Global J. Environ. Sci. Manage. 8(1): 27-44, Winter 2022, Serial #29

Global Journal of Environmental Science and Management
(GJESM)

Homepage: https://www.gjesm.net/

ORIGINAL RESEARCH PAPER

Forecasting particulate matter concentration using nonlinear autoregression with exogenous input model

M.I. Rumaling[1], F.P. Chee[1,*], H.W.J. Chang[2], C.M. Payus[1], S.K. Kong[3], J. Dayou[4], J. Sentian[1]

[1]*Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*

[2]*Preparatory Centre for Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*

[3]*Department of Atmospheric Sciences, National Central University, Taoyuan, 32001, Taiwan*

[4] *Energy, Vibration and Sound Research Group, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*

ARTICLE INFO

Article History:
Received 15 April 2021
Revised 10 July 2021
Accepted 08 August 2021

Keywords:
Artificial neural network (ANN)
Nonlinear autoregression with exogenous input (NARX)
Principal component analysis (PCA)
Rotated component matrix
Scree plot

ABSTRACT

BACKGROUND AND OBJECTIVES: Air quality in some developing countries is dominated by particulate matter, especially those with size 10 micrometers and smaller or PM_{10}. They can be inhaled and sometimes can get deep into lungs; some may even get into bloodstream and cause serious health problems. Therefore, future PM_{10} concentration forecasting is important for early prevention and in urban development planning, which is crucial for developing cities. This paper presents the development of PM_{10} forecasting model using nonlinear autoregressive with exogenous input model.

METHODS: To improve performance of nonlinear autoregressive with exogenous input model, principal component analysis is used prior to the model for variable selection. The first stage of principal component analysis involves Scree plot, which determines the number of principal components based on explained variance. This is then followed by selecting variables using a rotated component matrix, based on their strength of contribution towards variation of PM_{10} concentration. To test the model, PM_{10} data in Kota Kinabalu from 2003 – 2010 was used. Neural network models are developed using this data by varying number of input variables with the inclusion of temporal variables. The developed forecasting models are evaluated using data PM_{10} in the city from 2011 to 2012. Four performance indicators, namely root mean square error, mean absolute error, index of agreement and fractional bias are reported.

FINDINGS: Results from principal component analysis show that five variables including wind direction index, relative humidity, ambient temperature, concentration of nitrogen dioxide and concentration of ozone strongly contribute to the variation of PM_{10} concentration. By using these variables together with temporal variables as input in the nonlinear autoregressive with exogenous input models, the resultant model shows good forecasting performance, with root mean square error of 7.086±0.873 μg/m^3. The selection of significant variables helps in reducing input variables inside the forecast model without degrading its forecast performance.

CONCLUSION: This model shows very promising performance in forecasting PM_{10} concentration in Kota Kinabalu as it requires fewer input variables and does not require variable transformation.

DOI: 10.22034/gjesm.2022.01.03

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
49	8	7

*Corresponding Author:
Email: *fpchee06@gmail.com*
Phone: + 6016 8607 582
ORCID: 0000-0002-9782-5572

Note: Discussion period for this manuscript open until April 1, 2022 on GJESM website at the "Show Article.

