applied sciences

MDPI

Article

Short-Term Forecasting of Ozone Concentration in Metropolitan Lima Using Hybrid Combinations of Time Series Models

Natalí Carbo-Bustinza [1,*], Hasnain Iftikhar [2,3], Marisol Belmonte [4,5], Rita Jaqueline Cabello-Torres [6], Alex Rubén Huamán De La Cruz [7] and Javier Linkolk López-Gonzales [8,9,*]

1 Doctorado Interdisciplinario en Ciencias Ambientales, Universidad de Playa Ancha, Valparaíso 2340000, Chile
2 Department of Mathematics, City University of Science and Information Technology Peshawar, Peshawar 25000, Pakistan; hasnainchill3@gmail.com
3 Department of Statistics, Quaid-i-Azam University, Islamabad 45320, Pakistan
4 Laboratorio de Biotecnología, Medio Ambiente e Ingeniería (LABMAI), Facultad de Ingeniería, Universidad de Playa Ancha, Avda. Leopoldo Carvallo 270, Valparaíso 2340000, Chile; marisol.belmonte@upla.cl
5 HUB-Ambiental, Universidad de Playa Ancha, Avda. Leopoldo Carvallo 270, Valparaíso 2340000, Chile
6 Escuela de Ingeniería Ambiental, Universidad César Vallejo, Lima 15314, Peru; rcabello@ucv.edu.pe
7 E.P. de Ingenieria Ambiental, Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa, La Merced 15106, Peru; alebut2@hotmail.com
8 Vicerrectorado de Investigación, Universidad Privada Norbert Wiener, Lima 15046, Peru
9 Escuela de Posgrado, Universidad Peruana Unión, Lima 15468, Peru
* Correspondence: natali.carbo@alumnos.upla.cl (N.C.-B.); javierlinkolk@gmail.com (J.L.L.-G.)

Abstract: In the modern era, air pollution is one of the most harmful environmental issues on the local, regional, and global stages. Its negative impacts go far beyond ecosystems and the economy, harming human health and environmental sustainability. Given these facts, efficient and accurate modeling and forecasting for the concentration of ozone are vital. Thus, this study explores an in-depth analysis of forecasting the concentration of ozone by comparing many hybrid combinations of time series models. To this end, in the first phase, the hourly ozone time series is decomposed into three new sub-series, including the long-term trend, the seasonal trend, and the stochastic series, by applying the seasonal trend decomposition method. In the second phase, we forecast every sub-series with three popular time series models and all their combinations In the final phase, the results of each sub-series forecast are combined to achieve the results of the final forecast. The proposed hybrid time series forecasting models were applied to four Metropolitan Lima monitoring stations—ATE, Campo de Marte, San Borja, and Santa Anita—for the years 2017, 2018, and 2019 in the winter season. Thus, the combinations of the considered time series models generated 27 combinations for each sampling station. They demonstrated significant forecasts of the sample based on highly accurate and efficient descriptive, statistical, and graphic analysis tests, as a lower mean error occurred in the optimized forecast models compared to baseline models. The most effective hybrid models for the ATE, Campo de Marte, San Borja, and Santa Anita stations were identified based on their superior out-of-sample forecast results, as measured by RMSE (4.611, 3.637, 1.495, and 1.969), RMSPE (4.464, 11.846, 1.864, and 15.924), MAE (1.711, 2.356, 1.078, and 1.462), and MAPE (14.862, 20.441, 7.668, and 76.261) errors. These models significantly outperformed other models due to their lower error values. In addition, the best models are statistically significant ($p < 0.05$) and superior to the rest of the combination models. Furthermore, the final proposed models show significant performance with the least mean error, which is comparatively better than the considered baseline models. Finally, the authors also recommend using the proposed hybrid time series combination forecasting models to predict ozone concentrations in other districts of Lima and other parts of Peru.

Keywords: short-term ozone concentration forecasting; seasonal trend decomposition method; time series models; hybrid models

Citation: Carbo-Bustinza, N.; Iftikhar, H.; Belmonte, M.; Cabello-Torres, R.J.; De La Cruz, A.R.H.; López-Gonzales, J.L. Short-Term Forecasting of Ozone Concentration in Metropolitan Lima Using Hybrid Combinations of Time Series Models. *Appl. Sci.* **2023**, *13*, 10514. https://doi.org/10.3390/app131810514

Academic Editor: Dibyendu Sarkar

Received: 8 August 2023
Revised: 12 September 2023
Accepted: 12 September 2023
Published: 21 September 2023

1. Introduction

The stratosphere is the atmospheric layer characterized by the significant presence of ozone (O_3), which benefits all kinds of life on the planet due to the filtering process of solar ultraviolet radiation that occurs in the environment. Its presence in the biosphere is harmful to the health of all living beings and the environment because ozone is not only a greenhouse gas but also a powerful oxidant that contributes to global warming [1]. In addition, the impact caused by this atmospheric pollutant on crop production is known [2].

Recently, the atmospheric levels of ozone in the air have increased, affecting more and more people, especially in the cardiovascular system, causing inflammation, oxidative stress, and imbalances that have been related to mortality and morbidity [3]. It is important to develop stricter controls on O_3 precursors to mitigate the increased risks of ozone pollution episodes [4]. Tropospheric ozone monitoring represents a practical tool to analyze spatiotemporal trends in the behavior of this polluting agent in the air [5]. Thus, accurately forecasting ozone concentration is crucial to safeguarding vulnerable individuals, such as children, the elderly, and outdoor workers, from air pollution during hazardous periods of the day. Ground-level ozone concentrations are of significant concern due to their toxic agents, which can adversely affect the respiratory systems of people who inhale high ozone concentrations for extended periods. These adverse health effects can lead to decreased lung function, chest pain while breathing, coughing, throat infections, congestion, and worsening symptoms of asthma.

Time series record the observations made in a particular place and are associated with the evolution over time of a particular variable; observed behavior cannot be replicated with repeated experiments, and observations are often time-dependent. This information has allowed the development of traditional deterministic modeling and statistical models. Although chemical transport models have generally been applied to differentiate emission sources and meteorological variables to explain short- or long-term ozone fluctuations, temporal analysis can show spatial and seasonal changes in the distribution of ozone concentrations [6]. At the same time, statistical models are generated by relational analysis between factors influencing pollutants, producing powerful statistical prediction equations [7]. However, when you want to study the behavior in the spatiotemporal distribution of a pollutant, the problem lies in the variability of pollutant concentrations, which are strongly influenced by the fluctuations of the emitting sources and the meteorological state—hourly, daily, seasonal, and annual. Thus, the impact exerted by trends in the behavior of air pollutants may be beneficial to optimize the performance of modeling [7].

Currently, statistical modeling is evolving, including the management of time series that deserves to be compared with traditional models, mainly multiple linear regression. However, it is still necessary to continue exploring new studies to improve the prediction of reliable models, the reduction of noise through filters, and the organization of the numerical information of contaminants [7]. Decomposition is a methodology applied to analyze time series air pollutant data; the decomposition in ensemble empirical mode is counted to process these non-stationary and nonlinear signals and allows one to gradually separate the different fluctuation components [8]. Generally, the numerical data of the ozone time series have various types of patterns, so it is essential to break the database into several components or sub-classes in such a way that each one is a unique pattern of the data. Furthermore, the time series for an air pollutant is considered to be additive and may comprise elements over time [6]. Interrupted time series designs are a powerful tool for comparing the variation of levels and the trend of results [9].

According to Din [6], the ozone concentration at time t is given by the sum of each component (from decomposition). One component is given by the "trend" of time in the time series and is relevant to the persistent decrease or increase in ozone concentration driven via emission sources or meteorological variables. For its part, the second component, "seasonality", describes the fluctuations of the periodic seasons (decomposed), and the third, fixed by a short-term component, shows the "rest" of the random data once the first two components have been separated. In addition, other combined decomposition

methods or structures have been proposed in series and time convolution and long-term short-memory bidirectional networks [10]. Other models use a non-parametric Theil–Sen estimator as a robust Kendall [11] line-fit method or locally estimated scatterplot models for smoothing to filter the data obtained and subsequently decompose the time series models into trend, seasonal, and residual components of data and then recombine them appropriately [12]. After the decomposition of the time series into components or subseries, the data can be used in standardized time series modeling as linear, nonlinear autoregressive, or autoregressive moving averages. Linear models such as autoregressive are difficult to handle with nonlinear and time-varying data [13]. However, the application of combined auto-correlation function (ACF) and partial autocorrelation function (PACF) graphs overcomes the limitations of simple techniques by showing the correlation between the time series and the lags after excluding the contributions from previous lags [14]. Iftikhar et al. [15,16] applied a nonlinear autoregressive model relating a past value and smoothed functions of the original values of the time series. An autoregressive moving average was also applied to take into account the errors that make up the model, as well as linear models of combination for all lag observations and the lag error term. On the other hand, machine learning models have also been used to forecast ozone levels.For instance, the researchers in [17] proposed a deep learning model for the prediction of ozone levels in Aarhus using a grid search technique and implemented it as an accuracy tool for forecasting ozone levels in smart cities. The ozone concentration in India is predicted [18] using eight machine learning models, including XGboost, random forest, k-nearest neighbor, support vector regression, decision tree, Adaboost, linear regression, and bidirectional long-short-term memory, which achieved the predictive capabilities with a R^2 of 0.75 in winter. The researchers further divided the predicted capabilities in terms of season, and the winter season was found to be more predictable with 97.3%, post-monsoon 92.8%, monsoon 90.3%, and summer 88.9%. The authors in [19–21] applied time series, hybrid decomposition, machine learning, and deep learning models for forecasting ozone concentration in Tehran, Iran, in 11 municipal districts of Nanjing, China, and 8 out of 35 stations in Turkey.

Peru is a country located in South America in the Southeast Pacific Region, and its capital, Lima, is no stranger to ozone air pollution. Lima has become a megacity with more than 10 million inhabitants and severe air pollution problems. Romero et al. [22] evaluated the impact of meteorological variables on the ozone concentration and other pollutants present in the air through linear correlations made for data obtained between 2015 and 2018 at eight sampling stations in metropolitan Lima and reported that this pollutant increased with solar irradiation around 10:00 and 16:00 h, especially in spring, possibly caused by the interaction of primary NOx and hydrocarbon emissions from vehicle engines. Carbo-Bustinza et al. [23] instead studied the behavior of ozone in winter using machine learning algorithms in four stations in the city of Lima and found the highest critical levels (165.80 $\mu g/m^3$) in the Ate district (ATE). However, we observed, in general, a drop in values in the cold season ($O_3 < 100\ \mu g/m^3$), similar to another study [24]. At the same time, there is a need to comprehensively analyze the time series of the most polluted districts to optimize the prediction of ozone concentration. In this context, this research aims to propose an improved tool to forecast tropospheric ozone concentration using hybrid combinations of time series models in four districts of the megacity of Metropolitan Lima in a very precise way, through an innovative methodology based on the decomposition of a time series of data and the combination of traditional methods to achieve efficient predictions. The following are the contributions of this research:

- We improve the efficiency and accuracy of one-hour-ahead ozone concentration forecasting using a proposed hybrid combination of time series models based on the seasonal trend decomposition technique and various standard time series models.
- We apply the seasonal trend decomposition method of the ozone concentration database in four districts—ATE, Campo de Marte (CDM), San Borja (SB), and Santa Anita (STA)—with severe episodes of ozone contamination between 2017 and 2019.

