PREDICTION OF DAYTIME AND NIGHTTIME GROUND-LEVEL OZONE USING THE HYBRID REGRESSION MODELS

Aimi Nursyahirah Ahmad[1], Samsuri Abdullah[1], Amalina Abu Mansor[2], Nazri Che Dom[3], Ali Najah Ahmed[4], Nurul Ain Ismail[1] and Marzuki Ismail[2]

[1]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia
[2]Faculty of Science and Marine Environment, Universiti Malaysia Terengganu, Kuala Nerus, Malaysia
[3]Faculty of Health Sciences, Universiti Teknologi MARA, UiTM Cawangan Selangor, Puncak Alam, Selangor, Malaysia
[4]Department of Civil Engineering, Institute of Energy Infrastructure (IEI), College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang, Selangor Darul Ehsan, Malaysia
E-Mail: samsuri@umt.edu.my

ABSTRACT

Ozone is one of the major challenges for the air quality community due to its adverse impact on the environment and human health. This study seeks to improve the understanding of underlying mechanisms for several developed models for ozone prediction. We aim to establish a robust prediction model for ozone concentration up to the next four hours. Three years dataset including ozone (O_3), nitrogen oxide (NO_x), nitric oxide (NO), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), particulate matter (PM_{10}, $PM_{2.5}$), wind speed, solar radiation, temperature, and relative humidity (RH) were used in this study. The data were analyzed by using Multiple Linear Regression (MLR), Principal Component Regression (PCR), and Cluster-Multiple Linear Regression (CMLR) in predicting the next hours of O_3 concentration. Results show that the MLR models executed high accuracy for O_{3t+1} (R^2= 0.313), $O_{3,t+2}$ (R^2= 0.265), $O_{3,t+3}$ (R^2= 0.227) and $O_{3,t+4}$ (R^2= 0.217) as the best fitted-model. In conclusion, the MLR model is suitable for the next hour's O_3 concentration prediction.

Keywords: Prediction, multiple linear regression, cluster, principal component analysis.

Manuscript Received 24 February 2023; Revised 22 July 2023; Published 13 August 2023

INTRODUCTION

New Malaysia Ambient Air Quality Standard (NMAAQS) by the Department of Environment Malaysia (DOE) has suggested that the average O_3 concentration for 1 hour is 200 µg/m^3. Multiple Linear Regression (MLR) is commonly used based on multiple predictors for prediction. MLR was simple and easy to compute, leading to better and more commonly used mathematical modeling to explain the underlying influencing factors of O_3 variation (Napi *et al.,* 2021). However, the multicollinearity problem becomes the main concern in the MLR as it can reduce the reliability of the predictive model (Fong *et al.,* 2018). It is because when the independent variables in the regression model correlated with each other, the independent variables are supposed to be independent. Thus, the influences of meteorological factors and gaseous pollutants on O_3 are varied in different areas. Hence, the C-MLR and PCR were introduced to reduce the multicollinearity problem. PCR is the hybrid model combination of PCA with MLR (Fong *et al.,* 2018) and CMLR is another hybrid model combination of HCA with MLR (Leong *et al.,* 2015). PCA is a statistical technique capable of generating statistically uncorrelated principal components (PCs) which are the linear amalgamation of the original variables (catchment and climatic characteristics) (Nguyen *et al.,* 2020). PCA is also one of the methods for isolating independent factors that significantly clarify the variance of the dependent variable to evaluate the dependency of meteorological elements on particulate concentrations and is used as a predictor in a line regression (Zuska *et al.,* 2019). The clustering analysis technique applied to aggregate the air quality stations is hierarchical clustering and known as unsupervised learning (Stolz *et al.,* 2020). However, there are no consensus criteria for selecting the most proper technique. Clustering algorithms have been developed for data mining and different clustering algorithms or even different ways to use them on the same dataset can lead to different partition results but none of them have proved to be the best technique in a large configuration (Govender and Sivakumar, 2020). Prediction is important to provide cities and human settlements with inclusion, safe, resilience, and sustainability. These methods could help in obtaining accurate data for the trend of ozone. The trend of ozone would be different because the four areas are in different zones. Besides, to quantify the strength of the association of O_3 with these factors to better understand the underlying mechanisms responsible for the changes in the surface of ozone levels in Cheras, Terengganu, Sarawak, and Sabah as a background.

MATERIALS AND METHODS

Data Acquisition and Processing

The secondary data in terms of 3 years time span from 1st January 2018 until 31st December 2020 were used in this study. The study areas are Cheras (S1), Kuala Terengganu (S2), Kuching (S3) and Sabah (S4). Hourly data of air pollutants (O_3, SO_2, NO_2, NO, NO_X, CO, $PM_{10,}$ and $PM_{2.5}$) along with meteorological factors (wind speed, relative humidity, solar radiation, and temperature) were obtained from the Department of Environment (DOE).

The data are not available for several periods of hours because of missing due to several reasons. An incomplete dataset affects the quality of data and imputations are needed before further analysis. The deletion technique was used in this study due to the number of missing values of more than 40%. The deletion technique is suitable for missing data of more than 40% to avoid bias among the data. Then, the data set is used to analyze in terms of descriptive statistics, and inferential statistics.

The dependent and independent variables consist of different units and normalization of the data set was required as the normalization can generate results ranging from 0 to 1. The normalization is also can interpret all the relationships in the data precisely and reduce bias. This scaling is suitable for improving the accuracy of numeric computation carried out by the MLR and PCR models for better outputs utilizing through min-max technique (Whalley and Zandi, 2016). The equation of normalization procedure as in Equation 1:

$$Z_i = \frac{(x_i - min(x))}{(max(x) - min(x)\,)} \qquad (1)$$

Where $Z_i =$ the ith normalized value in the data set; $X_i =$ the ith value in the data set; $(x) =$ the minimum value in data set (O$_3$); $(x) =$ the maximum value of data set (O$_3$)

After deleting the missing value, the data set was divided into two data sets for the development of the models namely Multiple Linear Regression (MLR), Cluster Multiple Linear Regression (CMLR), and Principal Component Regression (PCR) models. Overall data were divided into 70% for model development and the remaining 30% was used for model validation (Roy and Ambure, 2016). The validations of the models are needed to deal with forecasting as it estimates the precision of developed or obtained models. Therefore, the performance of models during the development and validation phase was assessed by several performance indicators which frequently used in air pollution forecasting (Baklanov and Zhang, 2020).

Data Analysis

PCA is a mathematical method that transforms a collection of interrelated variables into a set of uncorrelated variables, the main components, using an orthogonal transformation (Mishra *et al.,* 2017). A linear combination of the initial predictor variables to account for the variation in the information of each of the principal components. Many of the major elements were orthogonal, meaning that they were uncorrelated to each other. The first key component was measured in such a way that inside the dataset it accounts for the largest possible variation, followed by the concurrent components. Because the variables were calculated in separate units, normalized data was required before an interpretation of the main factor was done, which involves scaling each variable to 0 and 1. The principal component model presents the principal component as a linear function of the p-measured variables as expressed in Equation 2 (Laban *et al.,* 2018).

$$Z_1 = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \qquad (2)$$

Where, Z = the principal component; α = the component loading; X = the measured variable.