INTRODUCTION

Particulate matter (PM) is one of the components that causes air pollution, along with other gaseous pollutants. In urban areas, substances such as metals, elemental carbon and organic matters make up PM, which can enter human respiratory system and causes various health problems depending on its size and composition (Kim *et al.*, 2015; Ul-Saufie *et al.*, 2013). The health effect is more apparent on children and infants compared to other age groups (Karri *et al.*, 2018). Due to negative health effects of PM_{10}, it has received much attention from scientific community in recent years. PM_{10} is constantly monitored around the world. In Malaysia, more than 50 air quality monitoring stations across the country collect and store hourly meteorological and air pollutant data, including PM_{10} under Continuous Air Quality Monitoring (CAQM) network (Dominick *et al.*, 2012). These stations are operated by Alam Sekitar Sdn. Bhd., under administration of Department of Environment (DOE). While Kota Kinabalu is continuously developing, public transportation is still poor in terms of efficiency, reliability, safety and availability (Besar *et al.*, 2020). Many commuters prefer private vehicles rather than public transportation, which consequently causes traffic congestion, especially during peak hours. High traffic density, especially in commercial areas, lead to high emission of air pollutant, mainly PM_{10} (Ul-Saufie *et al.*, 2013). Local authorities are required to manage public transportation and infrastructure in Kota Kinabalu. Furthermore, Kota Kinabalu is a developing city as certain roads and housing areas are currently under development in most areas. Therefore, it is essential to study the long-term forecast model. The forecasting and prediction model is developed as early preventive measures to curb negative impacts of PM_{10} on health, environment, and economy. Many statistical approaches have been employed in developing PM_{10} forecasting and prediction model. The most widely used method in model development is regression analysis, particularly the multiple linear regression (MLR) (Abdullah *et al.*, 2016; Özbay *et al.*, 2011; Ul-Saufie *et al.*, 2013). MLR is relatively simple and does not require data from past research (Shahraiyni and Sodoudi, 2016). Despite the popularity, MLR suffers high multicollinearity in which the predictor variables are highly correlated towards each other (Abdul Wahab *et al.*, 2005). Principal component regression (PCR) solves the problem by applying principal component analysis (PCA) before MLR (Gvozdić *et al.*, 2011). This is because PCA converts original predictor variables into principal components (PCs), reducing dimensionality in the process (Polat and Gunay, 2015). PCR is best suited for linear systems (Shahraiyni and Sodoudi, 2016). This does not consider non-linear relationship such as the behaviour of PM_{10} at different humidity levels (Lou *et al.*, 2017). To consider non-linear relationship involving PM_{10} concentration, artificial neural network (ANN) is applied in developing PM_{10} concentration forecast model. ANN is a machine learning technique for model development inspired by biological neural network (Franceschi *et al.*, 2018). Just like the human brain, ANN obtains relationship between input and output variables based on available data (Arhami *et al.*, 2013). Artificial neurons in ANN are connected by synaptic weights. Data propagates through neurons by passing through summation of input-weight product and activation function (Elangasinghe *et al.*, 2014), as shown in Fig. 1.

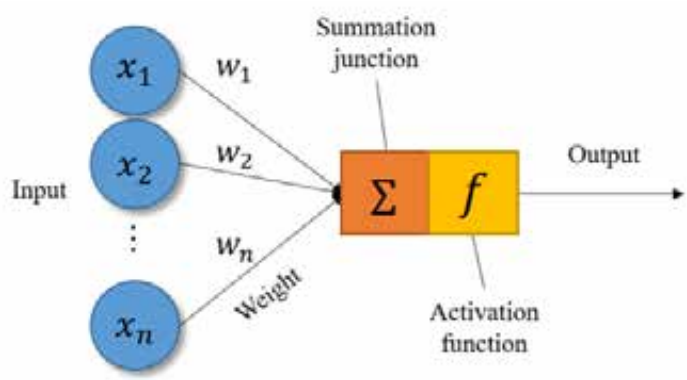


Fig 1: Basic parts of artificial neuron
(Ul-Saufie *et al.*, 2013)

Table 1: Several PM_{10} concentration prediction research using ANN model in Malaysia from 2011

Author (Year)	Study area	Best development method	RMSE (μg/m^3)
Ul Saufie *et al.* (2011)	Seberang Perai, Pulau Pinang	FFBP	8.369
Ul-Saufie *et al.* (2013)	Nilai, Negeri Sembilan	PCA-ANN and PCA-MLR	18.269
Ul-Saufie *et al.* (2015)	Seberang Perai, Pulau Pinang	FFBP	18.823
Abdullah *et al.* (2018)	Pasir Gudang, Johor	RBF with spread number of 0.3 and 12 hidden neurons	41.067
Abdullah *et al.* (2019)	Kuantan, Pahang	MLP	5.580

