Vidalon

ExtLearModelPM10_for_environments.pdf

Universidad Peruana Union

Detalles del documento

Identificador de la entrega

trn:oid:::29566:546759908

Fecha de entrega

16 ene 2026, 10:50 a.m. GMT-5

Fecha de descarga

20 abr 2026, 8:52 a.m. GMT-5

Nombre del archivo

ExtLearModelPM10_for_environments.pdf

Tamaño del archivo

626.5 KB

11 páginas

5726 palabras

28.540 caracteres

12% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado

Exclusiones

- N.º de fuentes excluidas

Grupos de coincidencias

67 Sin cita o referencia 12%
Coincidencias sin una citación ni comillas en el texto

0 Faltan citas 0%
Coincidencias que siguen siendo muy similar al material fuente

0 Falta referencia 0%
Las coincidencias tienen comillas, pero no una citación correcta en el texto

0 Con comillas y referencia 0%
Coincidencias de citación en el texto, pero sin comillas

Fuentes principales

6% Fuentes de Internet
10% Publicaciones
7% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Grupos de coincidencias

67 Sin cita o referencia 12%
Coincidencias sin una citación ni comillas en el texto

0 Faltan citas 0%
Coincidencias que siguen siendo muy similar al material fuente

0 Falta referencia 0%
Las coincidencias tienen comillas, pero no una citación correcta en el texto

0 Con comillas y referencia 0%
Coincidencias de citación en el texto, pero sin comillas

Fuentes principales

6% Fuentes de Internet
10% Publicaciones
7% Trabajos entregados (trabajos del estudiante)

Fuentes principales

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1 Publicación
Karlsson, Johannes. "Effective System Solutions for Battery Electric Trucks", Chal... <1%

2 Internet
eprints.soton.ac.uk <1%

3 Publicación
Vieri Mastropietro. "Ising Models with Four Spin Interaction at Criticality", Comm... <1%

4 Internet
cris.upeu.edu.pe <1%

5 Internet
purehost.bath.ac.uk <1%

6 Publicación
Hasnain Iftikhar, Josue E. Turpo-Chaparro, Paulo Canas Rodrigues, Javier Linkolk ... <1%

7 Trabajos entregados
Kamehameha School on 2009-09-09 <1%

8 Publicación
Wilfredo Meza Cuba, Juan Carlos Huaman Alfaro, Hasnain Iftikhar, Javier Linkolk ... <1%

9 Internet
www.ncbi.nlm.nih.gov <1%

10 Publicación
Natalí Carbo-Bustinza, Hasnain Iftikhar, Marisol Belmonte, Rita Jaqueline Cabello-... <1%

11 Publicación

Hasnain Iftikhar, Josue E. Turpo-Chaparro, Paulo Canas Rodrigues, Javier Linkolk ... <1%

12 Publicación

Tilahun, Seifu A., Christian D. Guzman, Assefa D. Zegeye, Dessalegn C. Dagnew, A... <1%

13 Trabajos entregados

Banaras Hindu University on 2024-04-29 <1%

14 Publicación

Huda M. Alshanbari, Hasnain Iftikhar, Faridoon Khan, Moeeba Rind, Zubair Ahma... <1%

15 Publicación

Chardin Hoyos Cordova, Manuel Niño Lopez Portocarrero, Rodrigo Salas, Romina ... <1%

16 Trabajos entregados

Saxion Brightspace on 2024-12-20 <1%

17 Internet

www.medrxiv.org <1%

18 Internet

worldwidescience.org <1%

19 Publicación

Bernard Hsieh, Mark Jourdan. "Chapter 2 An Operational Riverflow Prediction Sys... <1%

20 Internet

council.science <1%

21 Internet

journals.sta.uwi.edu <1%

22 Internet

www.devx.com <1%

23 Publicación

Hasnain Iftikhar, Faridoon Khan, Elías A. Torres Armas, Paulo Canas Rodrigues, Ja... <1%

24 Internet

millengustavo.github.io <1%

25 Internet

repository.charlotte.edu <1%

26 Trabajos entregados

Imam Abdulrahman Bin Faisal University on 2025-12-17 <1%

27 Trabajos entregados

Macquarie University on 2018-05-27 <1%

28 Internet

oa.upm.es <1%

29 Internet

umpir.ump.edu.my <1%

30 Publicación

Al Hadhrami, Siham. "A Numerical Modeling Study of the Impact of Air Quality in ... <1%

31 Publicación

Alexandra Abigail Encalada-Malca, Javier David Cochachi-Bustamante, Paulo Cana... <1%

32 Trabajos entregados

University of Cape Town on 2011-09-26 <1%

33 Internet

eares.org <1%

34 Internet

staging.preprints.org <1%

35 Internet

www.biorxiv.org <1%

36 Internet

www.gnedenko.net <1%

37 Publicación

Hasnain Iftikhar, Muhammad Daniyal, Moiz Qureshi, Kassim Tawiah, Richard Kwa... <1%