- We evaluate the performance of the proposed hybrid combination of time series models, by determining five different accuracy mean errors: two relative mean errors, two absolute mean errors, and one correlation measure, such as root mean square error, root mean square percentage error, mean absolute error, and mean absolute percentage error; a statistical test, the Diebold–Mariano test; and a visual evaluation.
- In this study, the results of the final best combination model are compared with the best model proposed in the literature as well as the considered baseline models and the comparative results are recorded. Based on these results, the proposed final best combination model from this work is highly accurate and efficient compared to the best models reported in the literature.
- We present a methodological proposal applicable to the environmental management system in order to mitigate ozone pollution aimed at the stakeholders of the national air quality program.
- Finally, the current work uses only the four district datasets in Lima, Peru. This can be extended to other districts of Lima, other regions of Peru, and even the world level to evaluate the performance of the proposed hybrid time series modeling and forecasting technique.

This article describes the proposed hybrid time series forecasting methodology and explains its construction step by step in Section 2. The results of the case study for each district studied are in Section 3. Discussion about the best combination model of this study versus the standard time series models is detailed in Section 4, and the conclusions, along with limitations and future challenges, are presented in Section 5.

2. The Proposed Hybrid Time Series Forecasting Methodology

Before starting the modeling, it often makes sense to prepare the data. The goal of preprocessing is usually to simplify the modeling of the data. To do this, the database is sorted, classified, and analyzed for each monitoring station, taking into account the winter period of the city, which runs from 21 June to 22 September, for ozone. From 2017 to 2019, four monitoring stations located at strategic points in the capital of Lima were considered. It should be noted that the number of monitoring stations in the capital of Lima is ten; however, four were selected due to a lack of data in the registry. The hourly ozone concentrations were measured with a Teledyne analyzer (an instrument with about 15 sensing technologies used in the monitoring and manufacturing of gas, liquid analysis, and medical fields). Analyzer operations include zero and span testing, calibration, and leak detection. Data are transmitted by telemetry to SENAMHI (National Meteorology and Hydrology Service of Peru) for validation after correcting zeros, duplicates, and/or anomalies. Similarly, SENAMHI has a systematic network of stations that normally and automatically monitor and report the variables studied to a processing center. These stations use high-quality instruments and sensors to measure temperature, relative humidity, wind speed, and direction on an hourly scale. In addition, an inductive algorithm called Multiple Imputation by Chained Equations was applied. This algorithm is based on a fully conditional specification, where each incomplete variable is specified by a separate model [25]. This performs multiple assignments to replace missing values in a dataset, in this case, for hourly rate records details (see Table 1).

Table 1. This table is based on 6768 observations taken throughout the winter season encompassing three years (2017, 2018, and 2019). It includes the percentage of imputation for each monitoring site.

Station	ATE	CDM	SB	STA
Total hours	6768	6768	6768	6768
Available hours	6654	6634	6614	6613
Imputed hours	114	134	154	155
Imputed%	1.68%	1.98%	2.27%	2.29%

Note: Campo de Marte (CDM), San Borja (SB), and Santa Anita (STA).

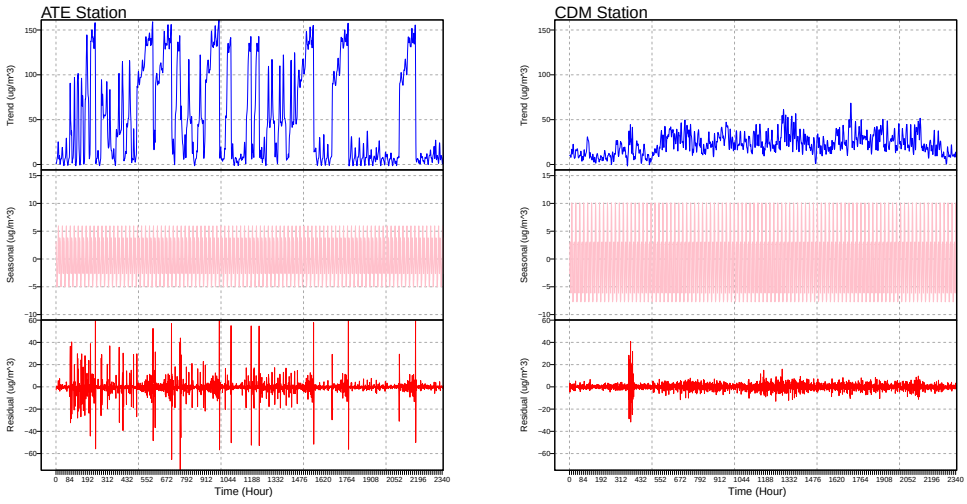
After obtaining the imputed ozone time series (free from missing values), we then proceed with the imputed ozone series and achieve a one-hour-ahead ozone concentration using the proposed hybrid combination of time series models. As explained previously, the hourly time series of ozone contains specific properties, such as a nonlinear long-run trend, an hourly cycle, and a different mean and variance. Considering these particular features in the model improves forecast accuracy significantly. To get these results, the ozone concentration in time series $(\mathfrak{C}_n)$ is divided into three new sub-series: the first is a long-run trend $(\mathfrak{l}_n)$, the second is a seasonal series $(\mathfrak{h}_n)$, and the third is a residual $(\mathfrak{r}_n)$ series. The mathematical description of the decomposed subsequence is given by

$$\mathfrak{C}_n = \mathfrak{l}_n + \mathfrak{h}_n + \mathfrak{r}_n \tag{1}$$

however, these sub-series are obtained using the seasonal trend decomposition method described in the following subsection.

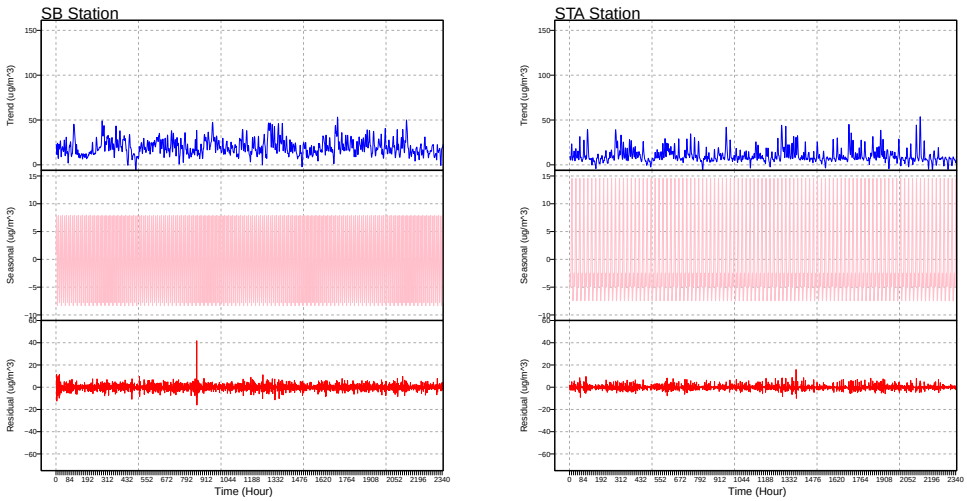
2.1. Seasonal Trend Decomposition Method

Cleveland et al. [26] proposed the decomposition technique where a seasonal time series model is split into three components of trend, seasonal, and stochastic. Seasonal trend decomposition (STLD) uses losses to decompose the seasonal component of a time series into other three components, including seasonal, trend, and stochastic. In particular, the steps included in STLD are: first de-trending; second cyclic smoothing of a sub-sequence, which creates the sequence of each seasonal component and smooths them individually; third, the regular sub-string is smoothed by a low-pass filter, which recombines and smooths sub-strings; fourth, we clean up the season series; fifth, the seasonal component computed in the previous step is used to de-trend the original series, and sixth, the seasonal sequence smoothing is used to get the trend component. To graphically explore the performance of the STLD method described above, the decomposed sub-series are shown in Figure 1. In each sub-figure (a to d) over a year (only winter season), the top panel indicates the long-term trend $(\mathfrak{l}_n)$, the seasonal component is shown in the middle panel $(\mathfrak{h}_n)$, and the residual component is presented in the bottom panel $(\mathfrak{r}_n)$. Hence, the STLD technique was applied to decompose $(\mathfrak{C}_n)$ to properly extract the long-term trend and hourly cycle in the ozone concentration time series. Moreover, the considered decomposition method extracts the specific features in all four station ozone concentration time series very well.



(**a**) (**b**)

Figure 1. *Cont.*



(**c**) (**d**)

Figure 1. Ozone concentration in the metropolitan area of Lima ($\mu g/m^3$): the hourly ozone concentration of the decomposed time series by the STLD method; ATE (**a**), Campo de Marte (**b**), San Borja (**c**), and Santa Anita (**d**), in each sub-figure, the top panel shows the long-run trend ($\mathfrak{l}_n$), the middle shows the seasonal ($\mathfrak{h}_n$) component, and the bottom shows the residual ($\mathfrak{r}_n$) component over a year.

2.2. Modeling the Decomposed Sub-Series

Once the sub-series are obtained from the hourly ozone concentration time series using the STL decomposition technique, the extracted sub-series are fit by applying the three considered standard time series models, including linear autoregressive (AR), nonlinear autoregressive (NLAR), and autoregressive moving averages (ARMA) [27,28]. These three models are explained in the following subsections.

2.2.1. Autoregressive Model

The autoregressive model (AR) model uses a linear combination of x lagged observations of $\mathfrak{C}_n$ to explain the short-term dynamics of $\mathfrak{C}_n$ [29] and can be expressed as

$$\mathfrak{C}_n = I + \xi_1 \mathfrak{C}_{n-1} + \xi_2 \mathfrak{C}_{n-2} + ... + \xi_x \mathfrak{C}_{n-x} + \epsilon_n, \quad (2)$$

where ξ_i $(i = 1, 2, \ldots, r)$ are the parameters of AR model and ϵ_m denotes the white noise process. In the present study, the maximum likelihood method is used for parameter estimation. The lags 1, 2, 3, 4, and 5 were included in the model due to their significant results after the plotting of autocorrelation function (ACF) and partial autocorrelation function (PACF) for the series.

2.2.2. Nonlinear Autoregressive Model

The nonlinear autoregressive model (NLAR) is the additive counterpart of the AR model, in which there is no specific linear form between $\mathfrak{z}_n$ and its corresponding lag values [30]. Mathematically, it can be expressed as

$$\mathfrak{C}_n = w_1(\mathfrak{C}_{n-1}) + w_2(\mathfrak{C}_{n-2}) + \ldots + w_x(\mathfrak{C}_{n-x}) + \epsilon_n, \quad (3)$$

where w_i represents each lag value, and smoothing function $\mathfrak{C}_n$ expresses the relationship between $\mathfrak{C}_n$. In this study, the function w_i is described by a cubic regression spline, and lags 1, 2, 3, 4, and 5 are used for NLAR modeling.

2.2.3. Autoregressive Moving Average Model

The autoregressive moving average (ARMA) model includes both error terms and lagged values of the time series. In this work, the sub-series are modeled with a linear

combination of x lagged values and delayed error terms [31]. Mathematically, the model equation can be expressed as

$$\mathfrak{C}_{\mathfrak{n}} = \mu + \xi_1 \mathfrak{C}_{\mathfrak{n}-1} + \xi_2 \mathfrak{C}_{\mathfrak{n}-2} + \ldots + \xi_x \mathfrak{C}_{\mathfrak{n}-\mathfrak{x}} + \epsilon_n + \psi_1 \epsilon_{n-1} + \psi_2 \epsilon_{n-2} + \ldots + \varphi \epsilon_{n-s}, \quad (4)$$

where μ is the intercept, ξ_i $(i = 1, 2, \ldots, x)$ and ψ_j $(j = 1, 2, \ldots, s)$ are the parameters for the MA and AR models, respectively, and $\epsilon_n \sim N(0, \sigma_\epsilon^2)$. In this work, the descriptive and graphical analysis indicates that, in the MA part, the first two lags are significant, whereas in the AR part, only lags 1, 2, 3, 4, and 5 are significant.