Several studies had focused on building and evaluating models to predict PM_{10} concentration in Malaysia, especifically in Peninsular Malaysia. Some of the predictions include long-term and short-term forecasting. Usage of ANN in these studies is more embraced in PM_{10} concentration forecast studies. Ul-Saufie *et al.,* (2011) studied on developing a prediction model of PM_{10} concentration in Seberang Perai, Pulau Pinang using MLR and feedforward backpropagation (FFBP) neural network. The prediction result showed that the FFBP neural network outperforms MLR due to lower RMSE of the FFBP neural network (8.369 μg/m3) compared to the MLR model (9.938 μg/m^3). Ul-Saufie *et al.,* (2013) then conducted another study on developing daily PM_{10} concentration prediction for Nilai, Negeri Sembilan. Prediction models are developed using MLR and ANN, as well as incorporating both methods with PCA. The result showed that models that incorporate PCA have the best prediction performance, with PCA-ANN having the lowest RMSE for next day prediction (11.1071 μg/m^3) and PCA-MLR having the lowest RMSE for the next two-day (RMSE = 14.4758 μg/m^3) and three-day prediction (RMSE = 18.2686 μg/m^3). Then, another study conducted by Ul-Saufie *et al.,* (2015) compared the performance of FFBP neural network and general regression neural network (GRNN) in predicting hourly PM_{10} concentration for the next three days in Seberang Jaya, a suburban located in Pulau Pinang. It is proven that FFBP generally performed better than GRNN in predicting the next three days of PM_{10} concentration. Abdullah *et al.,* (2018) developed a daily PM_{10} concentration forecast model for monitoring stations in Pasir Gudang, Johor using the radial basis function (RBF) model. While RBF model shows good performance during training, its performance significantly plummets during testing. In the following year, Abdullah *et al.* (2019) conducted another study developing a PM10 concentration forecast model in Kuantan, Pahang using the multilayer perceptron (MLP) model, with varying numbers of neurons in the hidden layer and activation function. Although studies on predicting future PM_{10} concentration in Malaysia was conducted extensively in the past years, the development and evaluation of models have not been conducted in Sabah yet. Table 1 presents several PM_{10} concentration prediction research using ANN model in Malaysia up to 2011.

This study aims to present performance evaluation results for forecast model of PM_{10} concentration in Kota Kinabalu, Sabah in Malaysia. $PM_{2.5}$ is not focused in this study as monitoring stations across Malaysia do not measure $PM_{2.5}$ concentration as of the year 2012. Nonlinear autoregressive with exogenous input (NARX) network was used in several studies in forecasting future PM_{10} concentration data with various sets of inputs (Abdulkadir and Yong, 2014; Saxena and Mathur, 2017; Vijayaraghavan and Mohan, 2016). Due to its performance, NARX network was used in this study. Models were tested for forecasting performance with different sets of input variables along with principal component analysis (PCA) as the accompanying method. This study was carried out in monitoring station in Kota Kinabalu of Malaysia in 2020.

MATERIALS AND METHODS

Study area and location

Kota Kinabalu (5.98°N, 116.07°E) is the capital city of Sabah, located at west coast of North Borneo at an altitude of 13 m above sea level, as shown in its map location in Fig. 2. Based on Köppen climate classification, Kota Kinabalu is categorized under tropical rainforest climate (Chang *et al.*, 2018). Kota Kinabalu experiences a hot and humid climate and seasonal circulation of monsoons (Djamila *et al.*, 2011). It is the busiest cities in Sabah as activities such as industry, trading and tourism are concentrated