MLR research leads to a deeper and commonly used interpretation of the underlying driving factors of O$_3$ heterogeneity by statistical simulation through quick and fast computation (Cifuentes *et al.,* 2021). Future concentrations of O$_3$ were important for prediction since effective steps may be suggested by the local authority to improve air quality at a given location and can be used for precautionary measures. MLR is a linear regression method for multiple explanatory variables and can be developed as the Equation (3):

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i \qquad (3)$$

Where, x$_i$ = the explanatory variable of i (or independent variable); y = the dependent variable; β$_i$= the regression coefficient; ε$_i$ = the residual.

The clustering ensemble method represents an aggregation of multiple clustering techniques. D'Urso *et al.,* (2017) found that HC is the most widely used approach for aggregating air quality stations. It is commonly complemented with the application of the PCA technique for redundant analysis and selection. Furthermore, when there is no prior knowledge about an issue, k-means is the clustering technique with the greatest simplicity, efficiency, and low computational complexity (Alqurashi and Wang, 2018) as is the case with aggregation of monitoring sites as shown in Equation 4.

$$CR_k^n(\%) = 100\left[\frac{1}{3}P_{k,}^n\, PCA + \frac{1}{3}P_{k,}^n k - means + \frac{1}{3}\sum_{j=1}^{7}\frac{1}{7}P_{k,}^n HCj\right] \qquad (4)$$

Where $CR_k^n =$ measure of similarity among the selected clustering method; $P_{k,}^n =$ the possible kth partition of the network; $K =$ represents different individual results from the application of each clustering techniques; $n =$ clusters; $P_{k,}^n =$ the partition from the clustering technique

$P_{k,}^n$ takes a value of one if the partition is equal to at least one partition with exact n clusters within other clustering methods and zero if the partition does not coincide with any other partition from a different clustering technique. The CR_k^n computed iteratively from two clusters to the maximum number of principal components and its maximum value is 100%. This value is achieved when every technique chooses the same partition as the best choice, specifying n cluster. However, seven approaches to hierarchical clustering are possible(j), and a weight of (1/7) is applied to each (Stolz *et al.,* 2020).

The models were evaluated based on the model error and accuracy using various output measures, Normalized Absolute Error (NAE), Root Mean Square

Error (RMSE), Mean Absolute Error (MAE), Index of Agreement (IA), and Coefficient of determination (R^2). The best-fitted model was picked because it had high accuracy in which the IA and R^2 were closer to 1, and the best error of RMSE, MAE, and NAE were close to 0. The specification of performance metrics used in this analysis were tabulated in Table-1.

Table-1. Performance indicator.

Performance Indicator	Equation	Description
Normalized Absolute Error	$NAE = \frac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{\sum_{i=1}^{n} O_i}$	A value closer to zero is better
Root Mean Square Error (RMSE)	$RMSE = (\frac{1}{n}\sum_{i=1}^{n}[P_i - O_i]^2)^{1/2}$	A value closer to zero is better
Mean Absolute Error (MAE)	$MAE = \frac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{n}$	A value closer to zero is better
Index of Agreement (IA)	$IA = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert^2}$	Value closer to one is better
Coefficient of determination (R^2)	$R^2 = (\frac{\sum_{i}^{n}(P_i - \underline{P})(O_i - \underline{O})}{n.S_{pred.}S_{obs}})^2$	Value closer to one is better

Where, n = total number of data; P_i = predicted values; O_i = observed values; P = mean of predicted; $\bar{O}$ = mean of observed values; S_{pred} = standard deviation of predicted values; S_{obs} = standard deviation of observed values

RESULT AND DISCUSSIONS

Models Development

The MLR models were developed, and the model summary was depicted in Table-2. The model was developed for predicting the next hour of $O_{3,\ t+1}$ concentration up to $O_{3,\ t+4}$ concentration, as to discern the range of significant leads during daytime and night-time. The value of Variance Inflation Factor (VIF) for the independent variables for prediction $O_{3,\ t+n}$, where n= 1, 2, 3, 4 is lower than 10 which indicates there is no multicollinearity between the independent variables. Durbin Watson shows that the models did not have any first autocorrelation problems as the values were in between 0 and 4. During the daytime S1, S2, S3 and S4 models show that $O_{3,\ t+1}$ having a higher coefficient of determination of 0.313, 0.580, 0.492 and 0.514, respectively as compared to model $O_{3,\ t+2}$, $O_{3,\ t+3}$ and $O_{3,\ t+4}$ for each area as tabulated in Table-2. Meanwhile, throughout the night-time, the S1, S2, S3 and S4 models demonstrate that $O_{3,\ t+1}$ has a higher coefficient of determination of 0.389,0.429,0.346, and 0.436 for each region than models $O_{3,\ t+2}$, $O_{3,\ t+3}$, and $O_{3,\ t+4}$, as shown in Table-2. The lowest R^2 was clarify by model $O_{3,\ t+4}$ for each area during daytime, S1 (0.217), S2 (0.424), S3 (0.388) and S4 (0.257) while during night-time, S1 (0.285), S2 (0.256), S3 (0.209) and S4 (0.301). It demonstrates that the MLR models' ability to estimate ozone concentrations is greater for the next hour and lower for the following three hours (Awang *et al.,* 2015). Based on this study, shows that different ozone precursors and meteorological factors are influencing $O_{3,\ t+1}$ concentration during the daytime and night-time. Optimal conditions played an important role in the photochemical interaction of ozone and ozone precursors occurring during the daytime, which will raise the concentration of O_3 in the atmosphere (Napi *et al.,* 2021).

The study proceeds using Principal Component Regression (PCR), which reduces the total number of parameters into a smaller number of principal components to reduce the multicollinearity. The PCR models were developed, and the model summary was depicted in Table-3. The value of Variance Inflation Factor (VIF) for the independent variables for prediction $O_{3,\ t+n}$, where n= 1, 2, 3, 4 is 1 which is lower than 10 which indicates there is no multicollinearity between the independent variables. The Durbin Watson also shows the range between 0.987 and 2.405 and the models did not have any first autocorrelation problems as the values were between 0 and 4. PCR is applied to analyze and forecast O_3 concentrations for the next hour of $O_{3,\ t+1}$ concentration up to $O_{3,\ t+4}$ concentration, to identify the range of significant daytime and night-time evidence using gaseous pollutants and meteorological data as independent variables. During the daytime, the S1, S2, S3, and S4 models show that $O_{3,\ t+1}$ have a higher coefficient of determination of 0.220, 0.419, 0.343, and 0.328, respectively as compared to model $O_{3,\ t+2}$, $O_{3,\ t+3}$, and $O_{3,\ t+4}$ for each area. Meanwhile, throughout the night-time, the S1, S,2, S,3, and S4 models demonstrate that $O_{3,\ t+1}$ has a lower coefficient of determination with 0.302, 0.267, 0.190 and 0.34 for each region, respectively to models $O_{3,\ t+2}$, $O_{3,\ t+3}$, and $O_{3,\ t+4}$. The lowest R^2 was clarified by model $O_{3,\ t+4}$ for each area during daytime, the S1 (0.150), S2 (0.332), S3 (0.274), and S4 (0.173) while during night-time, the S1 (0.190), S2 (0.181), S3 (0.097) and S4 (0.276). The models showed that all inputs (PC-1, PC-2, PC-3, and PC-4) were chosen as important predictors for the O_3 concentration over the following four hours.