In this research study, each combined model is denoted with the STLD method by ${}^{\mathfrak{l}_{\mathfrak{n}}}\mathrm{STLD}^{\mathfrak{h}_{\mathfrak{n}}}_{\mathfrak{r}_{\mathfrak{n}}}$, where the $\mathfrak{l}_{\mathfrak{n}}$ in the top left corner represent the long-run component/sub-series, the $\mathfrak{h}_{\mathfrak{n}}$ in the top right indicates the seasonal component/sub-series, and the residual component/sub-series is represented in the bottom right by $\mathfrak{r}_{\mathfrak{n}}$. In the forecasting models, we assign the codes "a", "b", and "c" to the autoregressive, the nonlinear autoregressive, and the autoregressive moving average models, respectively. For example, ${}^{\mathrm{a}}\mathrm{STLD}^{\mathrm{b}}_{\mathrm{c}}$ describes the estimate of the long-term trend ($\mathfrak{l}_{\mathfrak{n}}$) with AR model, the seasonal series ($\mathfrak{h}_{\mathfrak{n}}$) estimated with the NLAR model, and the residual series ($\mathfrak{r}_{\mathfrak{n}}$) estimated by using ARMA. The individual forecast models are combined to obtain the final one-hour-ahead forecasts of ozone concentration.

$$\hat{\mathfrak{C}}_{n+1} = \left(\hat{\mathfrak{l}}_{n+1} + \hat{\mathfrak{h}}_{n+1} + \hat{\mathfrak{r}}_{n+1}\right) \quad (5)$$

2.3. Accuracy Measures

In order to check the performance of the forecasting models in previous studies, many researchers used various performance measures and statistical tests [32–35]. Hence, in this study, for model evaluation, first, we used five accuracy mean errors: two relative mean errors, two absolute mean errors, and one correlation measure for observed versus forecasted values, such as root mean square error (RMSE), root mean square percentage error (RMSPE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The mathematical formula for accuracy means errors are expressed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathfrak{C}_{\mathfrak{i}} - \hat{\mathfrak{C}}_i)^2}, \quad (6)$$

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(\mathfrak{C}_{\mathfrak{i}} - \hat{\mathfrak{C}}_i)^2}{\mathfrak{C}_{\mathfrak{i}}}\right) \times 100}, \quad (7)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left(|\mathfrak{C}_{\mathfrak{i}} - \hat{\mathfrak{C}}_i|\right), \quad (8)$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|\mathfrak{C}_{\mathfrak{i}} - \hat{\mathfrak{C}}_i|}{|\mathfrak{C}_{\mathfrak{i}}|}\right) \times 100, \quad (9)$$

$$\mathrm{CC} = \mathrm{correlation}\left(\mathfrak{C}_{\mathfrak{i}}, \hat{\mathfrak{C}}_i\right). \quad (10)$$

Here, the observed value is $\mathfrak{C}_{\mathfrak{i}}$ of the time series, and $\hat{\mathfrak{C}}_i$ represent the forecasted ozone concentration value of the ith observation (i = 1, 2, . . ., n), with the size of n in the testing set. Second, the Diebold and Mariano (DM) test [36] was conducted to test the significance of the differences among the performance of the forecasting models. The DM test is a broadly used statistical test for the comparison of forecasts extracted from various models [37–39]. To understand it, consider two forecasts, $\hat{\mathfrak{C}}_{1n}$ and $\hat{\mathfrak{C}}_{2n}$, that are available for the time series $\mathfrak{C}_n$ for $n = 1, \ldots, N$. The associated forecast errors are $\mathrm{e}_{1n} = \mathfrak{C}_n - \hat{\mathfrak{C}}_{1n}$ and $\mathrm{e}_{2n} = \mathfrak{C}_n - \hat{\mathfrak{C}}_{2n}$. Let the loss associated with forecast error $\{\mathrm{e}_{in}\}_{i=1}^{2}$ be $L(\mathrm{e}_{in})$. For example, the absolute loss in time n would be $L(\mathrm{e}_{in}) = |\mathrm{e}_{in}|$, and the differential loss between forecast 1 and forecast 2 for time t is then $\mathrm{w}_n = L(\mathrm{e}_{1n}) - L(\mathrm{e}_{2n})$. The null hypothesis of equal forecast accuracy

for two forecasts is $E[\mathrm{w}_n] = 0$. The DM test needs the differential loss to be covariance stationary, i.e.,

$$E[\mathrm{w}_n] = \mu, \quad \forall\, n \tag{11}$$

$$\mathrm{cov}(\mathrm{w}_n - \mathrm{w}_{n-\tau}) = \gamma(\tau), \quad \forall\, n \tag{12}$$

$$\mathrm{var}(\mathrm{w}_n) = \sigma_{\mathrm{w}}, \quad 0 < \sigma_{\mathrm{w}} < \infty \tag{13}$$

Under these assumptions, the DM test of equal forecast accuracy is

$$\mathrm{DM} = \frac{\bar{\mathrm{w}}}{\hat{\sigma}_{\bar{\mathrm{w}}}} \xrightarrow{d} N(0,1)$$

where $\bar{\mathrm{w}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{w}_n$ is the differential loss of the sample mean, and $\hat{\sigma}_{\bar{\mathrm{w}}}$ is a consistent estimate of standard error w_n. Finally, we verify the superiority of the proposed hybrid combination of time series forecasting models using various figures, such as the box plot, line plot, bar plot, and dot plot in this work. To conclude this section, the design of the proposed hybrid combination of time series modeling and forecasting technique is presented in Figure 2.

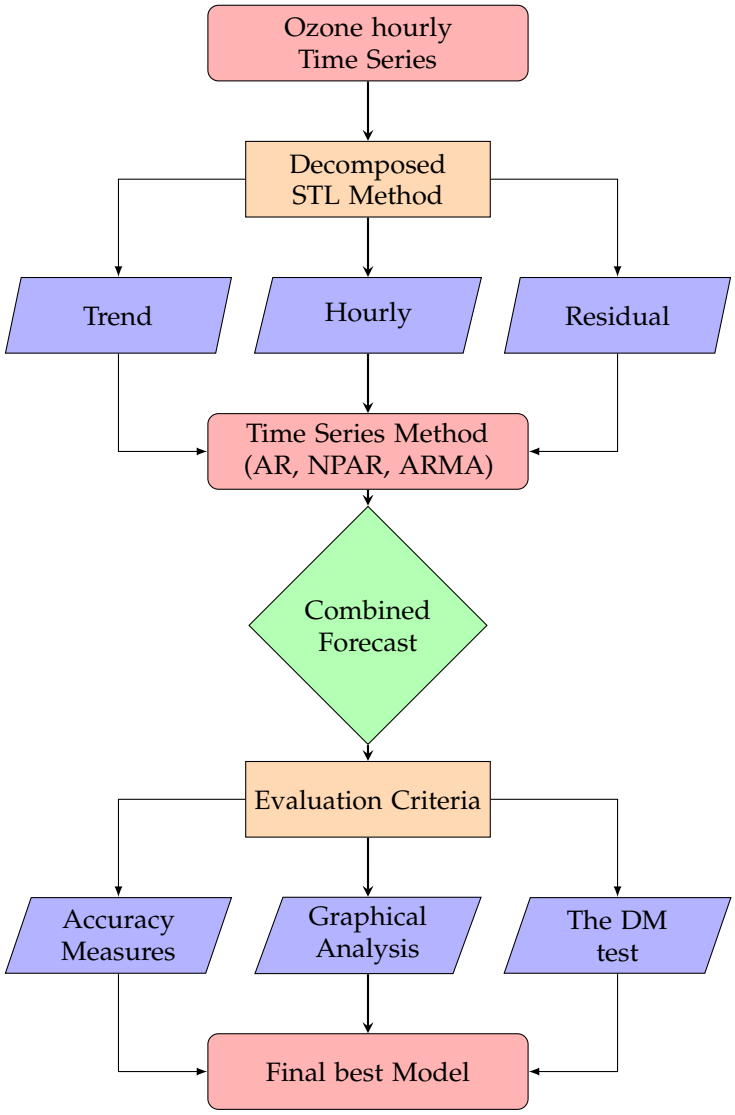


Figure 2. A flowchart of the proposed forecasting methodology.

3. Case Study Results

This work uses hourly ozone concentration datasets from four monitoring stations: ATE, CDM, SB, and Santa Anita, in Metropolitan Lima, for the duration of three consecutive years: 2017, 2018, and 2019. Within each year, only winter days are considered. Therefore, there are 6768 data points for one station. The graphic presentation of all four stations' hourly time series can be seen in Figure 3. The descriptive statistics and non-stationary statistics (augmented Dickey–Fuller (ADF) [40] test) for all four stations' imputed hourly ozone time series and the log imputed hourly ozone time series are listed in Table 2. Hence, descriptive metrics are a collection of methods for summarising and describing the key characteristics of a dataset, such as its central tendency, variability, and distribution. These statistics give an overview of the data and aid in determining the presence of patterns and linkages. It can be seen from Table 2 that the clear effect of the log and without log time series is in terms of all descriptive statistics, especially the variance and standard deviation stabilization. To conclude, the log-filtered series has the least descriptive statistic values. In addition to the above, we check the unit root issue for all four stations' imputed hourly ozone time series and the log imputed hourly ozone time series statistically by the ADF test. The results (statistic values), listed in Table 2, suggest that both the log-filtered imputed hourly ozone time series and the log-imputed hourly ozone time series have a higher negative statistic value, which indicates that the series is stationary. Therefore, once the database addresses all the essential treatments, we proceed further, and for forecasting and model estimation purposes, the data are divided into two parts: a training part (for model fit) and a testing part (for out-of-sample forecast). The training part contains the data for 5424 h, which is about 80% of the overall data, and 1344 h are used as the out-of-sample (testing).

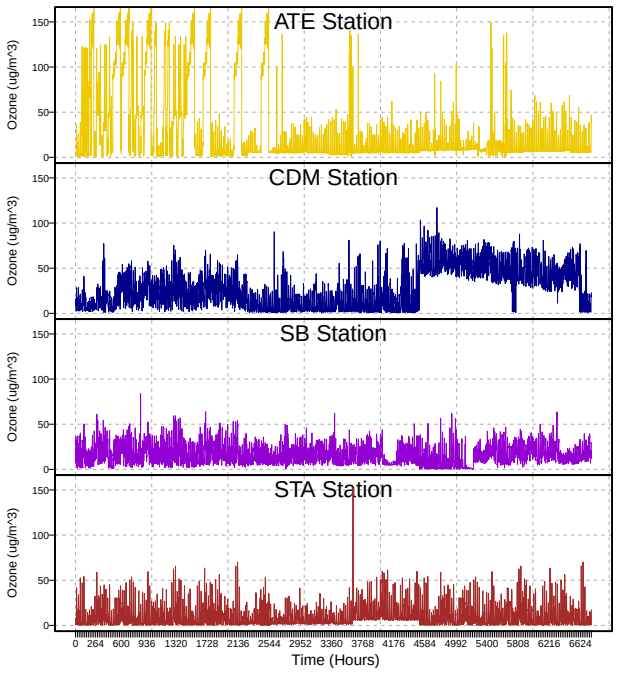


Figure 3. Ozone concentration in the metropolitan area of Lima ($\mu g/m^3$): the hourly ozone concentration time series for ATE (1st panel), Campo de Marte (2nd panel), San Borja (3rd panel), and Santa Anita (4th panel).