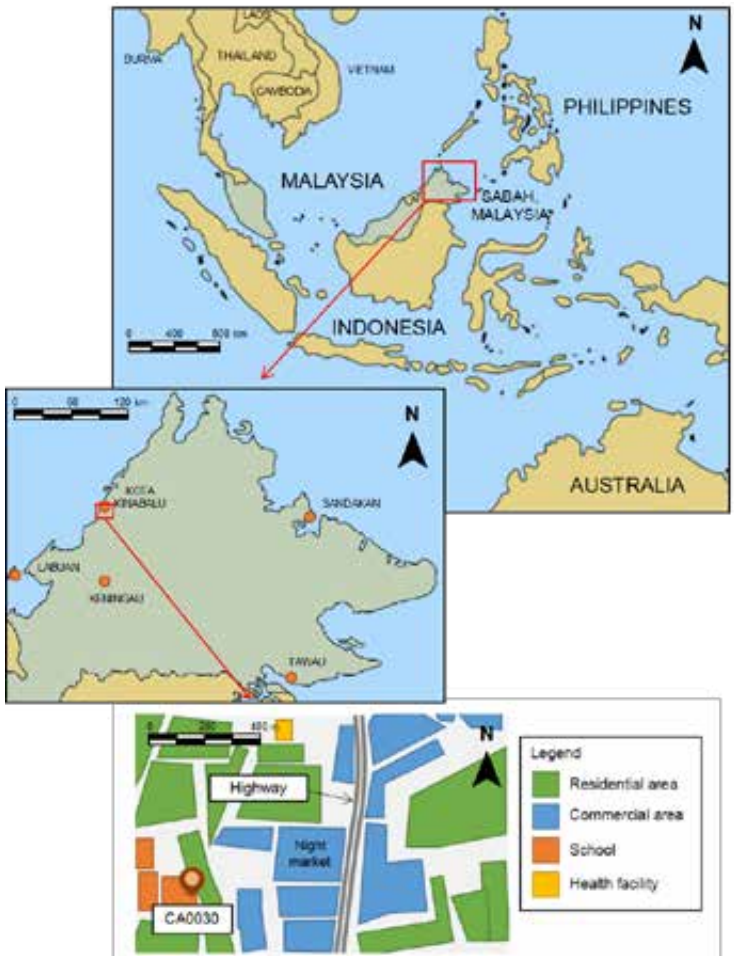


Fig. 2: Geographic location of the study area in Kota Kinabalu, Malaysia

here (Noor *et al.*, 2014). CA0030 is located in Putatan district, 10 km away from Kota Kinabalu city. Several buildings and infrastructure such as night markets and highway are the possible sources of PM_{10} measured by CA0030.

One of the monitoring stations in CAQM network, namely CA0030, collects meteorological and air pollutant data in Kota Kinabalu. CA0030 is located in the vicinity of SMK Putatan, approximately 10 km from Kota Kinabalu. This monitoring station collects four meteorological parameters namely wind speed (WS), wind direction (WD), relative humidity (RH) and ambient temperature (Temp). It also collects five concentrations of air pollutants which are carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), sulphur dioxide (SO_2) and PM_{10}. These data are collected at 1-hour interval. 10-year data ranging from 2003 to 2012 were used in this study. Table 2 highlights the descriptive statistics of the PM_{10} data used.

Table 2: Descriptive statistics of PM_{10} concentration in Kota Kinabalu from 2003 to 2012

Descriptive statistics	Values
Mean (μg/m^3)	35.90
Standard Deviation (μg/m^3)	19.10
Skewness	2.32
Kurtosis	16.59
Minimum (μg/m^3)	5
Maximum (μg/m^3)	495
1st Quartile (μg/m^3)	24
Median (μg/m^3)	33
3rd Quartile (μg/m^3)	42

Data preparation

Pre-processing of data is required before it is applied for any calculation or forecast modelling. The pre-processing work includes converting WD to wind direction index (WDI), converting hourly data to daily data, adding day of year (DOY) and month of year (MOY), and removing missing data. Raw data includes WD has discontinuity at 360°. The magnitude of wind direction does not reflect strength of wind itself. To remove discontinuity at 360°, WD was converted into a new variable namely WDI. WD was converted using Eqs. 1 and 2 (Vlachogianni *et al.*, 2011):