38 Trabajos entregados

Osmania University, Hyderabad on 2025-07-25 <1%

39	Trabajos entregados		
University of Southampton on 2018-09-09			<1%

MDPI

Article

Modeling and forecasting air pollution using extreme learning model

Diana Sánchez [1], María Vidalón [1], Ivan Medina [3], Hasnain Iftikhar [2] and Javier Linkolk López-Gonzales [1,*]

1 Escuela de Posgrado, Universidad Peruana Unión, Lima, Peru
2 Affiliation 2; e-mail@e-mail.com
* Correspondence: javierlinkolk@gmail.com

Received:
Revised:
Accepted:
Published:

Citation: Lastname, F.; Lastname, F.; Lastname, F. Modeling and forecasting air pollution using extreme learning model. *Journal Not Specified* **2025**, *1*, 0. https://doi.org/

1. Introduction

Metropolitan Lima faces a serious environmental pollution problem, being considered one of the most polluted cities in Latin America. The main sources of this pollution are vehicle emissions and industrial activities, which release gases, vapors, and solid particles into the atmosphere that are harmful to human health. This situation has led to poor air quality in the city, exceeding the limits recommended by the World Health Organization [1]

Air pollution is a significant environmental and public health problem in Metropolitan Lima, where concentrations of particulate matter less than 10 micrometers (PM10) often exceed the limits recommended by the World Health Organization (WHO). This pollutant is associated with various respiratory and cardiovascular conditions, highlighting the need to develop accurate models to forecast its levels and design effective mitigation strategies.

In Lima, PM10 forecasting and modeling face challenges due to factors such as complex topography, uncontrolled urban growth, and variability in emission sources. These factors hinder the accuracy of predictive models, underscoring the importance of appropriately selecting and adapting modeling methodologies.

Across South America and globally, various approaches have been implemented to model and forecast PM10 concentrations. For example, multiple linear regression (MLR) has been used to identify linear relationships between PM10 concentrations and meteorological variables. However, their ability to capture nonlinear relationships is limited. On the other hand, models such as extreme learning machines (ELM) and autoregressive neural networks (ARN) have demonstrated greater ability to model complex and nonlinear relationships in air quality data. In studies conducted in China, these approaches have shown significant improvements in PM10 forecasting accuracy compared to traditional models.

Furthermore, autoregressive integrated moving average (ARIMA) and nonparametric autoregressive (NAR) models have been applied in various regions for PM10 time series forecasting. For example, in Europe, the ARIMA model has been effective in capturing temporal patterns in PM10 concentrations, while NAR models have offered flexibility to adapt to the specific characteristics of local data. The application of these methods in Metropolitan Lima could improve forecast accuracy and contribute to the development of more informed public policies for air quality management.

2. Methodology

In this section, we will provide a comprehensive overview of the various models and methods used to construct the proposed modeling and forecasting approach for hourly (PM_{10}) forecasts. Thus, the subsequent subsections are for detailed information on each model and method.

2.1. The proposed forecasting technique

The primary goal of this study was to predict the particulate matter concentration one day ahead at five monitoring stations: ATE, CDM, CRB, HCH, and SMP in Lima, Peru. Let Z_h represent the particulate matter concentration for the hour h^{th}. To accurately account for the changes in (PM_{10}) concentration over time, we suggest modeling Z_h in the following way:

$$Z_h = z_h + e_h \tag{1}$$

The PM_{10} concentration series (μgm^3) is split into two parts: a deterministic component (denoted as z_h) and a stochastic component (denoted as e_h). z_h includes the trend (long-term pattern), hourly cycles, yearly, and weekly, while e_h represents random fluctuations. Mathematically, z_h is defined as follows:

$$z_h = l_h + a_h + w_h + h_h \tag{2}$$

The symbol l_h, a_h, w_h, and h_h represents the long-term trend, the yearly, weekly, and hourly periodicities, respectively. On the other hand, e_h is a stochastic component, also known as residuals, that defines the random dynamics. A multiple linear regression model is used to estimate the deterministic component z_h. In order to estimate stochastic components, this study examines four distinct models for univariate different models: two artificial-intelligent base models, such as extreme learning machine and artificial neural network autoregressive models, and two standard time series models, such as nonparametric autoregressive and autoregressive moving average models. As a result, there are four possible combinations for comparison purposes when deterministic and stochastic models are combined.

2.1.1. Modeling the Deterministic Component

This section will discuss how to estimate the deterministic component using a multiple linear regression model. To achieve this, we will model the response variable Z_h linearly by estimating the trend (long-run) component l_h through linear regression for time h. Additionally, we will describe the yearly, weekly, and hourly periodicity using dummies:

$$z_h = l_h + \sum_{i=1}^{3} \zeta_i Y_{i,h} + \sum_{i=4}^{11} \zeta_i W_{i,h} + \sum_{i=12}^{35} \zeta_i I_{i,h} \tag{3}$$

The variable $Y_{i,h}$ is assigned a value of 1 when h refers to the ith year of the series and 0 otherwise. The variable $W_{i,h}$ is assigned a value of 1 when h refers to the ith day of the week and 0 otherwise. The variable $I_{i,h}$ is assigned a value of 1 when h refers to the ith hour of the day and 0 otherwise. The regression coefficients associated with these components are determined using the ordinary least square method. After obtaining all the regression coefficients, the estimated equation is presented.