To obtain the forecast for ozone concentration one step ahead of an hour using the proposed hybrid methodology time series forecasting presented in Section 2, the given steps need to be followed: first, the STL method of decomposition was used to get a long-run trend ($\mathfrak{l}_n$), a seasonal ($\mathfrak{h}_n$), and the residual ($\mathfrak{r}_n$) of the time sub-series. Second, the previously explained three famous models of times series were used for each sub-series. Therefore,

the forecast of an hour ahead was obtained by using the rolling window technique for 1344 h and the models were estimated accordingly. Finally, the ozone concentration forecasts were achieved through Equation (5). The performance measures, including RMSE, RMSPE, MAE, MAPE, and CC, are then used for the evaluation and comparative performance of the models. Therefore, the following subsections detail the results from four monitoring stations: Ate, Campo de Marte, San Borja, and Santa Anita, all located in Metropolitan Lima.

Metropolitan Lima Stations

This subsection elaborates on the results and discussion about the Metropolitan Lima station. First, the hourly time series of the ATE, the CDM, the SB, and the SBA station's ozone concentration ($\mathfrak{C}_n$) are decomposed into a long-run trend ($\mathfrak{l}_n$), seasonal ($\mathfrak{h}_n$) and a residual sub-series ($\mathfrak{r}_n$); the STL decomposition method was implemented in this study. For obtaining the forecasts of the sub-series, three univariate time series models were used. Ensemble models for sub-series forecast of ($3^{\mathfrak{l}_n} \times 3^{\mathfrak{h}_n} \times 3^{\mathfrak{r}_n} = 27$) different combinations for all four considered monitoring stations were used. For these 27 different combination models, the performance measures (RMSE, RMSPE, MAE, MAPE, and CC) for one hour ahead of out-of-sample forecasts for the ATE, the CDM, the SB, and the SBA stations are listed in Table 3.

In the first attempt, the case study results of the ATE station accuracy performance measures (RMSE, RMSPE, MAE, MAPE, and CC) show that the $^{\mathbf{a}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ hybrid combination model produces the best forecasts compared to all other possible hybrid combinations of time series models. The $^{\mathbf{a}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ is the best forecasting model, which produced 4.611, 4.464, 1.711, 14.862, and 0.949 for RMSE, RMSPE, MAE, MAPE, and CC, respectively. However, the $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ (4.636, 4.480, 1.704, 14.985, 0.948), $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{b}}^{\mathbf{b}}$ (5.601, 4.817, 1.882, 16.179, 0.924), and $^{\mathbf{a}}\mathbf{STLD}_{\mathbf{b}}^{\mathbf{b}}$ (5.622, 4.906, 1.871, 16.081, 0.923) models produced the second, third, and fourth best results. Similarly, in the second attempt, the case study results of the CDM station and the results of the performance accuracy measures show that the $^{\mathbf{b}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{c}}$ model yields better forecasts compared to all other possible hybrid combination models. The best forecasting model, $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{c}}$, produced 3.637, 11.846, 2.356, 20.441, and 0.978 for RMSE, RMSPE, MAE, MAPE, and CC, respectively. However, the $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ (3.762, 11.689, 2.464, 20.847, 0.976), $^{\mathbf{a}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{c}}$ (3.746, 11.68, 2.458, 20.882, 0.976), and $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ (3.794, 11.906, 2.514, 21.323) models produced the second, third, and fourth best results. In the same way, in the third attempt, the case study results of the SB station and the results of the performance accuracy measures show that the $^{\mathbf{b}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ model yields better forecasts compared to all other possible combination models. The best forecasting model is $^{\mathbf{b}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$, which gives outcomes of 1.495, 1.864, 1.078, 7.668, and 0.989 for RMSE, RMSPE, MAE, MAPE, and CC, respectively. However, the $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ (1.559, 1.568, 1.136, 7.897, and 0.987), $^{\mathbf{a}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ (1.535, 1.644, 1.118, 7.793, and 0.987), and $^{\mathbf{b}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{c}}$ (1.721, 2.021, 1.301, 9.293, and 0.985) models produced the second, third, and fourth best results. Finally, in the fourth attempt, the case study results of the SB station and the results of the performance accuracy measures show that the $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$ model yields better forecasts compared to all other possible combination models. The best forecasting model is $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{b}}$, which gives outputs of 1.969, 15.924, 1.462, 76.261, and 0.989 for RMSE, RMSPE, MAE, MAPE, and CC, respectively. However, the $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{c}}$ (2.141, 19.925, 1.605, 88.958, and 0.988), $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{c}}^{\mathbf{a}}$ (2.143, 19.669, 1.603, 89.367, and 0.988), and $^{\mathbf{c}}\mathbf{STLD}_{\mathbf{b}}^{\mathbf{b}}$ (3.190, 21.490, 2.298, 95.063, and 0.972) models produced the second, third, and fourth best results.

Table 2. This table contains descriptive statistics for the time series of ozone concentration and the logarithmic time series of the ozone concentration for all considered monitoring stations.

Measure	Min	Q1	Median	Mean	Mode	Var	S.D	Skewness	Kurtosis	Q3	Max	ADF (Statistic)
ATE	0.80	5.50	8.50	28.36	5.20	1606.08	40.08	1.89	2.30	29.30	165.80	−8.61
log(ATE)	−0.22	1.70	2.14	2.55	1.65	1.50	1.23	0.49	−0.54	3.38	5.11	−8.70
CDM	0.80	8.98	24.50	28.13	1.00	454.02	21.31	0.53	−0.67	44.03	117.10	−6.03
log(CDM)	−0.22	2.19	3.20	2.86	0.00	1.37	1.17	−0.89	−0.19	3.78	4.76	−6.53
SB	0.20	8.30	15.10	17.09	6.50	122.05	11.05	0.83	0.51	24.00	83.90	−13.02
log(SB)	−1.61	2.12	2.71	2.56	1.87	0.78	0.88	−1.46	3.38	3.18	4.43	−10.35
STA	0.10	1.80	6.20	10.56	0.40	149.09	12.21	1.94	5.54	14.80	152.60	−16.17
log(STA)	−2.30	0.59	1.82	1.59	−0.92	2.02	1.42	−0.44	−0.61	2.69	5.03	−14.38

Table 3. Ozone concentration in four Metropolitan Lima stations ($\mu g/m^3$): out-of-sample one-hour ahead mean forecast error for all models combined with the STL decomposition method.

Station		ATE					Campo de Marte					San Borja					Santa Anita				
S.No	Models	RMSE	RMSPE	MAE	MAPE	CC	RMSE	RMSPE	MAE	MAPE	CC	RMSE	RMSPE	MAE	MAPE	CC	RMSE	RMSPE	MAE	MAPE	CC
1	$^{a}STLD^{a}_{a}$	5.529	5.414	2.209	20.827	0.932	5.073	16.53	3.329	25.735	0.957	2.115	2.600	1.587	11.217	0.975	5.279	40.464	3.958	196.406	0.916
2	$^{a}STLD^{a}_{b}$	5.699	4.828	2.076	18.005	0.921	5.145	16.406	3.354	24.908	0.955	2.081	2.417	1.547	10.817	0.976	5.338	40.070	3.959	188.568	0.913
3	$^{a}STLD^{a}_{c}$	4.675	4.628	1.913	18.257	0.947	3.993	12.117	2.719	22.96	0.973	1.818	1.854	1.376	9.768	0.982	3.958	33.264	2.965	158.543	0.954
4	$^{a}STLD^{b}_{a}$	5.410	4.976	1.950	17.562	0.937	4.889	16.474	3.136	25.196	0.96	1.974	2.497	1.448	10.279	0.979	5.224	36.786	3.878	179.823	0.917
5	$^{a}STLD^{b}_{b}$	5.622	4.906	1.871	16.081	0.923	4.921	16.336	3.108	24.118	0.959	1.973	2.333	1.439	10.174	0.979	5.310	36.972	3.907	172.380	0.914
6	$^{a}STLD^{b}_{c}$	4.611	4.464	1.711	14.862	0.949	3.774	11.89	2.504	21.293	0.976	1.535	1.644	1.118	7.793	0.987	3.909	30.357	2.937	148.290	0.955
7	$^{a}STLD^{c}_{a}$	5.529	5.414	2.209	20.827	0.932	4.717	16.474	2.848	26.253	0.963	2.115	2.600	1.587	11.217	0.975	5.277	40.431	3.959	197.407	0.916
8	$^{a}STLD^{c}_{b}$	5.699	4.828	2.076	18.005	0.921	4.685	16.271	2.764	24.623	0.963	2.081	2.417	1.547	10.817	0.976	5.337	40.071	3.963	190.219	0.913
9	$^{a}STLD^{c}_{c}$	4.675	4.628	1.913	18.258	0.947	3.746	11.68	2.458	20.882	0.976	1.818	1.854	1.376	9.767	0.982	3.958	33.365	2.970	159.531	0.954
10	$^{b}STLD^{a}_{a}$	5.607	5.313	2.277	21.015	0.933	5.485	16.957	3.697	26.817	0.949	2.213	2.872	1.664	11.793	0.974	5.319	41.263	3.977	199.487	0.915
11	$^{b}STLD^{a}_{b}$	5.730	4.845	2.067	17.830	0.922	5.579	16.845	3.776	26.481	0.947	2.231	2.711	1.680	11.819	0.973	5.375	40.769	3.979	191.434	0.912
12	$^{b}STLD^{a}_{c}$	4.709	4.683	2.033	19.601	0.947	4.187	12.464	2.984	24.15	0.971	1.721	2.021	1.301	9.293	0.985	3.991	33.742	2.979	160.368	0.953
13	$^{b}STLD^{b}_{a}$	5.509	4.895	2.047	17.770	0.937	5.247	16.859	3.458	26.089	0.954	2.132	2.803	1.597	11.384	0.976	5.257	37.440	3.894	182.865	0.916
14	$^{b}STLD^{b}_{b}$	5.672	4.951	1.909	16.143	0.924	5.305	16.733	3.507	25.446	0.952	2.182	2.661	1.643	11.671	0.975	5.339	37.506	3.927	175.424	0.913
15	$^{b}STLD^{b}_{c}$	4.669	4.552	1.893	16.799	0.949	3.886	12.183	2.687	22.163	0.975	1.495	1.864	1.078	7.668	0.989	3.933	30.607	2.941	148.480	0.954
16	$^{b}STLD^{c}_{a}$	5.607	5.313	2.277	21.015	0.933	4.921	16.764	2.979	26.353	0.959	2.213	2.872	1.664	11.793	0.974	5.317	41.231	3.978	200.433	0.915
17	$^{b}STLD^{c}_{b}$	5.730	4.845	2.067	17.830	0.922	4.921	16.575	2.995	25.119	0.959	2.231	2.711	1.680	11.820	0.973	5.374	40.771	3.984	193.141	0.912
18	$^{b}STLD^{c}_{c}$	4.709	4.683	2.033	19.601	0.947	3.637	11.846	2.356	20.441	0.978	1.721	2.021	1.301	9.293	0.985	3.991	33.843	2.985	161.576	0.953
19	$^{c}STLD^{a}_{a}$	5.545	5.581	2.197	20.797	0.932	5.092	16.544	3.34	25.732	0.956	2.124	2.506	1.606	11.322	0.975	3.267	25.624	2.460	125.732	0.973
20	$^{c}STLD^{a}_{b}$	5.678	4.753	2.081	17.964	0.922	5.166	16.42	3.363	24.898	0.955	2.075	2.316	1.554	10.821	0.976	3.289	25.472	2.435	117.338	0.971
21	$^{c}STLD^{a}_{c}$	4.700	4.659	1.900	18.266	0.946	4.013	12.134	2.73	22.965	0.973	1.858	1.807	1.421	10.163	0.981	2.143	19.669	1.603	89.367	0.988
22	$^{c}STLD^{b}_{a}$	5.427	5.143	1.940	17.535	0.936	4.909	16.487	3.146	25.199	0.96	1.965	2.384	1.441	10.090	0.979	3.125	20.593	2.277	99.420	0.975
23	$^{c}STLD^{b}_{b}$	5.601	4.817	1.882	16.179	0.924	4.942	16.35	3.118	24.128	0.959	1.947	2.212	1.415	9.869	0.979	3.190	21.490	2.298	95.063	0.972
24	$^{c}STLD^{b}_{c}$	4.636	4.480	1.704	14.985	0.948	3.794	11.906	2.514	21.323	0.976	1.559	1.568	1.136	7.897	0.987	1.969	15.924	1.462	76.261	0.989
25	$^{c}STLD^{c}_{a}$	5.545	5.581	2.197	20.797	0.932	4.734	16.481	2.855	26.204	0.962	2.124	2.506	1.606	11.322	0.975	3.262	25.637	2.460	126.306	0.973
26	$^{c}STLD^{c}_{b}$	5.678	4.753	2.081	17.963	0.922	4.704	16.28	2.771	24.576	0.963	2.075	2.316	1.554	10.821	0.976	3.286	25.540	2.432	116.842	0.971
27	$^{c}STLD^{c}_{c}$	4.700	4.659	1.900	18.267	0.946	3.762	11.689	2.464	20.847	0.976	1.858	1.807	1.421	10.163	0.981	2.141	19.925	1.605	88.958	0.988