The variability of ozone in the four areas was further investigated by conducting a Hierarchical Agglomerative Cluster Analysis (HACA). The results of the investigation during the daytime for S1 and S2 confirm having the same cluster which is C1 (PM_{10}, $PM_{2.5}$, SO_2, NO, NO_2, NO_X, CO). S3 and S4 also shows having the same cluster C1 (PM_{10}, $PM_{2.5}$, CO, NO, NO_X, NO_2, SO_2, WS, O_3), C1 (PM_{10}, $PM_{2.5}$, CO, NO, NO_X, NO_2, WS, O_3). There are different cluster (C2) for each site which are S1 (C2: WS, O_3), S2 (C2: O_3, WS, SR), S3 (C2: SR, TEMP) and S4 (C2: SO_2). S2, S3 and S4 confirm that have same cluster which are C3 (RH) while S1 (C3: SR, TEMP). There are different cluster for C4 (S1: RH), (S2: TEMP), and (S4: SR). S4 shows C5: Temp. During the night-time for each site of S1 and S2 (C1: PM_{10}, $PM_{2.5}$, SO_2, NO, SR, WS, O_3, NO_X, CO, TEMP), S3 (C1: $PM_{2.5}$, $SO_{2,}$ SR, NO, $O_{3,}$ WS, NO_X, CO, TEMP) and S4 (C1: PM_{10}, $PM_{2.5}$, NO_2, $NO_{X,}$ WS, NO, SR, O_3, CO, SO_2, Temp). Meanwhile, the S1, S2 and S4 confirm having the same cluster which is C2 (RH). S3 shows that having the C3 (PM_{10}, RH). Awang *et al.,* (2015) and Warmiński *et al.,* (2018) claimed that wind speed and wind direction are the important meteorological parameters that have contributed to the variation of O_3 concentration changes during night-time. In contrast to the daytime, solar radiation levels decrease at night, which is linked to the lack of significance of UVB radiation, temperature, and relative humidity on ozone concentration, this condition occurs of the dependence of temperature and relative humidity on sunlight, ozone concentrations, and other factors.

CMLR models were developed, and the model summary was depicted in Table-4. The value of Variance Inflation Factor (VIF) for the independent variables for prediction $O_{3,\ t+n}$, where n= 1, 2, 3, 4 is lower than 10 which are range between 1.434 and 2.597 during daytime. Meanwhile, during night-time the value of VIF is range between 1.481 and 2.565. These indicate there is no multicollinearity between the independent variables. Durbin Watson shows that the models did not have any first autocorrelation problems as the values were between 0 and 4. During daytime, the S1, S2, S3 and S4 models show that $O_{3,\ t+1}$ have a higher coefficient of determination of 0.313, 0.580, 0.492 and 0.508, respectively as compared to model $O_{3,\ t+2}$, $O_{3,\ t+3}$ and $O_{3,\ t+4}$ for each area. Meanwhile, throughout the night-time, the S1, S2, S3, and S4 models demonstrate that $O_{3,\ t+1}$ also has a higher coefficient of determination of 0.389, 0.429, 0.346, and 0.436 for each region than models $O_{3,\ t+2}$, $O_{3,\ t+3}$, and $O_{3,\ t+4}$. The lowest R^2 was clarified by model $O_{3,\ t+4}$ for each area during daytime, S1 (0.213), S2 (0.424), S3 (0.387), and S4 (0.256) while during night-time, S1 (0.283), S2 (0.256), S3 (0.209) and S4 (0.301).

Table-2. Summary of Multiple Linear Regression (MLR) models.