$$\hat{z}_h = \hat{\zeta}_0 l_h + \sum_{i=1}^{3} \hat{\zeta}_i Y_{i,h} + \sum_{i=4}^{11} \hat{\zeta}_i W_{i,h} + \sum_{i=12}^{35} \hat{\zeta}_i I_{i,h} \tag{4}$$

Once the estimated deterministic component is obtained, the residual or stochastic component can be derived as:

$$e_h = Z_h - \hat{z}_h \tag{5}$$

2.1.2. Modeling the Stochastic Component

The residual series was obtained from both models using a multiple linear regression model to estimate the stochastic component in this work. However, to model and forecast the stochastic component, four different univariate models are considered: two artificial-intelligence base models, such as extreme learning machine and artificial neural network autoregressive models, and two standard time series models, such as nonparametric autoregressive and autoregressive moving average models. Details on these models are the following:

Extreme-learning-Machine Model (ELM): An ML approach called (ELM) was first presented by [2] and is intended for classifying and regression-related tasks. With its unique training methodology, this particular sort of feed-forward neural network (FF-NN) sets itself apart from more conventional approaches. Mathematically, the model can be presented as

$$\hat{z} = h(x) \cdot \beta \tag{6}$$

where in Equation 6 $\hat{z}, h(x), \beta$ is predicted, the activation function $h(x) = g(W \cdot z + b)$ and the weight matrix of the series accordingly.

The Artificial Neural Network Auto-regressive Model: An artificial intelligence approach (ANN) uses historical data to predict future time series values. A mathematical function, represented by the notation

$$z_t = (\psi_{z-1}) + (\psi_{z-2}) + \cdots + (\psi_{z-n}) \tag{7}$$

where the time delay parameter is indicated by n, is used by the model to account for the historical values of the time series [3]. In order to minimize the discrepancy between predicted and actual values, the ANN model is developed via gradient descent and back-propagation (BP). The ANN model forecasts by first figuring out the AR order, which is trained on a data set, and then figuring out the ideal number of hidden nodes. In order to prevent overfitting the model, significant attention must be paid to the number of training iterations. This research uses a standard ANN conception with one hidden layer and three nodes.

Auto-regressive integrated moving average (ARIMA) model: A type of regression modeling that examines the strength of a single dependent variable when compared to other changing variables is the auto-regressive integrated moving average model[4].
This model is an addition of three components, which are defined as a model that displays a variable that evolves and regresses on its own lag values, which is known as an AR model. d represents the actual observational differencing that stabilizes the series. With lagged observations, the MA model considers the relationship between the actual time series and estimated residual errors. The optimal parameters of ARIMA models are estimates by (MLE) equations. Mathematically, the ARIMA model is demonstrated as

$$z_t = \alpha_0 + \alpha_1 z_{t-1} + \alpha_2 z_{t-2} + \cdots + \alpha_p z_{t-p} + \epsilon + \theta_1 z_{t-1} + \theta_2 z_{t-2} + \cdots + \theta_q z_{t-q} \tag{8}$$

Where in Equation (3)
The value of auto-regressive terms is indicated by p.
For stationary behavior (if needed), d specifies the non-seasonal deviations required.
q represents the value of lags in the estimated errors and ϵ represents the overall error term of the model.

Nonparametric Autoregressive Model: The additive nonparametric counterpart of the AR process leads to the additive model (NPAR), where the association between e_h and its previous terms do not have any specific parametric form, letting, probably, for any sort of nonlinearity which is stated as:

$$e_h = q_1(e_{d\text{-}1}) + q_2(e_{d\text{-}2}) + \ldots + q_n(e_{d\text{-}n}) + \varepsilon_d \tag{9}$$

where $q_j (j = 1, 2, \cdots, n)$ are showing smoothing functions and describe the association between e_h and its previous values[5]. In this work, functions q_i are denoted by cubic regression splines. As in the case of the parametric form, considered lags 5 while estimating NPAR.

2.2. Accuracy Measures

Four different standard accuracy measures were calculated to validate the performance of the proposed modeling and forecasting technique, including mean absolute error (MAE), Mean absolute percentage error (MAPE), mean deviation absolute error (MDAE), and Symmetric mean absolute percentage error (SMAPE)[6]. The MAE, MAPE, MDAE, and SMAPE formulae are shown below:

$$\mathrm{MAE} = \frac{1}{H}\sum_{h=1}^{H} |Z_h - \hat{Z}_h|, \tag{10}$$

$$\mathrm{MAPE} = \frac{1}{H}\sum_{h=1}^{H} \left|\frac{Z_h - \hat{Z}_h}{Z_h}\right|, \tag{11}$$

$$\mathrm{MDAE} = \frac{1}{H}\sum_{h=1}^{H} \frac{|Z_h - \hat{Z}_h|}{Z_h}, \tag{12}$$

$$\mathrm{SMAPE} = \frac{1}{H}\sum_{h=1}^{H} \frac{|Z_h - \hat{Z}_h|}{(|Z_h| + |\hat{Z}_h|)/2}, \tag{13}$$

(14)

In the above equations, Z_h is observed and $\hat{Z}_h$ is forecasted (PM_{10}) cencertaron values for h^{th} observation (h=1, 2, . . ., 2256=H).