From all twenty-seven models, in each monitoring station, the best four hybrid combination models are selected for comparison and compared with other models in each case. The outcome of all these best hybrid combination models is tabulated in Table 4. For example, in the case of the ATE station, based on the performance accuracy measure findings, it is evident that the ${}^{a}\mathbf{STLD}_{c}^{b}$ give the least values (RMSE = 4.611, RMSPE = 4.464, MAE = 1.711, MAPE = 14.862, and CC = 0.949). Therefore, it is concluded that the ${}^{a}\mathbf{STLD}_{c}^{b}$ is the best model among the best models as well as all twenty-seven models. In the same way, in the case of the CDM station, from Table 4, it is confirmed that the ${}^{b}\mathbf{STLD}_{c}^{c}$ produced the smallest values (RMSE = 3.637, RMSPE = 11.846, MAE = 2.356, MAPE = 20.441, and CC = 0.978). Hence, it is concluded that the ${}^{b}\mathbf{STLD}_{c}^{c}$ is the best model among the best models as well as all twenty-seven models. However, in the case of the SB station results, it is evident that the ${}^{c}\mathbf{STLD}_{c}^{b}$ produced the smallest values (RMSE=1.969, RMSPE = 15.924, MAE = 1.462, MAPE = 76.261, and CC = 0.989) within the final best hybrid combination models. Thus, it is concluded that the ${}^{c}\mathbf{STLD}_{c}^{b}$ is the best hybrid combination model among the best models as well as all twenty-seven models. Likewise, within the best hybrid combination model outcomes from the STA stations, the ${}^{b}\mathbf{STLD}_{c}^{b}$ produced the smallest values (RMSE = 1.495, RMSPE = 1.864, MAE = 1.078, MAPE = 7.668, and CC = 0.989). Based on these results, it is concluded that the ${}^{b}\mathbf{STLD}_{c}^{b}$ is the best model among the best models as well as all twenty-seven models.

Table 4. Ozone concentration in four Metropolitan Lima stations ($\mu g/m^3$): mean forecast error of one-hour-ahead post-sample for the best four models among all twenty-seven models.

ATE Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
${}^{a}\mathbf{STLD}_{c}^{b}$	4.611	4.464	1.711	14.862	0.949
${}^{c}\mathbf{STLD}_{c}^{b}$	4.636	4.480	1.704	14.985	0.948
${}^{c}\mathbf{STLD}_{b}^{b}$	5.601	4.817	1.882	16.179	0.924
${}^{a}\mathbf{STLD}_{b}^{b}$	5.622	4.906	1.871	16.081	0.923
Campo de Marte Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
${}^{b}\mathbf{STLD}_{c}^{c}$	3.637	11.846	2.356	20.441	0.978
${}^{c}\mathbf{STLD}_{c}^{c}$	3.762	11.689	2.464	20.847	0.976
${}^{a}\mathbf{STLD}_{c}^{c}$	3.746	11.68	2.458	20.882	0.976
${}^{c}\mathbf{STLD}_{c}^{b}$	3.794	11.906	2.514	21.323	0.976
San Borja Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
${}^{b}\mathbf{STLD}_{c}^{b}$	1.495	1.864	1.078	7.668	0.989
${}^{c}\mathbf{STLD}_{c}^{c}$	1.559	1.568	1.136	7.897	0.987
${}^{a}\mathbf{STLD}_{c}^{b}$	1.535	1.644	1.118	7.793	0.987
${}^{b}\mathbf{STLD}_{c}^{c}$	1.721	2.021	1.301	9.293	0.985
Santa Anita Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
${}^{c}\mathbf{STLD}_{c}^{b}$	1.969	15.924	1.462	76.261	0.989
${}^{c}\mathbf{STLD}_{c}^{c}$	2.141	19.925	1.605	88.958	0.988
${}^{c}\mathbf{STLD}_{c}^{a}$	2.143	19.669	1.603	89.367	0.988
${}^{c}\mathbf{STLD}_{b}^{b}$	3.190	21.490	2.298	95.063	0.972

To confirm the dominance of models for all monitoring stations (the ATE, the CDM, the SB, and the STA) listed in Table 4, in this work, we performed the DM test on each pair of models. The null hypothesis is that the two models on the columns and rows are equally accurate, and the alternative hypothesis is that the model on the columns is more accurate than the model on the rows (using the loss-squared function). The results (p-values) of the DM test are given in Table 5 for all four stations (ATE, CDM, SB, and STA) of Metropolitan Lima. The results of the ATE station show that the final best (${}^{a}\mathbf{STLD}_{c}^{b}$) model within all four best models is statistically superior to the other best combination models at the 5% level of significance. However, in the CDM, the SB, and the STA stations, the final best combination models, the (${}^{b}\mathbf{STLD}_{c}^{c}$), the (${}^{b}\mathbf{STLD}_{c}^{b}$), and (${}^{c}\mathbf{STLD}_{c}^{b}$), are statistically superior to the other best combination models at the 5% level of significance.

Once the proposed hybrid time series combination models' performance was evaluated by accuracy performance measures (RMSPE, RMSE, MAE, MAPE, and CC) and a statistical test (the DM test), we then processed the models for graphic analysis. For instance, a graphical representation of mean errors (RMSE, RMSPE, MAE, and MAPE) for all twenty-seven models is shown in Figure 4a for the ATE station, Figure 4b for the CDM station, Figure 4c for the SB station, and Figure 4d for the STA station. From Figure 4a–d, we can see that within all twenty-seven models, the $^{c}\mathbf{STLD}_{c}^{b}$ model in the ATE station, the $^{c}\mathbf{STLD}_{c}^{b}$ model in the CDM station, the $^{c}\mathbf{STLD}_{c}^{b}$ model in the SB station, the $^{c}\mathbf{STLD}_{c}^{b}$ model in the STA station produce the highest accuracy measures (RMSE, RMSPE, MAE, and MAPE) in comparison to the rest of all combination models. On the other hand, from all twenty-seven models in each monitoring station, the best four hybrid combination models are selected for comparison and compared with other models in each station. The results of all these best hybrid combination models are plotted in Figure 5. For example, see the ATE station in Figure 5a, the CDM station in Figure 5b, the SB station in Figure 5c, and the STA station in Figure 5d It can be observed from these plots that the $^{a}\mathbf{STLD}_{c}^{b}$, $^{a}\mathbf{STLD}_{c}^{b}$, and $^{a}\mathbf{STLD}_{c}^{b}$ show the least mean errors, respectively. In addition to the above, we plot the scatter diagrams for each station using their respective best model, which were obtained previously. For instance, Figure 6 displays the scatter plots for all considered monitoring stations. This figure showed that the best model produces greater correlation coefficient values, and it indicates that the correlation between forecast and actual ozone concentration values is highly significant. In the same way, the forecasted and observed values for the supermodel in each monitoring station are plotted in Figure 7. In Figure 7, forecasts of the best models follow the observed concentration of ozone very closely; from this, we can conclude that the supermodel in each considered station has accurate and efficient forecasts. Thus, from the descriptive statistical analysis, tests, and graphical results, we can conclude that the proposed hybrid combination of time series models is highly efficient and accurate in forecasting hourly ozone concentration.

Table 5. Ozone concentration in four Metropolitan Lima stations ($\mu g/m^3$): results (p-value) of the DM test for the best four models given in Table 4.

ATE Station				
Models	$^{a}\mathbf{STLD}_{c}^{b}$	$^{c}\mathbf{STLD}_{c}^{c}$	$^{c}\mathbf{STLD}_{c}^{b}$	$^{a}\mathbf{STLD}_{b}^{b}$
$^{a}STLD_{c}^{b}$	-	0.229	0.988	0.992
$^{c}STLD_{c}^{c}$	0.771	-	0.991	0.993
$^{c}STLD_{c}^{b}$	0.012	0.009	-	0.332
$^{a}STLD_{b}^{b}$	0.008	0.007	0.668	-
Campo de Marte Station				
Models	$^{b}\mathbf{STLD}_{c}^{c}$	$^{c}\mathbf{STLD}_{c}^{c}$	$^{a}\mathbf{STLD}_{c}^{c}$	$^{c}\mathbf{STLD}_{c}^{b}$
$^{b}\mathbf{STLD}_{c}^{c}$	-	0.965	0.944	0.963
$^{c}\mathbf{STLD}_{c}^{c}$	0.036	-	0.000	0.716
$^{a}\mathbf{STLD}_{c}^{c}$	0.056	1.000	-	0.806
$^{c}\mathbf{STLD}_{c}^{b}$	0.037	0.284	0.194	-
San Borja Station				
Models	$^{b}\mathbf{STLD}_{c}^{b}$	$^{c}\mathbf{STLD}_{c}^{b}$	$^{a}\mathbf{STLD}_{c}^{b}$	$^{b}\mathbf{STLD}_{c}^{c}$
$^{b}\mathbf{STLD}_{c}^{b}$	-	0.989	0.945	1.000
$^{c}\mathbf{STLD}_{c}^{b}$	0.011	-	0.005	1.000
$^{a}\mathbf{STLD}_{c}^{b}$	0.055	0.996	-	1.000
$^{b}\mathbf{STLD}_{c}^{c}$	0.000	0.000	0.000	-
Santa Anita Station				
Models	$^{b}\mathbf{STLD}_{c}^{b}$	$^{c}\mathbf{STLD}_{c}^{b}$	$^{a}\mathbf{STLD}_{c}^{b}$	$^{b}\mathbf{STLD}_{c}^{c}$
$^{b}\mathbf{STLD}_{c}^{b}$	-	1.000	1.000	1.000
$^{c}\mathbf{STLD}_{c}^{b}$	0.000	-	0.704	1.000
$^{a}\mathbf{STLD}_{c}^{b}$	0.000	0.296	-	1.000
$^{b}\mathbf{STLD}_{c}^{c}$	0.000	0.000	0.000	-