Model	R	R^2	Durbin-Watson
Daytime; Site 1			
$O_{3,\ t+1} = 0.461O_3 - 0.495CO + 0.182PM_{2.5} - 0.182SO_2 + 0.222NO_X + 0.030WS + 0.224$	0.559	0.313	2.235
$O_{3,\ t+2} = 0.369O_3 - 0.505CO + 0.113WS - 0.217SO_2 + 0.085NO_2 + 0.507PM_{2.5} - 0.052RH - 0.343PM_{10} + 0.168NO_X + 0.255$	0.515	0.265	1.523
$O_{3,\ t+3} = 0.336O_3 - 0.590CO + 0.076WS - 0.234SO_2 + 0.499PM_{2.5} + 0.327NO_X - 0.309PM_{10} - 0.030RH + 0.281$	0.476	0.217	1.403
$O_{3,\ t+4} = 0.326O_3 - 0.556CO + 0.103WS + 0.099NO_2 - 0.185SO_2 + 0.825PM_{2.5} - 0.677PM_{10} - 0.040RH + 0.225NO_X + 0.021SR + 0.256$	0.466	0.217	1.403
Daytime; Site 2			
$O_{3,\ t+1} = 0.659O_3 - 0.089RH - 0.132NO + 0.084NO_2 - 0.068PM_{10} + 0.161$	0.762	0.580	2.390
$O_{3,\ t+2} = 0.550O_3 - 0.142RH - 0.118PM_{10} - 0.107NO + 0.074NO_2 + 0.233$	0.701	0.492	1.435
$O_{3,\ t+3} = 0.480O_3 - 0.111RH - 0.119PM_{10} - 0.119NO + 0.058NO_2 + 0.059TEMP + 0.031WS + 0.196$	0.666	0.444	1.258
$O_{3,\ t+4} = 0.461O_3 + 0.118TEMP - 0.150PM_{10} - 0.142NO + 0.099NO_2 - 0.080RH + 0.153$	0.652	0.424	1.122
Daytime; Site 3			
$O_{3,t+1} = 0.574O_3 + 0.814PM_{2.5} + 0.042SR - 0.096TEMP - 0.108RH + 0.102WS + 0.119NO_2 - 0.702PM_{10} + 0.172$	0.702	0.492	2.406
$O_{3,\ t+2} = 0.525O_3 + 0.148WS + 0.133NO_2 + 0.065SR - 0.099TEMP - 0.092RH + 0.771PM_{2.5} - 0.681PM_{10} - 0.058NO + 0.163$	0.671	0.450	1.581
$O_{3,\ t+3} = 0.493O_3 + 0.065SR + 0.155NO_2 + 0.128WS - 0.091TEMP - 0.112RH - 0.067NO + 0.740PM_{2.5} - 0.689PM_{10} + 0.183$	0.646	0.417	1.479
$O_{3,\ t+4} = 0.509O_3 + 0.066SR + 0.102WS - 0.073TEMP - 0.077RH + 0.092NO_2 - 0.068NO + 0.155$	0.623	0.388	1.392
Daytime; Site 4			
$O_{3,\ t+1} = 0.667O_3 + 0.031SR + 0.048NO_2 - 0.134TEMP - 0.126RH - 0.036WS + 0.232$	0.717	0.514	2.352
$O_{3,\ t+2} = 0.563O_3 + 0.040SR + 0.048NO_2 - 0.118RH - 0.107TEMP - 0.032WS + 0.238$	0.639	0.408	1.346
$O_{3,\ t+3} = 0.475O_3 + 0.044SR - 0.083RH - 0.068TEMP - 0.031WS + 0.223$	0.557	0.310	1.117
$O_{3,\ t+4} = 0.426O_3 + 0.044SR - 0.447PM_{10} + 0.455PM_{2.5} + 0.072NO_2 - 0.040CO - 0.012SO_2 + 0.177$	0.506	0.257	0.990
Nighttime; Site 1			
$O_{3,\ t+1} = 0.495O_3 + 0.231SR + 0.104NO_2 - 0.158RH - 0.083TEMP + 0.042WS - 0.111NO + 0.121CO - 0.111PM_{2.5} + 0.150$	0.623	0.389	2.223
$O_{3,\ t+2} = 0.433O_3 + 0.185NO_2 + 0.244SR + 0.090WS - 0.134RH - 0.094TEMP - 0.128PM_{10} - 0.121NO_X + 0.099CO + 0.130$	0.560	0.314	1.517
$O_{3,\ t+3} = 0.395O_3 + 0.251SR + 0.133NO_2 + 0.088WS - 0.157PM_{10} - 0.094RH - 0.074SO_2 - 0.105NO + 0.102CO - 0.049TEMP + 0.097$	0.526	0.276	1.430
$O_{3,\ t+4} = 0.397O_3 + 0.317SR + 0.179NO_2 + 0.094WS - 0.339PM_{10} - 0.084SO_2 + 0.193PM_{2.5} - 0.070RH - 0.054TEMP + 0.104CO - 0.093NO_X + 0.075$	0.534	0.285	1.363
Nighttime; Site 2			
$O_{3,\ t+1} = 0.566O_3 + 0.166NO_2 + 0.161SR - 0.171TEMP - 0.184RH - 0.391PM_{2.5} + 0.291PM_{10} + 0.135CO - 0.131NO + 0.046WS + 0.237$	0.655	0.429	2.205
$O_{3,\ t+2} = 0.439O_3 + 0.126NO_2 + 0.181SR - 0.205RH - 0.150TEMP + 0.107WS + 0.204CO - 0.359PM_{2.5} - 0.184NO + 0.274PM_{10} + 0.252$	0.583	0.340	1.439
$O_{3,\ t+3} = 0.380O_3 + 0.274CO + 0.201SR - 0.187RH - 0.108TEMP + 0.169NO_2 + 0.103WS - 0.080SO_2 - 0.194NO_X - 0.077PM_{2.5} + 0.236$	0.536	0.288	1.274
$O_{3,\ t+4} = 0.383O_3 + 0.213SR + 0.152NO_2 - 0.114RH - 0.092SO_2 + 0.063WS - 0.048TEMP + 0.188$	0.506	0.256	1.1230
Nighttime; Site 3			
$O_{3,\ t+1} = 0.634O_3 + 0.264NO_2 + 0.267SR + 0.059CO - 0.005$	0.588	0.346	2.199
$O_{3,\ t+2} = 0.572O_3 + 0.257NO_2 + 0.310SR - 0.022TEMP + 0.061CO - 0.141SO_2 + 0.018$	0.535	0.286	1.581
$O_{3t+3} = 0.512O_3 + 0.245NO_2 + 0.333SR - 0.088PM_{10} - 0.019TEMP + 0.118$	0.490	0.240	1.448
$O_{3t+4} = 0.455O_3 + 0.220NO_2 + 0.276SR + 0.053WS - 0.028TEMP - 0.042RH + 0.076$	0.457	0.209	1.358
Nighttime; Site 4			
$O_{3,\ t+1} = 0.466O_3 + 0.279NO_2 - 0.091RH + 0.031SO_2 + 0.034TEMP + 0.118SR - 0.036WS - 0.024CO + 0.282PM_{10} - 0.240PM_{2.5} + 0.075$	0.661	0.436	2.270
$O_{3,\ t+2} = -0.125RH + 0.413O_3 + 0.033SO_2 - 0.078CO + 0.419NO_X + 0.045PM_{10} - 0.256NO + 0.092SR + 0.139$	0.640	0.410	1.603
$O_{3t+3} = -0.094RH + 0.361O_3 - 0.095CO + 0.035SO_2 + 0.050TEMP + 0.470NO_X - 0.296NO - 0.048WS + 0.105$	0.591	0.350	1.475
$O_{3,\ t+4} = -0.087RH + 0.292O_3 + 0.040SO_2 + 0.061TEMP - 0.069CO + 0.144SR + 0.436NO_X - 0.288NO + 0.090$	0.549	0.301	1.379

Table-3. Summary of Principal Component Regression (PCR) models.

Model	R	R^2	Durbin-Watson
Daytime; Site 1			
$O_{3t+1} = -0.056PC1 - 0.046PC2 + 0.030PC3 + 0.316$	0.469	0.220	1.759
$O_{3t+2} = - 0.055PC1 - 0.034PC2 + 0.019PC3 + 0.316$	0.442	0.195	1.362
$O_{3t+3} = -0.055PC1 - 0.034PC2 + 0.019PC3 + 0.316$	0.403	0.162	1.317
$O_{3t+4} = -0.053PC1 - 0.034PC2 + 0.017PC3 + 0.316$	0.388	0.150	1.274
Daytime; Site 2			
$O_{3t+1} = -0.102PC1 + 0.031PC2 - 0.012PC3 +0.316$	0.647	0.419	1.525
$O_{3t+2} = -0.098PC1 + 0.022PC2 - 0.007PC3 + 0.316$	0.610	0.372	1.115
$O_{3t+3} = -0.096PC1 + 0.018PC2 - 0.007PC3 + 0.316$	0.590	0.348	1.032
$O_{3t+4} = -0.094PC1 + 0.016PC2 - 0.006PC3 + 0.316$	0.576	0.332	0.969
Daytime; Site 3			
$O_{3t+1} = 0.045PC1 - 0.036PC2 + 0.063PC3 + 0.263$	0.586	0.343	1.653
$O_{3t+2} = 0.039PC1 - 0.036PC2 + 0.061PC3 + 0.263$	0.556	0.309	1.220
$O_{3t+3} = 0.035PC1 - 0.033PC2 + 0.61PC3 + 0.263$	0.536	0.287	1.163
$O_{3t+4} = 0.031PC1 - 0.035PC2 + 0.060PC3 + 0.263$	0.524	0.274	1.150
Daytime; Site 4			
$O_{3t+1} = 0.051PC1 + 0.026PC2 - 0.005PC3 + 0.002PC4 + 0.292$	0.573	0.328	1.528
$O_{3t+2} = 0.047PC1 + 0.022PC2 - 0.005PC3 +0.292$	0.525	0.276	1.053
$O_{3t+3} = 0.043PC1 + 0.017PC2 - 0.004PC3 + 0.292$	0.462	0.214	0.953
$O_{3t+4} = 0.039PC1 + 0.012PC2 - 0.005PC3 + 0.292$	0.416	0.173	0.881
Nighttime; Site 1			
$O_{3t+1} = -0.015PC1 + 0.071PC2 + 0.010PC3 + 0.110$	0.550	0.302	1.796
$O_{3t+2} = -0.010PC1 + 0.062PC2 + 0.007PC3 + 0.110$	0.477	0.228	1.363
$O_{3t+3} = - 0.009PC1 + 0.058PC2 + 0.004PC3 + 0.110$	0.443	0.196	1.265
$O_{3t+4} = - 0.005PC1 + 0.057PC2 + 0.003PC3 + 0.110$	0.436	0.190	1.226
Nighttime; Site 2			
$O_{3t+1} = 0.077PC1 - 0.022PC2 + 0.008PC3 +0.195$	0.516	0.267	1.628
$O_{3t+2} = -0.015PC1 + 0.009PC2 + 0.071PC3 + 0.195$	0.468	0.219	1.193
$O_{3t+3} = -0.010PC1 + 0.011PC2 + 0.066PC3 + 0.195$	0.434	0.188	1.125
$O_{3t+4} = -0.014PC1 + 0.005PC2 + 0.065PC3 + 0.195$	0.425	0.181	1.093
Nighttime; Site 3			
$O_{3t+1} = -0.028PC1 + 0.036PC2 + 0.035PC3 + 0.164$	0.435	0.190	1.620
$O_{3t+2} = -0.023PC1 + 0.032PC2 + 0.029PC3 + 0.164$	0.370	0.137	1.271
$O_{3t+3} = -0.020PC1 + 0.030PC2 + 0.027PC3 + 0.164$	0.344	0.118	1.184
$O_{3t+4} = -0.019PC1 + 0.023PC2 + 0.028PC3 + 0.164$	0.311	0.097	1.161
Nighttime; Site 4			
$O_{3t+1} = 0.028PC1 + 0.056PC2 - 0.003PC3 + 0.102$	0.620	0.384	1.963
$O_{3t+2} = 0.027PC1 + 0.054PC2 - 0.002PC3 + 0.102$	0.599	0.358	1.495
$O_{3t+3} = 0.025PC1 + 0.051PC2 + 0.102$	0.554	0.307	1.409
$O_{3t+4} = 0.023PC1 + 0.048PC2 + 0.002PC3 + 0.102$	0.525	0.276	1.329