$$WDI = 1 + \sin\left(WD - \varphi\right) \quad (1)$$

$$\varphi = WD_{max} - 90° \quad (2)$$

WD_{max} in Eq. 2 represents wind direction at maximum PM_{10} concentration, while ö in Eq. 1 is the angle shift caused by WD_{max}. Both values can be obtained from polar plot of PM_{10} concentration against wind direction. Average values of PM_{10} concentration for every direction of 16 wind compass points are used for polar plots. Long-term forecasting model is defined as a model that uses dataset with temporal resolution of at least one day (Shahraiyni and Sodoudi, 2016). To develop the model for PM_{10} concentration, dataset obtained from the monitoring station was converted from hourly to daily temporal resolution. Arithmetic mean was used as 24-hour average value for all variables. As for WD, the average value was calculated using circular mean, which can be calculated using Eq. 3. "atan2" is a MATLAB function that returns the value ranging from 0° to 360°. This function is different from inverse tangent which only returns values in quadrant I (angles below 90°) and IV (angles above 270°). To preserve variation of WDI dataset, Eq. 3 was applied before WD is converted into WDI. 24 h dataset that consists of only missingness (CAL or N/A) is denoted as missing daily data.

$$WD_{daily} = \operatorname{atan2}\left(\frac{1}{n}\sum_{i=1}^{n}\sin WD_i, \frac{1}{n}\sum_{i=1}^{n}\cos WD_i\right) \quad (3)$$

PM_{10} concentration in Kota Kinabalu exhibits daily and yearly variations (Muhammad Izzuddin *et al.*, 2019). In order to take temporal variations into account, temporal variables such as day of year (DOY) and month of year (MOY) are introduced. Temporal variables are calculated using Eqs. 4 and 5, where d_{th} represents integer day in a year, T represents number of days in a year, and m_{th} represents integer month in a year. For example, 1st February is represented by d_{th} = 32 and m_{th} = 2. Study by Arhami *et al.* (2013) claimed that the consideration of temporal variables improves forecast performance of model.

$$DOY = \cos\left(\frac{2\pi d_{th}}{T}\right) \quad (4)$$

$$MOY = \cos\left(\frac{2\pi m_{th}}{12}\right) \quad (5)$$

Missing data leads to error in estimation during forecasting PM_{10} concentration (Cabaneros *et al.*, 2017). Previous studies showed that nearest neighbour method (NNM) imputes data with better performance compared to expectation-maximization (EM) algorithm (Muhammad Izzuddin *et al.*, 2020). This method applies for Fourier analysis which requires continuous stream of dataset. In this study, missingness is removed instead in forecast model development as data imputation may distort variation and correlation when used incorrectly (Graham, 2009). Exclusion of missingness is appropriate as only 6% of data are missing.

Principal component analysis (PCA)

Principal component analysis (PCA) is an assisting method in time series forecasting that transforms input variables X into a new set of variables known as principal components (PCs) (Gvozdić *et al.*, 2011). PCA accompanies ANN by reducing the complexity of the neural network model by determining relevant inputs in forecasting future PM_{10} concentration (Ul-Saufie *et al.*, 2013). Several studies show that PCA improves forecasting performance of ANN model (Azid *et al.*, 2014; Cabaneros *et al.*, 2017; Ul-Saufie *et al.*, 2013; Voukantsis *et al.*, 2011). PCA converts input variables into PCs by evaluating loading factor l in a way that every PCs are orthogonal to each other (Ul-Saufie *et al.*, 2013). The relationship between X, l and PC is given in Eq. 6 (Dominick *et al.*, 2012).

$$PC_i = \sum_{j=1}^{n} l_{ji} X_i \quad (6)$$

In this paper, PCA was executed using SPSS software version 25.0. Dataset consisting of meteorological variables (WS, WDI, RH, Temp) and air pollutant variables (CO, NO, O_3, SO_2) are fed into PCA. PM_{10} is not included in PCA as only input variables are considered in the selection of variables (Gvozdić *et al.*, 2011). Apart from loading factor 1 in the form of component matrix, PCA produces eigenvalue that describes the significance of each of the PCs towards PM_{10} concentration. Eigenvalue is plotted against respective PCs in Scree plot along with Kaiser criterion which states that only eigenvalue above 1 is selected (Azid *et al.*, 2014). Kaiser criterion implies that the variation of such PCs is considered to be significant (Franceschi *et al.*, 2018). The component matrix contains factor loadings describes the strength of variables in contribution towards variation of certain PCs (Dominick *et al.*, 2012). To better interpret factor loadings in this component matrix, Varimax rotation is applied to convert the loading factor to Varimax factors (VF) by raising the value of more significant loadings and lowering the value of smaller loadings (Azid *et al.*, 2014). The VFs reflect the strength of the contribution of a variable towards particular PCs and similarity of one variable towards the other (Voukantsis *et al.*, 2011). A variable has strong contribution when its VF has value larger than 0.75, moderate for 0.50 – 0.75, and weak for 0.30 – 0.49 (Dominick *et al.*, 2012).