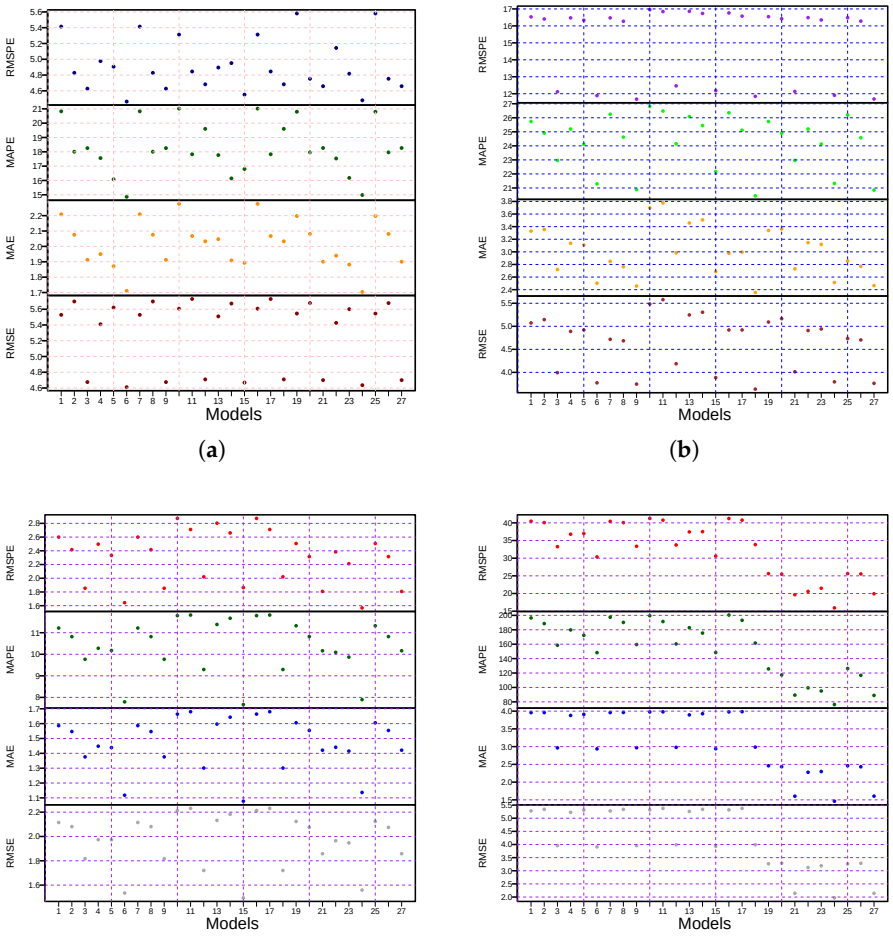


Figure 4. Ozone concentration ($\mu g/m^3$) in four Metropolitan Lima stations: (**a**) ATE, (**b**) Campo de Marte, (**c**) San Borja, and (**d**) Santa Anita; the RMSPE (1st panel), MAPE (2nd panel), MAE (3rd panel), and RMSE (4th panel) for all twenty-seven combination models using the proposed forecasting methodology.

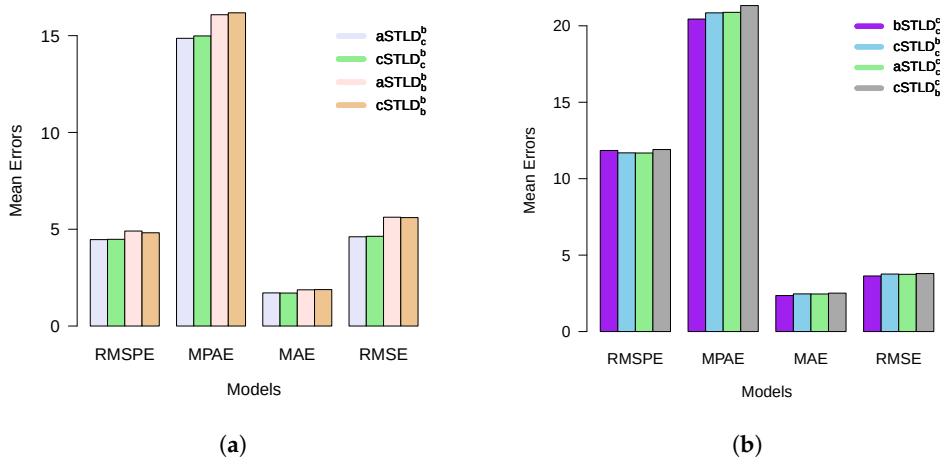
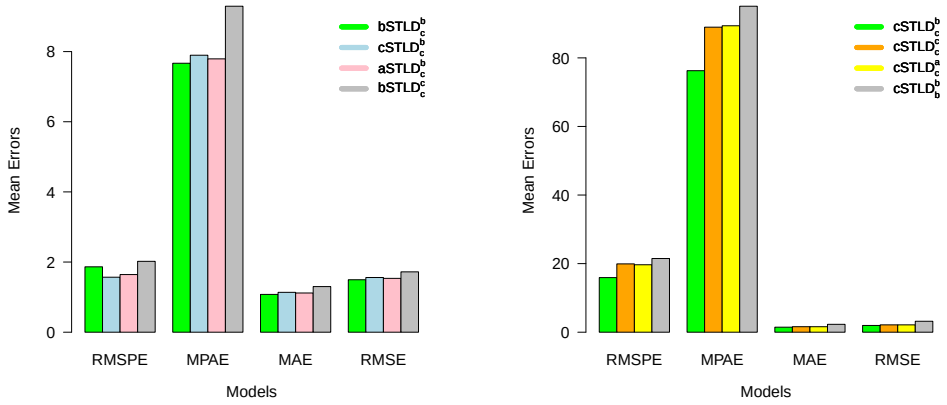


Figure 5. *Cont.*



(**c**) (**d**)

Figure 5. Ozone concentration in four Metropolitan Lima stations (µg/m^3): (**a**) Ate, (**b**) Campo de Marte, (**c**) San Borja, and (**d**) Santa Anita evaluation measures; the barplot for the best four models among all twenty-seven models.

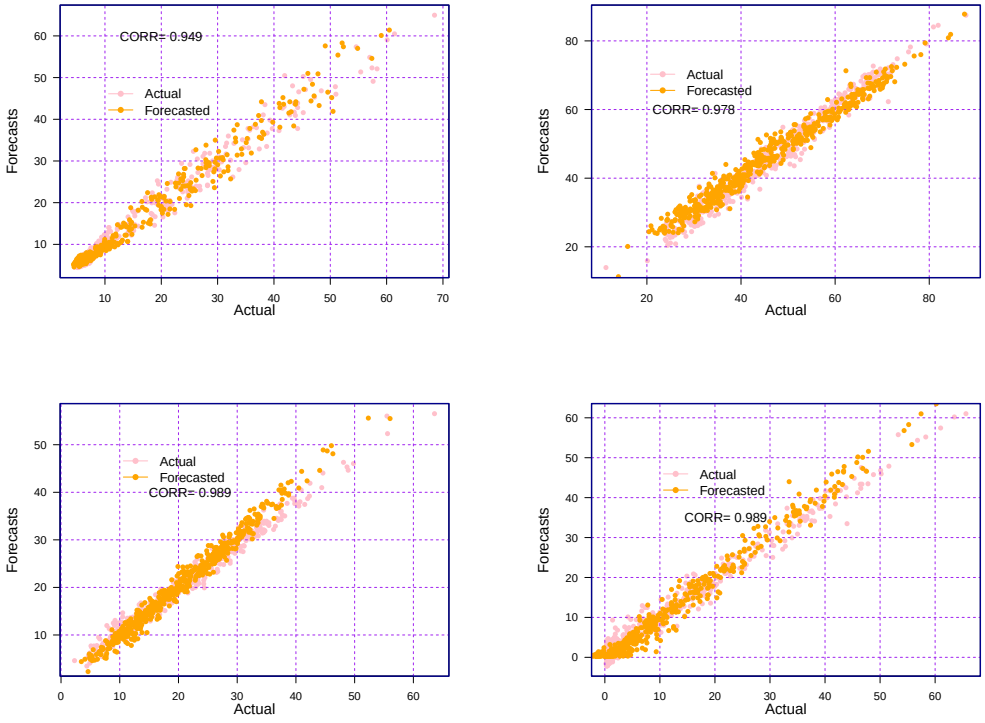


Figure 6. Correlation plots for the ozone concentration (µg/m^3) in all four Metropolitan Lima stations using their respective best hybrid models, including (**1st**) ATE ($^{\mathbf{a}}$**STLD**$_{\mathbf{c}}^{\mathbf{b}}$), (**2nd**) Campo de Marte ($^{\mathbf{b}}$**STLD**$_{\mathbf{c}}^{\mathbf{c}}$), (**3rd**), San Borja ($^{\mathbf{b}}$**STLD**$_{\mathbf{c}}^{\mathbf{b}}$), and (**4th**) Santa Anita ($^{\mathbf{c}}$**STLD**$_{\mathbf{c}}^{\mathbf{b}}$).

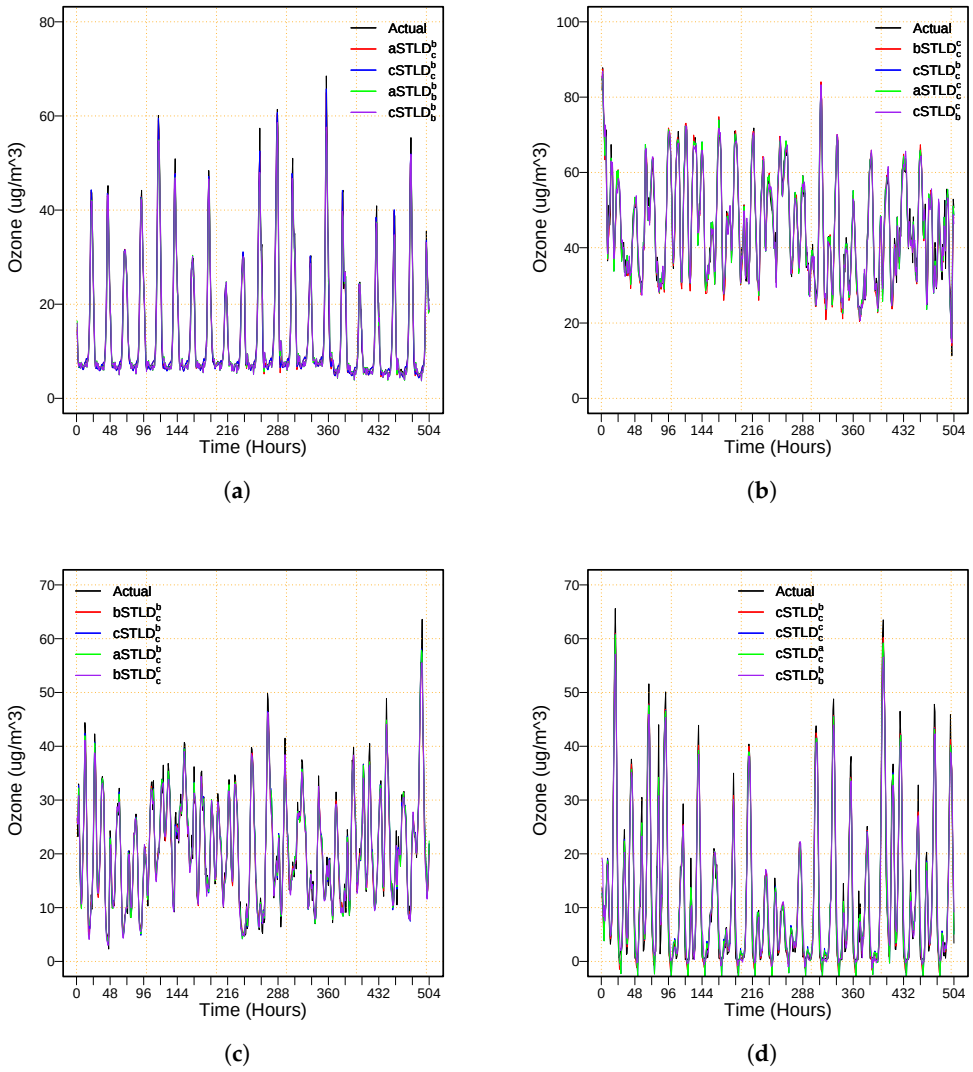


Figure 7. Ozone concentration in four Metropolitan Lima stations (µg/m^3): (**a**) Ate, (**b**) Campo de Marte, (**c**) San Borja, and (**d**) Santa Anita: actual and forecasted ozone concentration values for four of the best models over three weeks.