Table-4. Summary of Cluster-Multiple Linear Regression (C-MLR) models.

Model	R	R^2	Durbin-Watson
Daytime; Site 1			
O_{3t+1} = $-0.251NO + 0.162PM_{10} - 0.473CO - 0.191SO_2 + 0.441NO_X - 0.065NO_2 + 0.447O_3 + 0.037WS + 0.020SR + 0.211$	0.560	0.313	2.237
O_{3t+2} = $-0.197NO - 0.471CO + 0.135PM_{10} - 0.229SO_2 + 0.373NO_X - 0.033NO_2 + 0.376O_3 + 0.110WS + 0.025SR + 0.215$	0.513	0.263	1.526
$O_{3t+3} = -0.086NO - 0.557CO + 0.158PM_{10} - 0.250SO_2 + 0.315NO_X + 0.331O_3 + 0.076WS + 0.020SR + 0.252$	0.475	0.225	1.435
$O_{3t+4} = -0.133NO - 0.538CO + 0.139PM_{2.5} + 0.363NO_X - 0.209SO_2 + 0.324O_3 + 0.093WS + 0.033SR + 0.239$	0.461	0.213	1.403
Daytime; Site 2			
$O_{3t+1} = -0.132NO - 0.068PM_{10} + 0.084NO_2 + 0.659O_3 + 0.000114SR - 0.089RH + 0.161$	0.762	0.580	2.390
$O_{3t+2} = -0.106NO - 0.118PM_{10} + 0.074NO_2 - 0.007SO_2 + 0.550O_3 -0.000192SR - 0.142RH + 0.234$	0.701	0.492	1.435
$O_{3t+3} = -0.118NO + 0.057NO_2 - 0.125PM_{10} + 0.489O_3 + 0.005SR - 0.106RH + 0.059TEMP + 0.198$	0.666	0.443	1.260
$O_{3t+4} = -0.156NO + 0.100NO_2 - 0.154PM_{2.5} + 0.453O_3 + 0.003SR - 0.073RH + 0.130TEMP + 0.138$	0.651	0.424	1.124
Daytime; Site 3			
$O_{3t+1} = 0.574O_3 + 0.812PM_{2.5} + 0.103WS + 0.118NO_2 - 0.706PM_{10} + 0.009CO + 0.042SR - 0.096TEMP - 0.109RH + 0.172$	0.702	0.492	2.406
$O_{3t+2} = 0.539O_3 + 0.151WS + 0.116NO_2 + 0.783PM_{2.5} - 0.711PM_{10} + 0.017CO + 0.063SR - 0.103TEMP - 0.094RH + 0.161$	0.670	0.449	1.581
$O_{3t+3} = 0.534O_3 + 0.118WS + 0.115NO_2 + 0.060SR - 0.082TEMP - 0.096RH + 0.155$	0.641	0.411	1.472
$O_{3t+4} = 0.523O_3 + 0.103WS + 0.070NO_2 + 0.064SR - 0.075TEMP - 0.077RH + 0.153$	0.622	0.387	1.391
Daytime; Site 4			
$O_{3t+1} = 0.710O_3 + 0.017SR - 0.023TEMP + 0.091$	0.713	0.508	2.356
$O_{3t+2} = 0.570O_3 - 0.003CO - 0.113RH + 0.031SR - 0.112TEMP + 0.237$	0.638	0.406	1.345
$O_{3t+3} = 0.491O_3 - 0.347PM_{10} + 0.379PM_{2.5} - 0.018CO - 0.009SO_2 + 0.035SR + 0.152$	0.560	0.314	1.129
$O_{3t+4} = 0.427O_3 - 0.435PM_{10} + 0.449PM_{2.5} - 0.030CO - 0.012WS - 0.011SO_2 + 0.041SR + 0.181$	0.506	0.256	0.988
Nighttime; Site 1			
O_{3t+1} = $0.495O_3 + 0.231SR + 0.104NO_2 - 0.083TEMP + 0.042WS - 0.111NO - 0.111PM_{2.5} + 0.121CO - 0.158RH + 0.150$	0.623	0.389	2.223
O_{3t+2} = $0.433O_3 + 0.185NO_2 + 0.244SR + 0.090WS - 0.128PM_{10} - 0.121NO_X + 0.099CO - 0.094TEMP - 0.134RH + 0.130$	0.560	0.314	1.517
$O_{3t+3} = 0.395O_3 + 0.251SR + 0.133NO_2 + 0.088WS - 0.157PM_{10} - 0.105NO + 0.102CO - 0.074SO_2 - 0.049TEMP - 0.094RH + 0.097$	0.526	0.276	1.430
$O_{3t+4} = 0.382O_3 + 0.311SR + 0.173NO_2 + 0.103WS - 0.379PM_{10} - 0.093SO_2 + 0.285PM_{2.5} - 0.023RH + 0.023$	0.532	0.283	1.360
Nighttime; Site 2			
$O_{3t+1} = 0.576O_3 + 0.167NO_2 + 0.165SR - 0.028SO_2 - 0.380PM_{2.5} + 0.282PM_{10} + 0.128CO - 0.119NO - 0.179RH - 0.168TEMP + 0.240$	0.655	0.429	2.203
$O_{3t+2} = 0.437O_3 + 0.125NO_2 + 0.179SR + 0.206CO + 0.107WS - 0.044SO_2 - 0.179NO - 0.346PM_{2.5} + 0.267PM_{10} - 0.204RH - 0.143TEMP + 0.252$	0.584	0.341	1.438
$O_{3t+3} = 0.380O_3 + 0.274CO + 0.201SR + 0.169NO_2 - 0.080SO_2 + 0.103WS - 0.194NO_X - 0.077PM_{2.5} - 0.187RH - 0.108TEMP + 0.236$	0.536	0.288	1.274
$O_{3t+4} = 0.398O_3 + 0.220SR + 0.150NO_2 - 0.093SO_2 - 0.109RH - 0.048TEMP + 0.190$	0.505	0.256	1.230
Nighttime; Site 3			
$O_{3t+1} = 0.634O_3 + 0.264NO_2 + 0.267SR + 0.059CO - 0.005$	0.588	0.346	2.199
$O_{3t+2} = 0.572O_3 + 0.257NO_2 + 0.310SR - 0.022TEMP + 0.061CO - 0.141SO_2 + 0.018$	0.535	0.286	1.581
$O_{3t+3} = 0.512O_3 + 0.246NO_2 + 0.333SR + 0.078PM_{2.5} - 0.020TEMP + 0.031$	0.490	0.240	1.447
$O_{3t+4} = 0.455O_3 + 0.220NO_2 + 0.276SR + 0.053WS - 0.028TEMP - 0.042RH + 0.076$	0.457	0.209	1.358
Nighttime; Site 4			
$O_{3t+1} = 0.466O_3 + 0.279NO_2 + 0.034TEMP + 0.031SO_2 + 0.118SR - 0.036WS + 0.282PM_{10} - 0.240PM_{2.5} - 0.024CO - 0.091RH + 0.075$	0.661	0.436	2.270
$O_{3t+2} = 0.419O_3 + 0.002TEMP + 0.033SO_2 - 0.079CO + 0.040PM_{10} + 0.425NO_X - 0.259NO - 0.033WS + 0.093SR - 0.124RH + 0.140$	0.640	0.410	1.605
$O_{3t+3} = 0.360O_3 + 0.049TEMP + 0.035SO_2 - 0.094CO - 0.049WS + 0.469NO_X - 0.295NO + 0.079SR - 0.093RH + 0.104$	0.592	0.350	1.476
$O_{3t+4} = 0.292O_3 + 0.061TEMP + 0.040SO_2 - 0.069CO + 0.144SR + 0.436NO_X - 0.288NO - 0.087RH + 0.090$	0.549	0.301	1.379