2.3. Expanding window method

In this work, we used the expanding window method to forecast one, two, three, and seven days ahead forecast. Hence, the expanding window method is a tool for forecasting and time series analysis. Using a fixed-size window of previous data that "expands" when new data becomes available is how it works. The effectiveness of forecasting models is frequently evaluated using this technique by comparing predicted values to actual values as the window advances in time. It is useful to replicate scenarios for real-time forecasting and assess how well the model can adjust to shifting patterns and trends. To do this, the expanding window approach involves the following main steps:

1. Select a Starting Point: Choose a starting point in your time series data.
2. Define Window Size: Decide on the initial window size (e.g., the number of data points) that will use to train your model.
3. Train and Forecast: Train your forecasting model using the data within the initial window, and then use the model to make predictions for the next data point or period.
4. Evaluate: Compare the model's predictions against the data for the next period. Calculate an error metric (such as mean absolute error or root mean squared error) to quantify the accuracy of the model's predictions.
5. Expand the Window: Move the window forward by adding one more data point to the training set. This means you'll have a larger training window for the next iteration.

6. Repeat: Train the model again using the updated training window and make forecasts for the next period. Re-evaluate the model's performance.
7. Continue: Repeat steps 5 and 6, gradually increasing the size of the training window and assessing the model's performance as it adapts to the expanding dataset. However, a graphical presentation of the expanding window forecasting technique is shown in Figure 1.

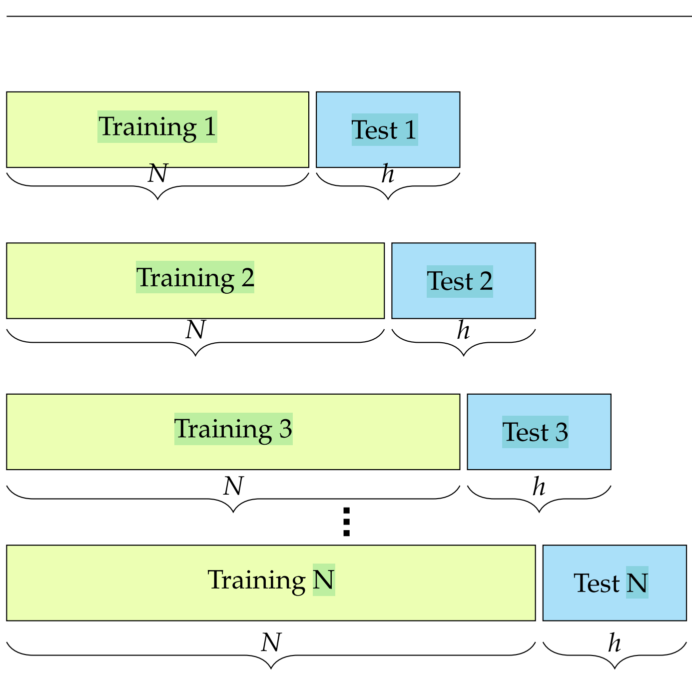

Figure 1. Expanding window method

3. Forecasting Results and Discussion

Table 1 displays statistical summaries for five datasets (ATE, CDM, CRB, HCH, and SMP) for both the original and log-transformed versions. The statistics provided include minimum, first quartile, median, mean, mode, variance, standard deviation, skewness, kurtosis, third quartile, and maximum.

In the original datasets, the mean and median values indicate a degree of skewness, as most means are higher than their respective medians. The variance and standard deviation reflect the variability of the data, with HCH showcasing the highest variance (8404.3375) and standard deviation (91.6752), while CDM displays the lowest variance (605.5354) and standard deviation (24.6076). The values for skewness and kurtosis reveal the distributions' shapes; the positive skewness across the board indicates right-skewed distributions, while high kurtosis values (particularly for ATE and CRB) point to heavy-tailed distributions with significant extremes.

When the log transformation is applied, the statistical characteristics change. The log-transformed datasets exhibit considerably lower variance and standard deviation, indicating a decrease in dispersion. Skewness values decline significantly, with some even turning negative, suggesting a more symmetrical distribution. Additionally, kurtosis values are reduced, indicating a distribution that is closer to normal. These alterations demonstrate that the log transformation effectively normalizes the data by alleviating skewness and extreme values, thereby making it better suited for statistical analysis.