4. Discussion

Finally, according to the results (descriptive statistical analysis, tests, and visual analysis), it is concluded that the final best models for forecasting hourly ozone concentration were the $^{a}\mathbf{STLD}^{b}_{c}$, the $^{b}\mathbf{STLD}^{c}_{c}$, the $^{b}\mathbf{STLD}^{b}_{c}$, and the $^{c}\mathbf{STLD}^{b}_{c}$ for the ATE, the CDM, the SB, and the STA, respectively. However, to verify the superiority of these final best models, we compare them with some standard baseline time series models, including parametric autoregressive (PAR), nonparametric autoregressive (NPAR), and autoregressive integrated moving averages (ARIMA) models. For example, the comparative results are presented in Table 6 for all four monitoring stations. The results show that the considered baseline time series models are significantly outperformed by the best-proposed model in each station. In addition, to confirm the dominance of the best-proposed models given in Table 6 for each station, we performed a statistical DM test on each pair of models. The results (p-values) of the DM test are listed in Table 7, indicating that the baseline time series (PAR, NPAR, and ARIMA) models performed poorly in comparison to our best-proposed models in the considered stations at the 5% level of significance. To conclude, based on overall results, the performance measures of accuracy for the proposed methods of fore-

casting are comparatively better and more efficient than all other benchmark models in the competition.

Table 6. Ozone concentration in four Metropolitan Lima stations (μg/m^3): mean accuracy measures of the proposed versus the baseline models.

ATE Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
^{a}STLD$^{b}_{c}$	4.611	4.464	1.711	14.862	0.949
PAR	5.607	5.313	2.277	21.015	0.933
NPAR	5.730	4.845	2.067	17.830	0.922
ARIMA	4.709	4.683	2.033	19.601	0.947
Campo de Marte Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
^{b}STLD$^{c}_{c}$	3.637	11.846	2.356	20.441	0.978
PAR	5.485	16.957	3.697	26.817	0.949
NPAR	5.579	16.845	3.776	26.481	0.947
ARIMA	4.187	12.464	2.984	24.150	0.971
San Borja Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
^{b}STLD$^{b}_{c}$	1.495	1.864	1.078	7.668	0.989
PAR	2.213	2.872	1.664	11.793	0.974
NPAR	2.231	2.711	1.680	11.819	0.973
ARIMA	1.721	2.021	1.301	9.293	0.985
Santa Anita Station					
Models	**RMSE**	**RMSPE**	**MAE**	**MAPE**	**CC**
^{c}STLD$^{b}_{c}$	1.969	15.924	1.462	76.261	0.989
PAR	5.319	41.263	3.977	199.487	0.915
NPAR	5.375	40.769	3.979	191.434	0.912
ARIMA	3.991	33.742	2.979	160.368	0.953

Table 7. Ozone concentration in four Metropolitan Lima stations (μg/m^3): results (*p*-value) of the DM test for the final best-proposed model versus the baseline models given in Table 6.

ATE Station				
Models	**^{a}STLD$^{b}_{c}$**	**PAR**	**NPAR**	**ARIMA**
^{a}STLD$^{b}_{c}$	-	0.999	0.995	0.927
PAR	0.001	-	0.662	0.001
NPAR	0.005	0.338	-	0.006
ARIMA	0.073	0.999	0.995	-
Campo de Marte Station				
Models	**^{b}STLD$^{c}_{c}$**	**PAR**	**NPAR**	**ARIMA**
^{b}STLD$^{c}_{c}$	-	1.000	1.000	1.000
PAR	0.000	-	0.926	0.000
NPAR	0.000	0.074	-	0.000
ARIMA	0.000	1.000	1.000	-
San Borja Station				
Models	**^{b}STLD$^{b}_{c}$**	**PAR**	**NPAR**	**ARIMA**
^{b}STLD$^{b}_{c}$	-	1.000	1.000	1.000
PAR	0.000	-	0.907	0.000
NPAR	0.000	0.093	-	0.000
ARIMA	0.000	1.000	1.000	-
Santa Anita Station				
Models	**^{b}STLD$^{b}_{c}$**	**PAR**	**NPAR**	**ARIMA**
^{b}STLD$^{b}_{c}$	-	1.000	1.000	1.000
PAR	0.000	-	0.995	0.000
NPAR	0.000	0.005	-	0.000
ARIMA	0.000	1.000	1.000	-

In addition to the above, in the literature, Carbo-Bustinza [23] explored the correlations between ozone and meteorological variables and predicted ozone concentration for the same sites and winter periods selected in this study. They used models such as linear regression, support vector regression, decision trees, random forest, and multilayer perceptron and based their arguments on R^2, MSE, and MAE. The linear model presented the highest prediction performance for all the places evaluated (R^2: 0.9849–9923), supported by the lowest calculated errors (MAE: 0.0087–0.0724 and MSE: 0.0036–0.0087). Conversely, when

the ozone concentration model is represented exclusively as a function of time as a relevant factor without considering meteorological factors, the decomposition methods have shown great performance, since in this investigation the significant models ($p < 0.05$; R^2 max: 0.949) with errors less than 20% (RMSE, RMSPE, MAE, MAPE) showed great performance. These errors have been comparable to other STL decomposition studies that used root mean square error (RMSE: 6.8%) and mean absolute percentage error (MAPE: 10.49%) as benchmarks for forecast reliability for ozone [10]. This evaluation of tropospheric ozone explains its long-term and seasonal behavior with temporary ozone patterns [41], in accordance with what was demonstrated by Carbo-Bustinza [23] for the winter months in these geographic areas. This approach has presented high precision and strong performance that allows for preventing serious tropospheric ozone pollution events and optimizing the powers of the authorities and actors involved in decision making, especially at the urban level.

5. Conclusions

An improved tool for forecasting ozone concentration has been proposed using hybrid combinations of time series models in four districts of Metropolitan Lima between the years 2017 and 2019. It was shown that the combination of the models through the decomposition of the series ozone temporal data into "long-term trend", "seasonal", and "stochastic" series, by the use of the seasonal trend decomposition method, produced efficient model performance. The combinations made of the autoregressive models, nonlinear autoregressive models, and autoregressive moving average models generated 27 combinations for each sampling station. They demonstrated significant forecasts of the sample based on highly accurate and efficient descriptive, statistical, and graphic analysis tests, as a lower mean error occurred in the optimized forecast models compared to traditional models. Thus, the best hybrid models for the ATE ($^{a}\mathbf{STLD}^{b}_{c}$), CDM ($^{b}\mathbf{STLD}^{c}_{c}$), SB ($^{b}\mathbf{STLD}^{b}_{c}$), and Santa Anita ($^{c}\mathbf{STLD}^{b}_{c}$) stations were presented because they showed the best forecast reflected in the measurement of RMSE, RMSPE, MAE, MAPE, and CC, which were very small compared to the other models. The confirmation of the best models was statistically significant ($p < 0.05$), being superior to the other models. The graphical representation of the mean errors (RMSE, RMSPE, MAE, MAPE, and CC) for the twenty-seven models at each sampling station presented a better precision for the supermodels compared to the rest of all the models combined. These statistical tests and graphical results show that the proposed forecast methodology is highly accurate and efficient in predicting hourly ozone concentration, which meant that the independent AR, NPAR, and ARIMA models were outperformed by our best models ($p < 0.05$).

The main drawback of this study is that it only provides hourly data on ozone concentration. It can be extended to include additional exogenous factors such as wind speed, temperature, wind direction, and humidity, which may improve the short-term forecast of ozone concentration. In addition, the current work uses only four district datasets in Lima, Peru. This can be extended to other districts of Lima (San Juan de Lurigancho, Chorrillos, Comas, San Juan de Miraflores, etc.) or to different regions of Peru (Huánuco, Coyhaique, Traiguén, Padre Las Casas, Santiago, etc.). It could also be extended to the world level (Mexico, China, Japan, Malaysia, Pakistan, etc.) to evaluate the performance of the proposed hybrid time series modeling and forecasting technique. Moreover, only univariate time series models were used in this study, which should be extended by machine learning models such as deep learning and artificial neural networks. They can also be considered in the current hybrid time series forecasting framework. It can also be extended and applied to other approaches and datasets (for example, energy [42–44], air pollution [45,46], solid waste [47], and academic performance [48]).

Author Contributions: Conceptualization, N.C.-B., H.I., M.B. and J.L.L.-G.; methodology, software, and validation, H.I.; formal analysis, H.I. and J.L.L.-G.; investigation, N.C.-B., H.I. and J.L.L.-G.; resources, N.C.-B., H.I., M.B., R.J.C.-T. and J.L.L.-G.; data curation, N.C.-B., H.I. and J.L.L.-G.; writing—original draft preparation, N.C.-B., H.I., M.B., R.J.C.-T., A.R.H.D.L.C. and J.L.L.-G.; writing—review and editing, N.C.-B., H.I., M.B., R.J.C.-T., A.R.H.D.L.C. and J.L.L.-G.; visualization, N.C.-B., M.B., H.I. and J.L.L.-G.; supervision, M.B. and J.L.L.-G.; project administration, H.I., M.B. and J.L.L.-G.; funding acquisition, M.B. and R.J.C.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The collection and statistical processing of the data was carried out under the authorization of *Servicio Nacional de Meteorología e Hidrología del Perú*, a specialized technical agency of the Peruvian State that provides information on weather forecasting, as well as scientific studies in the areas of hydrology, meteorology, and environmental issues. The datasets are available in the repository, https://www.senamhi.gob.pe/site/descarga-datos/ (accessed on 21 July 2022).

Acknowledgments: The authors would like to thank the "INVESTIGA UCV" Teaching Research Support Fund of the Universidad César Vallejo for the financial support for the publication of this research.