Development of ANN model

Nonlinear autoregression with exogenous input (NARX) is a type of neural network model that falls under recurrent neural network (RC-NN), whose signal is processed via feedforward and is back-propagated into input level (Biancofiore *et al.*, 2017). NARX is used in air quality forecast research by using observed values from the past to predict PM_{10} concentration data in several time steps ahead (Vijayaraghavan and Mohan, 2016). Exogenous input in NARX model is fed with observed values of meteorological and air pollutant data. Previous research study has shown that NARX model has better forecasting performance of PM_{10} concentration compared to other models such as feed-forward (FF) neural network. NARX model is employed in developing forecast model in this research and its topology is illustrated in Fig. 3, where X represents exogenous input and Y represents PM_{10} concentration data. NARX has a feedback loop which sends output data back to input level. A variation of NARX model is known as nonlinear autoregression (NAR), in which the forecasting of PM_{10} concentration data depends only on past values of itself (Potdar and Pardawala, 2017).

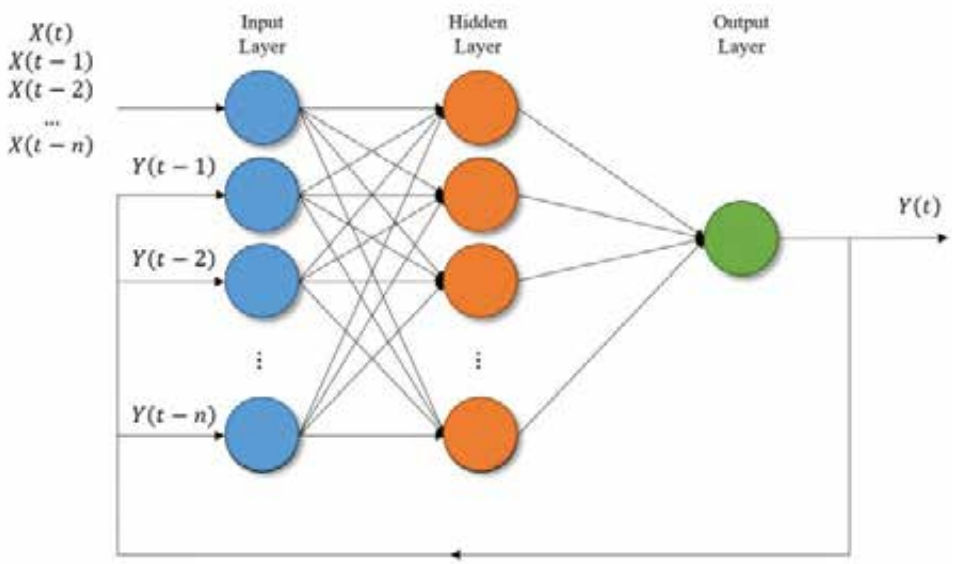


Fig. 3: NARX neural network architecture
(Abdulkadir and Yong, 2014)