Table 1. Statistical Summary of Original and Log Series

Statistic	Original Series					Log Series				
	ATE	CDM	CRB	HCH	SMP	ATE	CDM	CRB	HCH	SMP
Minimum	6.41	6.08	5.44	5.21	7.77	1.8579	1.805	1.6938	1.6506	2.0503
1st Quartile	82.9	35.84	31.49	62.1	61.95	4.4176	3.5791	3.4497	4.1287	4.1263
Median	112.05	47.7	42.285	113.2	82.1	4.7189	3.8649	3.7444	4.7292	4.4079
Mean	121.5588	52.2989	48.6852	130.0331	86.0488	4.6865	3.8596	3.7579	4.5941	4.3651
Mode	90	82.3	28.47	85.7	86	4.4998	4.4104	3.3489	4.4509	4.3543
Variance	3635.7541	605.5354	806.0286	8404.3375	1276.4097	0.2438	0.1967	0.2466	0.6501	0.1952
Standard D	60.2972	24.6076	28.3906	91.6752	35.7269	0.4937	0.4435	0.4966	0.8063	0.4418
Skewness	2.0825	2.3034	3.2383	1.531	1.0046	-0.5954	-0.1263	0.1089	-0.6527	-0.6521
Kurtosis	11.0643	18.2321	22.2645	4.8828	2.861	1.865	0.5333	0.8773	0.1248	1.1329
3rd Quartile	148	63.45	58.45	176.5	105.1	4.9972	4.1503	4.0682	5.1733	4.6549
Maximum	931	463.6	488.02	974	426.8	6.8363	6.139	6.1904	6.8814	6.0563

The illustration displays hourly concentrations of particulate matter (PM_{10}) recorded at five monitoring sites in Lima, Peru- ATE, CDM, CRB, HCH, and SMP- from January 1, 2017, to December 31, 2018. Each site shows unique trends in PM_{10} levels, reflecting localized differences in air quality due to factors such as traffic volume, industrial operations, and weather conditions.

HCH exhibits the most significant variations of the five sites, with PM_{10} levels soaring to nearly 1000 $\mu g/m^3$ during peak times, indicating severe pollution events. ATE also displays elevated numbers, with some hourly measurements nearing 800 $\mu g/m^3$. Conversely, CDM and SMP present relatively lower PM_{10} levels, peaking at approximately 400 $\mu g/m^3$, implying better air quality in these regions. CRB shows mid-range values, topping off at around 500 $\mu g/m^3$.

The differences in PM_{10} levels across the monitoring sites underscore the spatial diversity of air pollution in Lima. These variations can be linked to local emission sources, such as vehicle traffic and industrial discharges, as well as meteorological aspects that affect the dispersal of pollutants. Grasping these trends is vital for formulating specific air quality management measures aimed at reducing health hazards tied to particulate matter exposure.

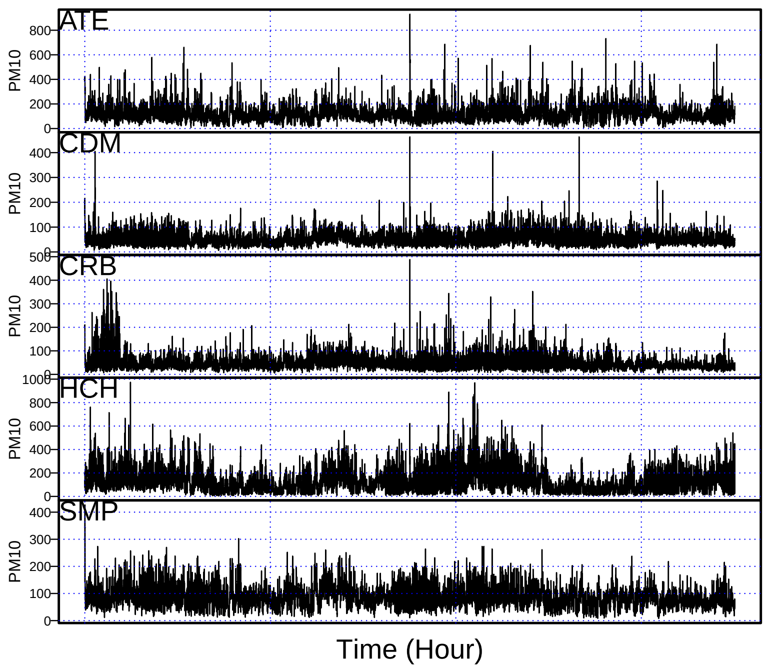

Figure 2. Hourly measurements of PM_{10} levels were recorded from January 1, 2017, to December 31, 2018, at Lima, Peru's monitoring locations (ATE, CDM, CRB, HCH, and SMP).

Table 2 illustrates the statistical outcomes of different forecasting models for predicting one-hour-ahead PM_{10} levels at five monitoring locations in Lima, Peru—ATE, CDM, CRB, HCH, and SMP—during the period from January 1, 2017, to December 31, 2018. The models assessed include the Extreme Learning Machine (ELM), Neural Network Autoregression (NNAR), Non-Parametric Autoregression (NPAR), Autoregressive Moving Average (ARMA), Seasonal Naïve (SNaive), and a basic Naïve model. The evaluation metrics utilized are Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Median Absolute Error (MDAE), and Root Mean Squared Error (RMSE).

Table 2. One-hour ahead PM_{10} forecasting performance at different monitoring sites.