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

References

1. Wang, T.; Xue, L.; Feng, Z.; Dai, J.; Zhang, Y.; Tan, Y. Ground-level ozone pollution in China: A synthesis of recent findings on influencing factors and impacts. *Environ. Res. Lett.* **2022**, *17*, 063003. [CrossRef]
2. Feng, Z.; Xu, Y.; Kobayashi, K.; Dai, L.; Zhang, T.; Agathokleous, E.; Calatayud, V.; Paoletti, E.; Mukherjee, A.; Agrawal, M.; et al. Ozone pollution threatens the production of major staple crops in East Asia. *Nat. Food* **2022**, *3*, 47–56. [CrossRef] [PubMed]
3. Jiang, Y.; Huang, J.; Li, G.; Wang, W.; Wang, K.; Wang, J.; Wei, C.; Li, Y.; Deng, F.; Baccarelli, A.A.; et al. Ozone pollution and hospital admissions for cardiovascular events. *Eur. Heart J.* **2023**, *44*, 1622–1632. [CrossRef]
4. Lei, Y.; Yue, X.; Liao, H.; Zhang, L.; Zhou, H.; Tian, C.; Gong, C.; Ma, Y.; Cao, Y.; Seco, R.; et al. Global perspective of drought impacts on ozone pollution episodes. *Environ. Sci. Technol.* **2022**, *56*, 3932–3940. [CrossRef] [PubMed]
5. Cabello-Torres, R.J.; Estela, M.A.P.; Sánchez-Ccoyllo, O.; Romero-Cabello, E.A.; Ávila, F.F.G.; Castañeda-Olivera, C.A.; Valdiviezo-Gonzales, L.; Eulogio, C.E.Q.; De La Cruz, A.R.H.; López-Gonzales, J.L. Statistical modeling approach for pm10 prediction before and during confinement by COVID-19 in South Lima, Perú. *Sci. Rep.* **2022**, *12*, 16737. [CrossRef] [PubMed]
6. Ding, J.; Dai, Q.; Fan, W.; Lu, M.; Zhang, Y.; Han, S.; Feng, Y. Impacts of meteorology and precursor emission change on O3 variation in Tianjin, China from 2015 to 2021. *J. Environ. Sci.* **2023**, *126*, 506–516. [CrossRef] [PubMed]
7. Wu, Q.; Lin, H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Sci. Total Environ.* **2019**, *683*, 808–821. [CrossRef]
8. Fu, H.; Zhang, Y.; Liao, C.; Mao, L.; Wang, Z.; Hong, N. Investigating PM 2.5 responses to other air pollutants and meteorological factors across multiple temporal scales. *Sci. Rep.* **2020**, *10*, 15639. [CrossRef]
9. Ewusie, J.E.; Soobiah, C.; Blondal, E.; Beyene, J.; Thabane, L.; Hamid, J.S. Methods, applications and challenges in the analysis of interrupted time series data: A scoping review. *J. Multidiscip. Healthc.* **2020**, *13*, 411–423. [CrossRef]
10. Li, W.; Jiang, X. Prediction of air pollutant concentrations based on TCN-BiLSTM-DMAttention with STL decomposition. *Sci. Rep.* **2023**, *13*, 4665. [CrossRef]
11. Tudor, C. Ozone pollution in London and Edinburgh: Spatiotemporal characteristics, trends, transport and the impact of COVID-19 control measures. *Heliyon* **2022**, *8*, e11384. [CrossRef]
12. Hong, J.; Wang, W.; Bai, Z.; Bian, J.; Tao, M.; Konopka, P.; Ploeger, F.; Müller, R.; Wang, H.; Zhang, J.; et al. The Long-Term Trends and Interannual Variability in Surface Ozone Levels in Beijing from 1995 to 2020. *Remote Sens.* **2022**, *14*, 5726. [CrossRef]
13. Chang, S.W.; Chang, C.L.; Li, L.T.; Liao, S.W. Reinforcement learning for improving the accuracy of pm 2.5 pollution forecast under the neural network framework. *IEEE Access* **2019**, *8*, 9864–9874. [CrossRef]
14. Gemst, M.V. Forecasting Stock Index Volatility—A Comparison of Models. Ph.D. Thesis, Universidade Nova de Lisboa, Lisbon, Portugal, 2020.
15. Iftikhar, H.; Bibi, N.; Canas Rodrigues, P.; López-Gonzales, J.L. Multiple Novel Decomposition Techniques for Time Series Forecasting: Application to Monthly Forecasting of Electricity Consumption in Pakistan. *Energies* **2023**, *16*, 2579. [CrossRef]
16. Iftikhar, H.; Turpo-Chaparro, J.E.; Canas Rodrigues, P.; López-Gonzales, J.L. Forecasting Day-Ahead Electricity Prices for the Italian Electricity Market Using a New Decomposition—Combination Technique. *Energies* **2022**, *15*, 3607. [CrossRef]

17. Ghoneim, O.A.; Manjunatha, B.R. Forecasting of ozone concentration in smart city using deep learning. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 1320–1326.
18. Juarez, E.K.; Petersen, M.R. A comparison of machine learning methods to forecast tropospheric ozone levels in Delhi. *Atmosphere* **2021**, *13*, 46. [CrossRef]
19. Chaloulakou, A.; Saisana, M.; Spyrellis, N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Sci. Total Environ.* **2003**, *313*, 1–13. [CrossRef]
20. Borhani, F.; Ehsani, A.H.; Hosseini Shekarabi, H.S. Prediction and spatiotemporal analysis of atmospheric Fine Particles and their effect on temperature and vegetation cover in Iran using Exponential Smoothing approach in Python. *J. Nat. Environ.* **2023**, *76*, 325–344.
21. Tang, H.; Bhatti, U.A.; Li, J.; Marjan, S.; Baryalai, M.; Assam, M.; Ghadi, Y.Y.; Mohamed, H.G. A New Hybrid Forecasting Model Based on Dual Series Decomposition with Long-Term Short-Term Memory. *Int. J. Intell. Syst.* **2023**, *2023*, 9407104. [CrossRef]
22. Romero, Y.; Diaz, C.; Meldrum, I.; Velasquez, R.A.; Noel, J. Temporal and spatial analysis of traffic–Related pollutant under the influence of the seasonality and meteorological variables over an urban city in Peru. *Heliyon* **2020**, *6*, e04029. [CrossRef]
23. Carbo-Bustinza, N.; Belmonte, M.; Jimenez, V.; Montalban, P.; Rivera, M.; Martínez, F.G.; Mohamed, M.M.H.; De La Cruz, A.R.H.; da Costa, K.; López-Gonzales, J.L. A machine learning approach to analyse ozone concentration in metropolitan area of Lima, Peru. *Sci. Rep.* **2022**, *12*, 22084. [CrossRef] [PubMed]
24. Leon, C.A.M.; Felix, M.F.M.; Olivera, C.A.C. Influence of Social Confinement by COVID-19 on Air Quality in the District of San Juan de Lurigancho in Lima, Perù. *Chem. Eng. Trans.* **2022**, *91*, 475–480.
25. Van Buuren, S.; Oudshoorn, C.G. *Multivariate Imputation by Chained Equations*; Netherlands Organization for Applied Scientific Research (TNO): The Hague, The Netherlands, 2000.
26. Cleveland, R.B.; Cleveland, W.S.; McRae, J.E.; Terpenning, I. STL: A seasonal-trend decomposition. *J. Off. Stat.* **1990**, *6*, 3–73.
27. Iftikhar, H.; Zafar, A.; Turpo-Chaparro, J.E.; Canas Rodrigues, P.; López-Gonzales, J.L. Forecasting Day-Ahead Brent Crude Oil Prices Using Hybrid Combinations of Time Series Models. *Mathematics* **2023**, *11*, 3548. [CrossRef]
28. Iftikhar, H.; Turpo-Chaparro, J.E.; Canas Rodrigues, P.; López-Gonzales, J.L. Day-Ahead Electricity Demand Forecasting Using a Novel Decomposition Combination Method. *Energies* **2023**, *16*, 6675. [CrossRef]
29. Davis, P.J.B.R.A. *Introduction to Time Series and Forecasting*; Springer: Berlin/Heidelberg, Germany, 2016.
30. Wasserman, L. *All of Nonparametric Statistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
31. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts: Melbourne, Australia, 2018.
32. Iftikhar, H. Modeling and Forecasting Complex Time Series: A Case of Electricity Demand. Master's, Thesis, Quaid-i-Azam University, Islamabad, Pakistan, 2018. Available online: https://www.researchgate.net/publication/372103958_Modeling_and_Forecasting_Complex_Time_Series_A_Case_of_Electricity_Demand (accessed on 28 July 2023).
33. Shah, I.; Iftikhar, H.; Ali, S.; Wang, D. Short-Term Electricity Demand Forecasting Using Components Estimation Technique. *Energies* **2019**, *12*, 2532. [CrossRef]
34. Shah, I.; Iftikhar, H.; Ali, S. Modeling and Forecasting Medium-Term Electricity Consumption Using Component Estimation Technique. *Forecasting* **2020**, *2*, 9. [CrossRef]
35. Shah, I.; Iftikhar, H.; Ali, S. Modeling and Forecasting Electricity Demand and Prices: A Comparison of Alternative Approaches. *J. Math.* **2022**, *2022*, 3581037. [CrossRef]
36. Diebold, F.; Mariano, R. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263.
37. Iftikhar, H.; Khan, M.; Khan, Z.; Khan, F.; Alshanbari, H.M.; Ahmad, Z. A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease. *Sustainability* **2023**, *15*, 2754. [CrossRef]
38. Iftikhar, H.; Khan, M.; Khan, M.S.; Khan, M. Short-Term Forecasting of Monkeypox Cases Using a Novel Filtering and Combining Technique. *Diagnostics* **2023**, *13*, 1923. [CrossRef] [PubMed]
39. Alshanbari, H.M.; Iftikhar, H.; Khan, F.; Rind, M.; Ahmad, Z.; El-Bagoury, A.A.A.H. On the Implementation of the Artificial Neural Network Approach for Forecasting Different Healthcare Events. *Diagnostics* **2023**, *13*, 1310. [CrossRef] [PubMed]
40. Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431. . [CrossRef]
41. Kawano, N.; Nagashima, T.; Sugata, S. Changes in the seasonal cycle of surface ozone over Japan during 1980–2015. *Atmos. Environ.* **2022**, *279*, 119108. [CrossRef]
42. Leite Coelho da Silva, F.; da Costa, K.; Canas Rodrigues, P.; Salas, R.; López-Gonzales, J.L. Statistical and artificial neural networks models for electricity consumption forecasting in the Brazilian industrial sector. *Energies* **2022**, *15*, 588. [CrossRef]
43. Gonzales, J.L.L.; Calili, R.F.; Souza, R.C.; Coelho da Silva, F.L. Simulation of the energy efficiency auction prices in Brazil. *Renew. Energy Power Qual. J.* **2016**, *1*, 574–579. [CrossRef]
44. López-Gonzales, J.L.; Castro Souza, R.; Leite Coelho da Silva, F.; Carbo-Bustinza, N.; Ibacache-Pulgar, G.; Calili, R.F. Simulation of the Energy Efficiency Auction Prices via the Markov Chain Monte Carlo Method. *Energies* **2020**, *13*, 4544. [CrossRef]
45. da Silva, K.L.S.; López-Gonzales, J.L.; Turpo-Chaparro, J.E.; Tocto-Cano, E.; Rodrigues, P.C. Spatio-temporal visualization and forecasting of PM 10 in the Brazilian state of Minas Gerais. *Sci. Rep.* **2023**, *13*, 3269. [CrossRef]
46. Jeldes, N.; Ibacache-Pulgar, G.; Marchant, C.; López-Gonzales, J.L. Modeling air pollution using partially varying coefficient models with heavy tails. *Mathematics* **2022**, *10*, 3677. [CrossRef]

47. Quispe, K.; Martínez, M.; da Costa, K.; Romero Giron, H.; Via y Rada Vittes, J.F.; Mantari Mincami, L.D.; Hadi Mohamed, M.M.; Huamán De La Cruz, A.R.; López-Gonzales, J.L. Solid Waste Management in Peru's Cities: A Clustering Approach for an Andean District. *Appl. Sci.* **2023**, *13*, 1646. [CrossRef]
48. Orrego Granados, D.; Ugalde, J.; Salas, R.; Torres, R.; López-Gonzales, J.L. Visual-Predictive Data Analysis Approach for the Academic Performance of Students from a Peruvian University. *Appl. Sci.* **2022**, *12*, 11251. [CrossRef]