Model Evaluation and Selection

Five performance indicators were used to measure the error and accuracy of the models to evaluate the performance and test the best-fitted model. The performance indicators for MLR, PCR, and CMLR models for O_3 prediction models are shown in Tables 5 and 6. The error of the model was calculated using Normalized Absolute Error (NAE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), with a value closer to zero indicating a better model. The accuracy of the models' output was calculated using the coefficient of determination (R^2) and index of agreement (IA), where a number nearer to 1 indicates more accuracy (Moursi *et al.,* 2022).

The best prediction model for O_3 concentration, $O_{3,\ t+n}$, where n= 1, 2, 3, 4 is MLR. Site 1 indicates the high and low error range from $1.00E^{-2}$ to 98.289 (NAE), 0.409 ppb to 2.642 ppb (RMSE) and 0.015 ppb to 0.019 ppb (MAE) whereas has high accuracy between 0.071 to 0.666 (IA) and 0.217 to 0.389 (R^2). Meanwhile, the error and accuracy values for PCR at Site 1 range between 2.197 and 70.588 (NAE), 0.107 ppb and 0.386 ppb (RMSE), 0.001 µg/m^3 and 0.003 ppb (MAE), 0.231 and 0.487 (IA) and 0.150 and 0.302 (R^2) compared with the range values of error and accuracy for CMLR, Site 1, NAE from $1.141E^{-2}$ to 98.289, RMSE from 0.063 ppb to 2.435 ppb , MAE (0.0004 ppb to 0.017 ppb), IA (0.165 to 0.746) and R^2 (0.213 to 0.389). MLR for Site 2 also composed the low error and high accuracy ranges between 1.201 and 40.110 (NAE), 0.096 ppb and 0.866 ppb (RMSE), 0.001 ppb and 0.006 ppb for MAE, 0.123 and 0.402 (IA) and R^2 (0.256 and 0.580) compared to PCR model with range between 3.340 and 4.276 (NAE), 0.111ppb and 0.187 ppb (RMSE), MAE equals to 0.001 ppb, 0.309 and 0.422 (IA)and R^2 (0.181 and 0.419), CMLR model with range from 1.103 and 40.110 (NAE), 0.090 ppb to 0.866 ppb (RMSE), 0.001 ppb to 0.006 ppb (MAE), 0.123 to 0.408 (IA) and R^2 (0.256 to 0.580).

The model was used for Site 3 and Site 4 MLR with ranges 0.913 and 72.263 (NAE), 0.054 ppb and 0.494 ppb (RMSE), 0.0003 ppb and 0.003 ppb (MAE), 0.092 and 0.576 (IA) and 0.257 and 0.514 (R^2), respectively compared to PCR for the same site range from 3.798 to 6.455 (NAE), 0.043 ppb to 0.166 ppb (RMSE), 0.0003 ppb to 0.001 ppb (MAE), 0.291 to 0.380 (IA), 0.097 to 0.343 (R^2) and 0.0512 to 10.321(NAE) and 0.042 ppb to 0.052 ppb (RMSE), 0.0002 ppb to 0.0004 ppb (MAE), 0.265 to 0.541 (IA), 0.173 to 0.384 (R^2), and also compared with CMLR for the same site with a range between 0.835 and 18.639 (NAE), 0.021 ppb and 0.093 ppb (RMSE), 0.001 ppb and 0.002 ppb (MAE), 0.151 and 0.611 (IA), 0.209 and 0.492 (R^2) and 1.367 and 71.501 (NAE), 0.022 ppb and 0.489 ppb (RMSE), 0.0003 ppb and 0.001 ppb (MAE), 0.100 and 0.924 (IA), 0.256 and 0.508 (R^2), respectively.

To demonstrate that the model could accurately forecast the following hours of O_3 concentration, the best-selected model (MLR) was assigned. The results of performance measures for the predicted O_3 concentration to the next hours by using MLR detected has a small error measurement compared to the result of performance measures by using PCR and CMLR. The accuracy measure (IA and R^2) also proves that the MLR model indicates high agreement between the observed and predicted data with increased accuracy compared to the results of accuracy PCR and CMLR. It was demonstrated that the MLR models can be utilized to forecast the O_3 concentration because of the good agreement between the predicted data and the observed data (Abdullah *et al.,* 2019; Awang *et al.,* 2015).

One of the most used techniques for forecasting ozone concentrations of weather variables and several atmospheric pollutants is multiple linear regression (Hashim *et al.,* 2022). Furthermore, MLR enables the prediction of maximum O_3 concentration in urban areas several hours in advance (Silva *et al*., 2022). According to Verma *et al.,* (2015); Laban *et al.,* (2018), and Silva *et al.,* (2022), the relationship between the meteorological conditions and peak O_3 concentration has been established using MLR analysis. Since MLR is a simple linear regression technique, it was frequently used to forecast O_3 concentration as well as other pollutants and weather parameters (Abdullah *et al.,* 2019; Napi *et al.,* 2021). Hence, from the above-mentioned studies, it can be proven that the O_3 concentration was best explained by the MLR.