Models	MAE	MAPE	MDAE	RMSE
ATE				
ELM	20.8337	16.9198	15.8974	27.5653
NNAR	20.7188	16.8033	17.0075	27.8599
NPAR	27.1006	22.1046	21.588	35.7430
ARMA	25.8563	20.848	21.5277	34.1282
SNaive	23.528	19.575	18.8894	32.1302
Naive	35.3853	29.7344	29.426	46.4898
CDM				
ELM	6.318	12.9253	5.1624	8.3241
NNAR	6.726	13.9242	5.377	8.7704
NPAR	7.4166	16.0692	6.3092	9.2050
ARMA	7.326	15.6247	6.0393	9.484
SNaive	7.9667	15.6985	6.385	10.6473
Naive	11.7371	24.4998	9.3425	15.4628
CRB				
ELM	24.852	60.8873	30.2088	31.3642
NNAR	24.7711	60.6885	30.2088	31.3502
NPAR	25.3737	62.7827	30.2088	31.4666
ARMA	25.1251	61.9045	30.2088	31.423
SNaive	26.5902	66.6410	30.2088	31.8399
Naive	27.898	70.11	31.499	32.5102
HCH				
ELM	22.5719	20.666	16.334	29.0691
NNAR	22.9715	21.1081	15.5004	29.7016
NPAR	34.8553	30.2058	27.7463	44.3691
ARMA	30.4869	27.5275	23.7315	40.402
SNaive	24.0789	23.1074	18.6612	31.5933
Naive	51.6995	50.1085	45.705	64.934
SMP				
ELM	12.6341	17.8862	8.6133	17.6609
NNAR	13.0248	18.4791	8.6487	17.9268
NPAR	14.7091	21.3742	12.3948	19.232
ARMA	15.2844	23.1454	13.1293	19.7294
SNaive	18.0323	23.9148	13.582	23.5309
Naive	27.0142	40.976	22.6667	32.1947

Across the various monitoring locations, the ELM and NNAR models consistently exhibited lower error metrics than the other approaches, highlighting their superior forecasting performance. Conversely, the Naïve and SNaive models yielded the highest error metrics, indicating their ineffectiveness in capturing variations in PM_{10}. The ARMA model typically performed worse than the ELM and NNAR but provided better accuracy than the models based on the Naïve method. At the ATE station, for instance, the ELM model achieved the smallest RMSE (27.5653), while the NNAR model followed closely with an RMSE of 27.8599. Both ELM and NNAR recorded the lowest MAE and RMSE values for the CDM location, further confirming their reliability. Results from the CRB station also indicated that ELM was the top-performing model, achieving an RMSE of 31.3642, with NNAR closely trailing. At HCH, ELM again surpassed all other models with an RMSE of 29.0691, while the ARMA model showed relatively higher error metrics. Finally, ELM produced the lowest RMSE of 17.6609 for the SMP station, with NNAR not far behind. These findings suggest that machine learning-based models such as ELM and NNAR provide improved predictive capabilities for PM_{10} levels compared to traditional statistical models.

On the other hand, to check the performance of proposed forecasting models in graphical form, the bar plots shown in the Figure 3 display the MAE, MDAE, MAPE, and RMSE performance for various forecasting models—ELM, NNAR, NPAR, ARMA, SNaive, and Naive—across five monitoring locations: ATE, CDM, CRB, HCH, and SMP. The bar charts illustrating the forecasting performance metrics for PM_{10}- including MAE, MAPE, MDAE, and RMSE—offer valuable information regarding the accuracy of various forecasting models at different monitoring locations (ATE, CDM, CRB, HCH, SMP). The MAE values reflect the average deviation of the predicted PM_{10} levels from the actual measurements. Generally, the ELM and NNAR models show lower MAE values when compared to ARMA, SNaive, and Naive models, indicating that these machine learning approaches exhibit better performance in minimizing absolute forecasting inaccuracies. On the other hand, the MAPE figures represent the percentage divergence of forecasted figures from real values. The ATE and CRB monitoring locations show relatively elevated MAPE values, particularly with traditional statistical models (ARMA, SNaive, Naive), which implies that these models encounter difficulties in delivering precise predictions in those areas. ELM and NNAR models showcase lower MAPE values once again, highlighting their enhanced forecasting proficiency. Similarly, the MDAE measure emphasizes the robustness of the error distribution and exhibits a trend similar to MAE. Locations such as CRB and HCH display higher MDAE values, indicating greater variability in their forecasting inaccuracies. The machine learning models (ELM and NNAR) consistently outperform traditional models, achieving lower MDAE values and demonstrating their capability to maintain reliable predictions. However, RMSE values emphasize the squared discrepancies in forecasting, placing greater emphasis on larger errors. Throughout all monitoring sites, traditional forecasting models such as ARMA, SNaive, and Naive show significantly elevated RMSE values, suggesting a higher occurrence of substantial prediction errors. In contrast, ELM and NNAR models present lower RMSE values, further validating their effectiveness in delivering more consistent and precise forecasts.

