat. Assoc.* **71**, 581–587 (1976).
34. Cleveland, W. S., Freeny, A. E. & Graedel, T. The seasonal component of atmospheric co2: Information from new approaches to the decomposition of seasonal time series. *J. Geophys. Res.: Oceans* **88**, 10934–10946 (1983).
35. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, vol. 57, 61 (Austin, TX, 2010).
36. Allende, H., Moraga, C. & Salas, R. Artificial neural networks in time series forecasting: A comparative analysis. *Kybernetika* **38**, 685–707 (2002).
37. Allende, H., Salas, R., Torres, R. & Moraga, C. Modular neural network applied to non-stationary time series. In *Computational Intelligence, Theory and Applications, International Conference 8th Fuzzy Days, Dortmund, Germany, Sept. 29 - Oct. 01, 2004, vol. 33 of Advances in Soft Computing* (ed. Reusch, B.) 585–598 (Springer, Berlin, 2004). https://doi.org/10.1007/3-540-31182-3_54.
38. Veloz, A., Salas, R., Allende-Cid, H. & Allende, H. Sifar: Self-identification of lags of an autoregressive tsk-based model. In *2012 IEEE 42nd International Symposium on Multiple-Valued Logic*, 226–231 (IEEE, 2012).
39. Vivas, E., Allende-Cid, H., Salas, R. & Bravo, L. Polynomial and wavelet-type transfer function models to improve fisheries' landing forecasting with exogenous variables. *Entropy***21**, 1082. https://doi.org/10.3390/e21111082 (2019).
40. Vivas, E., Allende-Cid, H. & Salas, R. A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score. *Entropy* **22**, 1412. https://doi.org/10.3390/e22121412 (2020).
41. Morales, Y., Querales, M., Rosas, H., Allende-Cid, H. & Salas, R. A self-identification neuro-fuzzy inference framework for modeling rainfall-runoff in a Chilean watershed. *J. Hydrol.* **594**, 125910. https://doi.org/10.1016/j.jhydrol.2020.125910 (2021).
42. Xayasouk, T., Lee, H. & Lee, G. Air pollution prediction using long short-term memory (lstm) and deep autoencoder (dae) models. *Sustainability* **12**, 2570. https://doi.org/10.3390/su12062570 (2020).
43. Graves, A. Generating sequences with recurrent neural networks (2013).
44. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks (2015).
45. Bebis, G. & Georgiopoulos, M. Feed-forward neural networks. *IEEE Potentials* **13**, 27–31. https://doi.org/10.1109/45.329294 (1994).
46. Hornik, K. *et al.* Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).
47. Fu, M., Wang, W., Le, Z. & Khorram, M. S. Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. *Neural Comput. Appl.* **26**, 1789–1797 (2015).
48. Elbayoumi, M., Ramli, N. A. & Yusof, N. F. F. M. Development and comparison. Atmospheric of regression models and feedforward backpropagation neural network models to predict seasonal indoor pm2. 5–10 and pm2. 5 concentrations in naturally ventilated schools. *Pollut. Res.* **6**, 1013–1023 (2015).
49. Reddy, V., Yedavalli, P., Mohanty, S. & Nakhat, U. Deep air: Forecasting air pollution in beijing, china (2018).
50. Lipton, Z. C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning (2015).
51. Li, W. *et al.* Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (gru). *Inform. Process. Agric.* **8**, 185–193. https://doi.org/10.1016/j.inpa.2020.02.002 (2021).
52. Ayturan, Y. A., Ayturan, Z. C. & Altun, H. O. Air pollution modelling with deep learning: A review. *Int. J. Environ. Pollut. Environ. Model.* **1**, 58–62 (2018).
53. Yusof, N. F. F. M. *et al.* Monsoonal differences and probability distribution of pm 10 concentration. *Environ. Monit. Assess.* **163**, 655–667. https://doi.org/10.1007/s10661-009-0866-0 (2010).
54. Lasheras, F. S., Nieto, P. J. G., Gonzalo, E. G., Bonavera, L. & de Cos Juez, F. J. Evolution and forecasting of pm10 concentration at the port of Gijon (Spain). *Sci. Rep.* **10**, 1–12. https://doi.org/10.1038/s41598-020-68636-5 (2020).
55. Delgado-Villanueva, A. & Aguirre-Loayza, A. Modelamiento y evaluación del nivel de calidad del aire mediante el análisis de grey clustering, estudio de caso lima metropolitana. *Tecnia* **30**, 114–120. https://doi.org/10.21754/tecnia.v30i1.588 (2020).