Table-5. Results of model evaluation through performance indicators during daytime.

Site		Method	NAE	RMSE (ppb)	MAE (ppb)	IA	R^2
S1	O_{3t+1}	MLR	18.401	2.193	0.015	0.666	0.313
		PCR	2.396	0.368	0.003	0.306	0.220
		CMLR	17.276	2.047	0.014	0.744	0.313
	O_{3t+2}	MLR	19.058	2.297	0.016	0.076	0.265
		PCR	2.197	0.341	0.002	0.316	0.195
		CMLR	17.180	2.052	0.014	0.745	0.263
	O_{3t+3}	MLR	21.924	2.642	0.019	0.067	0.227
		PCR	2.453	0.386	0.003	0.299	0.162
		CMLR	20.333	2.435	0.017	0.746	0.225
	O_{3t+4}	MLR	20.658	2.483	0.017	0.071	0.217
		PCR	2.392	0.379	0.003	0.302	0.150
		CMLR	19.518	2.334	0.016	0.746	0.213
S2	O_{3t+1}	MLR	1.252	0.096	0.001	0.358	0.580
		PCR	4.586	0.187	0.001	0.313	0.419
		CMLR	1.252	0.096	0.001	0.358	0.580
	O_{3t+2}	MLR	1.766	0.146	0.001	0.309	0.492
		PCR	4.409	0.181	0.001	0.313	0.372
		CMLR	1.766	0.146	0.001	0.309	0.492
	O_{3t+3}	MLR	1.415	0.130	0.001	0.322	0.444
		PCR	4.320	0.179	0.001	0.315	0.348
		CMLR	1.294	0.109	0.001	0.345	0.443
	O_{3t+4}	MLR	1.201	0.113	0.001	0.338	0.424
		PCR	4.232	0.176	0.001	0.316	0.332
		CMLR	1.103	0.090	0.001	0.365	0.424
S3	O_{3t+1}	MLR	4.082	0.091	0.001	0.529	0.492
		PCR	5.946	0.166	0.001	0.291	0.343
		CMLR	2.691	0.071	0.001	0.510	0.492
	O_{3t+2}	MLR	3.903	0.107	0.001	0.542	0.450
		PCR	5.023	0.140	0.001	0.321	0.309
		CMLR	1.696	0.092	0.001	0.611	0.449
	O_{3t+3}	MLR	4.322	0.112	0.001	0.525	0.417
		PCR	4.425	0.125	0.001	0.345	0.287
		CMLR	2.438	0.090	0.001	0.579	0.411
	O_{3t+4}	MLR	2.194	0.096	0.001	0.047	0.388
		PCR	3.798	0.109	0.001	0.380	0.274
		CMLR	2.122	0.093	0.001	0.587	0.387
S4	O_{3t+1}	MLR	1.306	0.057	0.0004	0.527	0.514
		PCR	0.560	0.043	0.0003	0.541	0.328
		CMLR	0.295	0.022	0.0001	0.725	0.508
	O_{3t+2}	MLR	1.185	0.054	0.0003	0.576	0.408
		PCR	0.529	0.042	0.0002	0.538	0.276
		CMLR	1.419	0.058	0.0004	0.533	0.406
	O_{3t+3}	MLR	0.913	0.056	0.0003	0.570	0.310
		PCR	0.513	0.043	0.0002	0.529	0.214
		CMLR	1.367	0.060	0.0004	0.533	0.314
	O_{3t+4}	MLR	3.033	0.126	0.0009	0.324	0.257
		PCR	0.512	0.044	0.0003	0.505	0.173
		CMLR	2.297	0.096	0.0006	0.406	0.256

Table-6. Results of model evaluation through performance indicators during night-time.

Site		Method	NAE	RMSE (ppb)	MAE (ppb)	IA	R^2
S1	O_{3t+1}	MLR	$1.141E^{-2}$	0.537	0.004	0.165	0.389
		PCR	70.580	0.304	0.002	0.231	0.302
		CMLR	$1.141E^{-2}$	0.409	0.004	0.165	0.389
	O_{3t+2}	MLR	89.795	0.409	0.003	0.194	0.314
		PCR	61.382	0.260	0.002	0.245	0.228
		CMLR	89.795	0.409	0.003	0.194	0.314
	O_{3t+3}	MLR	98.289	0.457	0.003	0.181	0.276
		PCR	12.666	0.127	0.001	0.451	0.196
		CMLR	98.289	0.457	0.003	0.181	0.276
	O_{3t+4}	MLR	$1.00E^{-2}$	0.470	0.003	0.177	0.285
		PCR	8.691	0.107	0.001	0.487	0.190
		CMLR	4.262	0.063	0.0004	0.575	0.283
S2	O_{3t+1}	MLR	17.365	0.359	0.002	0.218	0.429
		PCR	4.276	0.158	0.001	0.361	0.267
		CMLR	16.315	0.336	0.002	0.226	0.429
	O_{3t+2}	MLR	28.527	0.606	0.004	0.159	0.340
		PCR	3.150	0.133	0.001	0.393	0.219
		CMLR	28.945	0.616	0.004	0.157	0.341
	O_{3t+3}	MLR	40.110	0.866	0.006	0.123	0.288
		PCR	2.199	0.111	0.001	0.422	0.188
		CMLR	40.110	0.866	0.006	0.123	0.288
	O_{3t+4}	MLR	3.375	0.120	0.001	0.402	0.256
		PCR	3.340	0.138	0.001	0.387	0.181
		CMLR	3.254	0.117	0.001	0.408	0.256
S3	O_{3t+1}	MLR	20.198	0.271	0.002	0.173	0.346
		PCR	6.455	0.074	0.001	0.307	0.190
		CMLR	18.325	0.279	0.002	0.151	0.346
	O_{3t+2}	MLR	18.639	0.280	0.002	1.014	0.286
		PCR	5.716	0.064	0.0004	0.317	0.137
		CMLR	18.639	0.280	0.002	0.152	0.286
	O_{3t+3}	MLR	0.773	0.030	0.0002	0.533	0.240
		PCR	5.331	0.059	0.0004	0.320	0.118
		CMLR	0.835	0.021	0.0001	0.609	0.240
	O_{3t+4}	MLR	1.792	0.054	0.0004	0.431	0.209
		PCR	3.799	0.043	0.0003	0.309	0.097
		CMLR	1.792	0.054	0.0004	0.431	0.209
S4	O_{3t+1}	MLR	21.145	0.151	0.001	0.235	0.436
		PCR	10.321	0.054	0.0003	0.265	0.384
		CMLR	21.145	0.151	0.001	0.235	0.436
	O_{3t+2}	MLR	64.027	0.436	0.003	0.101	0.410
		PCR	9.932	0.052	0.0004	0.268	0.358
		CMLR	64.648	0.440	0.003	0.100	0.410
	O_{3t+3}	MLR	72.263	0.494	0.003	0.092	0.350
		PCR	9.114	0.047	0.0003	0.274	0.307
		CMLR	71.501	0.489	0.003	0.924	0.350
	O_{3t+4}	MLR	53.610	0.370	0.003	0.118	0.301
		PCR	8.285	0.043	0.0003	0.279	0.276
		CMLR	53.610	0.370	0.003	0.118	0.301