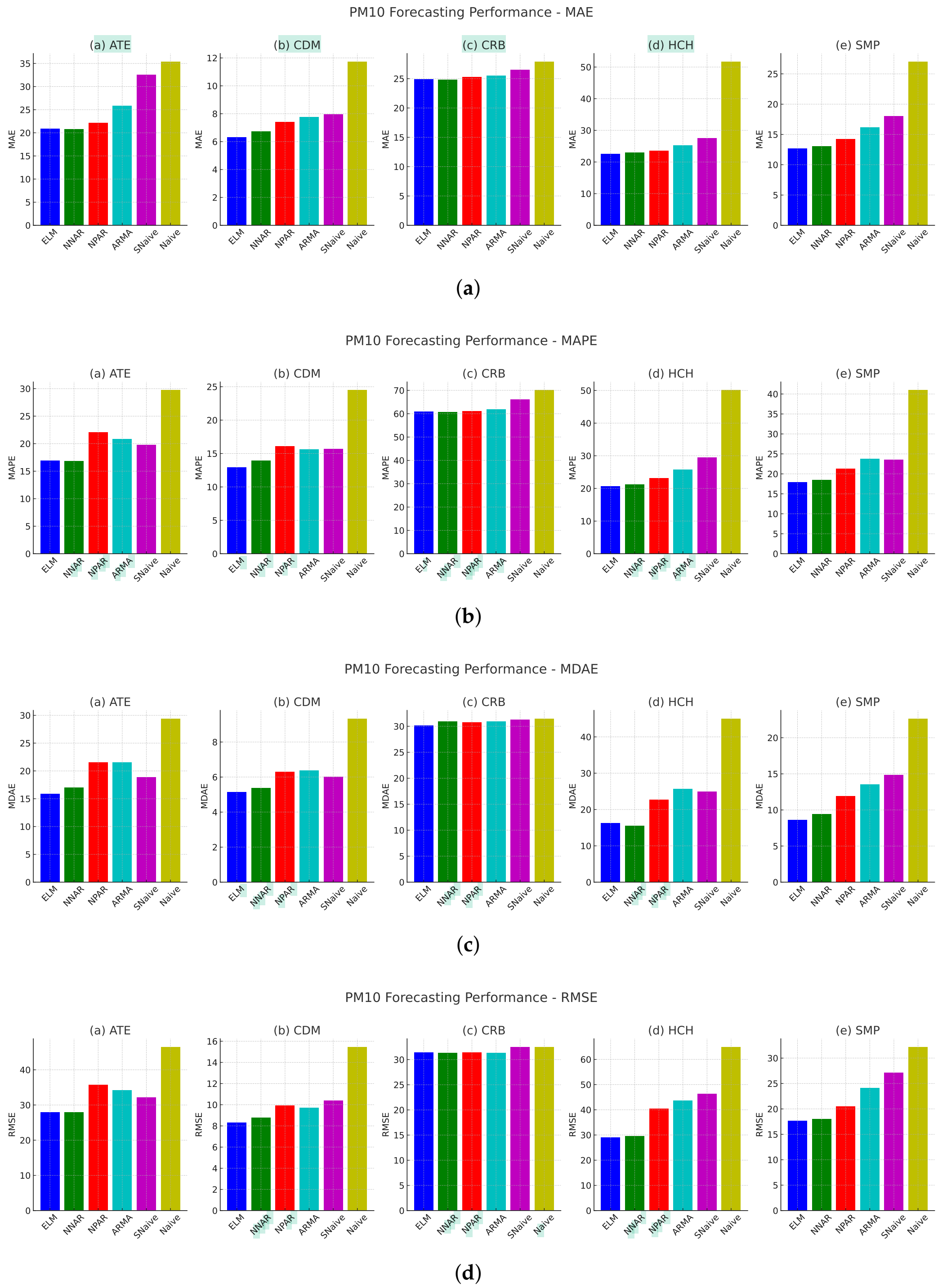

Figure 3. Performance metrics plots: The MAE, MAPE, MDAE, and RMSE, with subheadings for different monitoring sites.

The findings indicate that machine learning-based methodologies (ELM and NNAR) consistently surpass statistical models (ARMA, SNaive, and Naive) in PM_{10} forecasting across all evaluation metrics and monitoring sites. These models yield lower error values, making them more desirable for forecasting air pollution. Nonetheless, specific monitoring locations, such as CRB and HCH, display higher overall error rates, which may point to potential forecasting difficulties arising from local environmental or meteorological factors.

These findings hold significant importance for public health officials and policymakers in Lima, Peru, for multiple reasons:

1. Enhanced Air Quality Monitoring and Prediction The capacity to precisely forecast PM_{10} concentrations enables authorities to anticipate pollution patterns and implement preventative strategies. Given that the ELM and NNAR models yield the lowest error rates across various monitoring sites, decision-makers can incorporate these machine learning methods for immediate air quality predictions.
2. Protection of Public Health High levels of PM_{10} exposure are associated with respiratory and cardiovascular health issues. By utilizing accurate forecasting models, health authorities can provide prompt alerts, guide vulnerable groups (such as children, the elderly, and individuals with pre-existing respiratory conditions), and initiate emergency responses when pollution levels are projected to increase.
3. Urban Development and Traffic Regulation Because PM_{10} pollution is frequently linked to vehicular and industrial emissions, officials can utilize forecasting information to create policies aimed at reducing pollution. This could involve limiting vehicle access in areas at high risk during times of elevated pollution or encouraging the use of public transport and green infrastructure.
4. Adherence to Regulations and Policy Implementation These forecasting models can assist policymakers in ensuring adherence to air quality regulations and evaluating the effectiveness of pollution mitigation strategies. If certain locations consistently show predictions of high pollution levels, stricter rules concerning industrial emissions and urban development may be warranted.
5. Increasing Public Awareness and Promoting Behavior Change Reliable PM_{10} forecasts can educate the public about air quality status via mobile applications, websites, or news broadcasts. This empowers residents to take safety measures, such as wearing masks, limiting outdoor activities, or utilizing air purification systems, ultimately decreasing health hazards.
6. Climate Change and Environmental Policy Development The insights derived from these models can support wider environmental initiatives focused on reducing emissions and addressing climate change. By pinpointing pollution hotspots and their sources, authorities can pursue sustainable urban planning efforts and long-term strategies for improving air quality.

In summary, these forecasting findings provide a scientific basis for evidence-based decision-making, enabling Lima's policymakers to implement targeted strategies that foster public health, enhance environmental conditions, and promote sustainable urban growth.

4. Conclusions

This study introduces an innovative forecasting approach to predict hourly (PM_{10}) levels in Metropolitan Lima, Peru. It employs multiple linear regression, artificial intelligence, and time series models, analyzing data from four districts (ATE, CDM, CRB, HCH, and SMP) from January 1, 2017, to December 31, 2018. Initially, we decomposed the hourly (PM_{10}) time series into two distinct components: deterministic and stochastic. The deterministic component includes a long-term linear trend and yearly, weekly, and hourly seasonal components, whereas the stochastic component consists solely of the remainder. To estimate the deterministic portion, we applied a multiple linear regression model. In contrast, for the stochastic portion, we utilized four forecasting models: two artificial intelligence models, specifically the extreme learning model and the artificial neural network autoregressive model, alongside two standard time series models, namely nonparametric autoregressive and autoregressive moving average models. We calculated four different accuracy metrics to evaluate the effectiveness of the proposed component-based modeling and forecasting approach: mean absolute error, mean absolute percentage error, mean absolute deviation, and mean deviation absolute error, in addition to a statistical assessment using the bar plots. The findings indicate that machine learning-based methodologies (ELM and NNAR) consistently surpass statistical models (ARMA, SNaive, and Naive) in PM_{10} forecasting across all evaluation metrics and monitoring sites. These models yield lower error values, making them more desirable for forecasting air pollution. Nonetheless, specific monitoring locations, such as CRB and HCH, display higher overall error rates, which may point to potential forecasting difficulties arising from local environmental or meteorological factors.

Nonetheless, a significant limitation of this research is its reliance solely on hourly (PM_{10}) level data. Future improvements could involve integrating additional external variables, such as wind speed, temperature, wind direction, and humidity, to refine short-term predictions. Moreover, the analysis was confined to datasets from four districts in Lima, Peru. Still, it could be broadened to include other districts within Lima, additional regions in Peru, or even a global perspective to evaluate the effectiveness of the suggested component-based time series modeling and forecasting approach. Furthermore, the study was limited to univariate time series models, which might be enhanced by incorporating machine learning techniques such as deep learning and artificial neural networks. These advanced models could also be merged into the existing time series forecasting framework.

Author Contributions:

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement:

Acknowledgments:

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Moretti-Villegas, L.F.; Tafur-Anzualdo, V.I.; Valiente-Saldaña, Y.M. Air pollution in the city of Lima, Perú. *Revista Arbitrada Interdisciplinaria Koinonía* **2023**, *8*, 822–830.
2. Huang, G.B.; Wang, D.H.; Lan, Y. Extreme learning machines: A survey. *International Journal of Machine Learning and Cybernetics* **2011**, *2*, 107–122. https://doi.org/CrossRef.
3. Iftikhar, H.; Khan, M.; Turpo-Chaparro, J.E.; Rodrigues, P.C.; López-Gonzales, J.L. Forecasting stock prices using a novel filtering-combination technique: Application to the Pakistan stock exchange. *AIMS Math* **2024**, *9*, 3264–3288. https://doi.org/CrossRef.
4. Gonzales, S.M.; Iftikhar, H.; López-Gonzales, J.L. Analysis and forecasting of electricity prices using an improved time series ensemble approach: An application to the Peruvian electricity market. *AIMS Mathematics* **2024**, *9*, 21952–21971. https://doi.org/CrossRef.
5. Shah, I.; Iftikhar, H.; Ali, S. Modeling and forecasting electricity demand and prices: A comparison of alternative approaches. *Journal of Mathematics* **2022**, *2022*, 3581037.
6. Quispe, F.; Salcedo, E.; Iftikhar, H.; Zafar, A.; Khan, M.; Turpo-Chaparro, J.E.; Rodrigues, P.C.; López-Gonzales, J.L. Multi-step ahead ozone level forecasting using a component-based technique: A case study in Lima, Peru. *AIMS Environmental Science* **2024**, *11*.