56. Sahu, V., Elumalai, S. P., Gautam, S., Singh, N. K. & Singh, P. Characterization of indoor settled dust and investigation of indoor air quality in different micro-environments. *Int. J. Environ. Health Res.* **28**, 419–431 (2018).
57. Silva, J. S., Rojas, J. P., Norabuena, M. & Seguel, R. J. Ozone and volatile organic compounds in the metropolitan area of Lima-Callao, Peru. *Air Qual. Atmos. Health* **11**, 993–1008. https://doi.org/10.1007/s11869-018-0604-2 (2018).
58. Moreno Jiménez, A. *et al.* La concentración de partículas en el aire: análisis estadístico de la relación espacial entre medidas de superficie y del sensor modis para dos tipos de tiempo en la comunidad de madrid. *Investigaciones Geográficas* **73**, 189–209. https://doi.org/10.14198/INGEO2020.MJCTMA (2020).
59. Ahmadi, S. *et al.* Assessment of health impacts attributed to pm10 exposure during 2015–2017 in Zabol city, Iran. *Int. J. Environ. Sci. Technol.* **18**, 1–14 (2021).
60. Dahmardeh Behrooz, R., Kaskaoutis, D., Grivas, G. & Mihalopoulos, N. Human health risk assessment for toxic elements in the extreme ambient dust conditions observed in Sistan, Iran. *Chemosphere* **262**, 127835. https://doi.org/10.1016/j.chemosphere.2020.127835 (2021).
61. Sahin, F., Kara, M. K., Koc, A. & Sahin, G. Multi-criteria decision-making using gis-ahp for air pollution problem in Igdir province/Turkey. *Environ. Sci. Pollut. Res.* **27**, 36215–36230. https://doi.org/10.1007/s11356-020-09710-3 (2020).
62. Taheri Shahraiyni, H. & Sodoudi, S. Statistical modeling approaches for pm10 prediction in urban areas; a review of 21st-century studies. *Atmosphere* **7**, 15. https://doi.org/10.3390/atmos7020015 (2016).
63. Gautam, S., Talatiya, A., Patel, M., Chabhadiya, K. & Pathak, P. Personal exposure to air pollutants from winter season bonfires in rural areas of Gujarat, India. *Exposure and Health* **12**, 89–97 (2020).
64. Ilizarbe-Gonzáles, G. M. *et al.* Chemical characteristics and identification of pm10 sources in two districts of Lima, Peru. *Dyna* **87**, 57–65. https://doi.org/10.15446/dyna.v87n215.83688 (2020).
65. OTD. Tránsito de vehículos a nivel nacional aumentó 15,5%. Tech. Rep. 076, INEI (2018).
66. Jain, S., Sharma, S., Vijayan, N. & Mandal, T. Seasonal characteristics of aerosols (pm2.5 and pm10) and their source apportionment using pmf: A four year study over delhi, india. *Environ. Pollut.* **262**, 114337. https://doi.org/10.1016/j.envpol.2020.114337 (2020).
67. Alolayan, M. A., Brown, K. W., Evans, J. S., Bouhamra, W. S. & Koutrakis, P. Source apportionment of fine particles in Kuwait city. *Sci. Total Environ.* **448**, 14–25. https://doi.org/10.1016/j.scitotenv.2012.11.090 (2013).
68. Owoade, K. O. *et al.* Chemical compositions and source identification of particulate matter (pm2. 5 and pm2. 5–10) from a scrap iron and steel smelting industry along the ife–ibadan highway, nigeria. *Atmos. Pollut. Res.* **6**, 107–119. https://doi.org/10.5094/APR.2015.013 (2015).
69. Capel Molina, J. J. Lima, un clima de desierto litoral. *Anales de Geografía de la Universidad Complutense* **19**, 25 (1999).
70. Concepción, E. & Rodríguez, J. Informe nacional de la calidad del aire 2013-2014. https://bit.ly/36KTRAM/ (2014). Accessed on 18-07-2021.
71. Iparraguirre Medina, R. L. & Valdivia Torres, A. G. *Caracterización y problemática de las ladrilleras en Huachipa-Lurigancho-Lima. 2018*. Master's thesis, Universidad Católica Sedes Sapientiae, http://repositorio.ucss.edu.pe/handle/UCSS/735 (2018). Accessed on 18-07-2021.
72. Álvarez-Liébana, J. & Ruiz-Medina, M. Prediction of air pollutants pm 10 by arbx (1) processes. *Stoch. Env. Res. Risk Assess.* **33**, 1721–1736. https://doi.org/10.1007/s00477-019-01712-z (2019).
73. Reátegui-Romero, W. *et al.* Behavior of the average concentrations as well as their pm10 and pm2. 5 variability in the metropolitan area of Lima, Peru: Case study February and July 2016. *International Journal of Environmental Science and Development* **12** (2021).
74. Tikhe, S., Khare, K. & Londhe, S. Air quality forecasting using soft computing techniques (2020).
75. Mellado, D., Saavedra, C., Chabert, S., Torres, R. & Salas, R. Self-improving generative artificial neural network for pseudorehearsal incremental class learning. *Algorithms***12**, https://doi.org/10.3390/a12100206 (2019).
76. Veloz, A., Salas, R., Allende-Cid, H., Allende, H. & Moraga, C. Identification of lags in nonlinear autoregressive time series using a flexible fuzzy model. *Neural Process. Lett.* **43**, 641–666. https://doi.org/10.1007/s11063-015-9438-1 (2016).


## Acknowledgements

Javier Linkolk López-Gonzales acknowledges financial support from the ANID scholarship. The works of Chardin Hoyos Cordova and Manuel Niño Lopez Portocarrero were supported by the grant *Beca 18* of the national government. P.C. Rodrigues acknowledges financial support from the Brazilian National Council for Scientific and Technological (CNPq) grant "bolsa de produtividade PQ-2" 305852/2019-1. The authors are grateful to the *Servicio Nacional de Meteorología e Hidrología* (SENAMHI) for providing the air quality data used in this study.

## Author contributions

J.L.L-G., C.H.C., R.S. and M.N.L.P conceived the experiment(s), R.S., R.T. and P.C.R. conducted the experiment(s), C.H.C., R.S., R.T., J.L.L-G and P.C.R. analyzed the results. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.L.L.-G.