CONCLUSIONS

In conclusion, the development of MLR model revealed the range R^2 was between 0.209 and 0.580, IA (0.047 and 5.76), RMSE (0.030 ppb and 2.642 ppb), NAE (0.913 and 98.289) and MAE (0.0003 ppb and 0.019 ppb) while for PCR and CMLR, R^2 range from 0.097 to 0.419, IA (0.231 to 0.541), RMSE (0.042 ppb to 0.386 ppb), NAE (0.0512 to 70.588), MAE (0.0002 ppb to 0.003 ppb) and R^2 (0.209 to 0.580), IA (0.100 to 0.924), RMSE (0.0.021 ppb to 0.866 ppb), NAE (0.835 to 98.289), MAE (0.0003 ppb to 0.017 ppb) respectively. In comparison to PCR and CMLR, the created MLR models are the most suitable and have the lowest errors for forecasting O_3 concentration to the next hours.

ACKNOWLEDGEMENT

We acknowledge the Malaysian Ministry of Higher Education by providing a Fundamental Research Grant Scheme (FRGS) (FRGS/1/2022/TK08/UMT/02/8) (VOT: 59716) for funding this study.

REFERENCES

Abdullah S., Hamid F. F. A., Ismail, M., Ahmed, A. N. and Mansor, W. N. D. W. 2019. Data on Indoor Air Quality (IAQ) in kindergartens with different surrounding activities. Data in Brief. 25: 103969.

Abdullah S., Nasir N. H. A., Ismail M., Ahmed A. N. and Jarkoni M. N. K. 2019. Development of Ozone Prediction Model in Urban Area. International Journal of Innovative Technology and Exploring Engineering. 8: 2263–2267.

Alqurashi T. and Wang T. 2018. Clustering Ensemble Method. International Journal of Machine Learning and Cybernetics. 10: 1227-1246.

Awang N. R., Ramli N. A., Yahaya A. S. and Elbayoumi M. 2015. High Night-Time Ground-Level Ozone Concentrations in Kemaman: NO and NO_2 Concentrations Attributions. Aerosol and Air Quality Research. 15: 1357-1366.

Baklanov A. and Zhang, Y. 2020. Advances in Air Quality Modelling and Forecasting. Global Transitions. 2: 261-270.

Cifuentes F., Gálvez A., González C. M., Orozco-Alzate M. and Aristizábal B. H. 2021. Hourly Ozone and $PM_{2.5}$ Prediction Using Meteorological Data-Alternatives for Cities with Limited Pollutant Information. Aerosol and Air Quality Research. 21: 200471.

D'Urso P., Massari R., Cappeli C. and De Giovanni L. 2017. Autoregressive Metric-Based Trimmed Fuzzy Clustering with An Application to PM_{10} Time Series. Chemometrics and Intelligent Laboratory Systems. 161: 15-26.

Fong S. Y., Abdullah M. and Ismail M. 2018. Forecasting of Particulate Matter (PM_{10}) concentration based on gaseous pollutants and meteorological factors for different monsoons of urban coastal area in Terengganu. Journal of Sustainability Science and Management. 5: 3-17.

Govender P. and Sivakumar V. 2020. Application of K-Means and Hierarchical Clustering Techniques for Analysis of Air Pollution: A Review (1980-2019). Atmospheric Pollution Research. 11: 40-56.

Hashim N. M., Noor N. M., Ul-Saufie A. Z., Sandu A. V., Vizureanu P., Deák G. and Kheimi M. 2022. Forecasting Daytime Ground-Level Ozone Concentration in Urbanized Areas of Malaysia Using Predictive Models. Sustainability. 14: 7936.

Laban T. L., Zyl P. G. V., Beukes J. P., Vakkari V., Jaars K., Borduas-Dedekind N. and Laakso L. 2018. Seasonal Influences on Surface Ozone Variability in Continental South Africa and Implications for Air Quality. Atmospheric Chemistry and Physics. 18: 15491-15514.

Leong I. I., Lou I., Ung W. K. and Mok K. M. 2015. Using Principal Component Regression, Artificial Neural Network, And Hybrid Models for Predicting Phytoplankton Abundance in Macau Storage Reservoir. Environmental Modeling & Assessment. 20: 355-365.

Mishra S. P., Sarkar U. Taraphder S., Datta S., Swain D., Saikhom R. and Laishram M. 2017. Multivariate Statistical Data Analysis-Principal Component Analysis (PCA). International Journal of Livestock Research. 7: 60-78.

Moursi A. S. A., El-Fishawy N., Djahel S. and Shouman M. A. 2022. Enhancing $PM_{2.5}$ Prediction Using NARX-Based Combined CNN and LSTM Hybrid Model. Sensors. 22: 4418.

Napi N., Abdullah S., Mansor A. A., Ahmed A. and Ismail M. 2021. Development of Models for Forecasting of Seasonal Ground Level Ozone (O_3). Journal of Engineering Science and Technology. 16: 3136-3154.

Nguyen H., Drebenstedt C., Bui X. N. and Bui D. T. 2020. Prediction of Blast-Induced Ground Vibration in An Open-Pit Mine by A Novel Hybrid Model Based on Clustering and Artificial Neural Network. Natural Resource Research. 29, 691-709.

Roy K. and Ambure P. 2016. The Double Cross-Validation Software Tool for MLR QSAR Model Development. Chemometrics and Intelligent Laboratory Systems. 159: 108-126.

Silva R. C. and Pires J. C. 2022. Surface Ozone Pollution: Trends, Meteorological Influences, and Chemical Precursors in Portugal. Sustainability. 14: 2383.

Stolz T., Huertas M. E. and Mendoza A. 2020. Assessment Of Air Quality Monitoring Networks Using an Ensemble Clustering Method In The Three Major Metropolitan Areas Of Mexico. Atmospheric Pollution Research. 11: 1271-1280.

Verma N., Satsangi A., Lakhani A. and Kumari K. M. 2015. Prediction of Ground Level Ozone Concentration in Ambient Air Using Multiple Regression Analysis. Journal of Chemical, Biological and Physical Sciences. 5: 3685-3696.

Warmiński K. and Bęś A. 2018. Atmospheric Factors Affecting a Decrease in the Night-Time Concentrations of

Tropospheric Ozone in a Low-Polluted Urban Area. Water, Air, & Soil Pollution. 229: 1-13.

Whalley J. and Zandi S. 2016. Particulate Matter Sampling Techniques and Data Modelling Methods. In (Ed.), Air Quality - Measurement and Modeling. IntechOpen 2016.

Zuśka Z., Kopcińska J., Dacewicz E., Skowera B., Wojkowski J. and Ziernicka-Wojtaszek A. 2019. Application of the Principal Component Analysis (PCA) Method to Assess the Impact of Meteorological Elements on Concentrations of Particulate Matter (PM_{10}): A Case Study of the Mountain Valley (The Sącz Basin, Poland). Sustainability. 11: 6740.