Dataset from 2003 to 2010 (8-year data) are used to develop open-loop NAR and NARX models using MATLAB version 2018b. The 8-year data is further divided into 3 smaller sets which are training set (70% of 8-year dataset), validation set (15%) and testing set (15%). The proportion is MATLAB default setting and has been used in several previous studies (Cabaneros *et al.*, 2017; Ceylan and Bulkan, 2018; Feng *et al.*, 2015; Shekarrizfard *et al.*, 2012). Neurons at hidden and output layers have their activation function set as tansig (hyperbolic tangent) and purelin (identity linear), respectively. Hidden layer uses tansig as activation function because it can produce normalized values at both positive and negative ranges (Ul-Saufie *et al.*, 2013). As for the output layer, purelin activation function is used to optimize model performance (Wu *et al.*, 2019). NAR and NARX models are trained using Levenberg-Marquardt (LM) algorithm because it trains model at relatively short amount of time and guarantees convergence (Yu and Wilamowski, 2016). A total of 10 forecast models are developed to forecast PM_{10} concentration data from 2011 to 2012 (2 years of data), by varying inputs (U, M, G, P and S) in 5 ways and include or exclude temporal variables (MOY and DOY) for each way. U (univariate) model uses only PM_{10} concentration as inputs. M (meteorological) model includes four meteorological variables (WS, WD, RH, Temp) along with PM_{10} concentration data. G (gaseous) model further adds four air pollutant concentration variables (CO, NO_2, O_3, SO_2) along with meteorological variables and PM_{10} concentration. P (principal component) model uses PC scores obtained from PCA as inputs instead of original variables. Finally, S (selection) model uses only certain original variables selected by Scree plot and rotated component matrix. These models are labelled by two characters, in which the first character denotes variables used as neural network model inputs (U, M, G, P and S) and second character denotes inclusion of temporal variables as inputs (0: exclude temporal variables, 1: include temporal variables). 10 replicates shown in Table 3 together with their properties were developed for each model to verify their stability. N_s (number of selected variables) is used in Table 3 because it depends on result from rotated component matrix. The number

Table 3: List of developed forecast models

Model	Number of external inputs	Inclusion of temporal variables	Network type	Number of hidden neurons
U0	0	No	NAR	1
U1	2	Yes	NARX	5
M0	4	No	NARX	9
M1	6	Yes	NARX	13
G0	8	No	NARX	17
G1	10	Yes	NARX	21
P0	8	No	NARX	17
P1	10	Yes	NARX	21
S0	N_s	No	NARX	$2N_s + 1$
S1	$N_s + 2$	Yes	NARX	$2N_s + 5$

Table 4: Input variables for development of forecast models

Model	Input variables
U0	PM_{10}
U1	PM_{10}, MOY, DOY
M0	PM_{10}, WS, WDI, RH, Temp
M1	PM_{10}, WS, WDI, RH, Temp, MOY, DOY
G0	PM_{10}, WS, WDI, RH, Temp, CO, NO_2, O_3, SO_2
G1	PM_{10}, WS, WDI, RH, Temp, CO, NO_2, O_3, SO_2, MOY, DOY
P0	PM_{10}, PC_1, PC_2, PC_3, PC_4, ..., PC_8
P1	PM_{10}, PC_1, PC_2, PC_3, PC_4, ..., PC_8, MOY, DOY
S0	Significant variables from PCA
S1	Significant variables from PCA, MOY, DOY

of hidden neurons is set as 2N + 1 (N represents number of external inputs) following the study by Cabaneros *et al.* (2017) because it does not require determination of number of hidden neurons. The input variables for each model are listed in Table 4.

Evaluation of NAR and NARX models

To assess the forecast performance of NAR and NARX models, a set of performance indicators are used, namely root mean square error (RMSE), mean absolute error (MAE), index of agreement (IA) and fractional bias (FB). RMSE and MAE indicates accuracy of forecast model and are frequently used in many studies (Antanasijević *et al.*, 2013; Díaz-Robles *et al.*, 2008; Feng *et al.*, 2015; Grivas and Chaloulakou, 2006; Paschalidou *et al.*, 2011; Wu *et al.*, 2019). Both RMSE and MAE show better accuracy as their values approach zero. While RMSE tends to change with frequency distribution of error, MAE only depends on average magnitude of error (Willmott and Matsuura, 2005). IA expresses the difference between predicted and observed values, indicating the agreement of both datasets as the name suggests (Fan *et al.*, 2013). IA ranges from 0 to 1 with better agreement indicated by higher performance. FB indicates underestimation and overestimation of forecast model (Biancofiore *et al.*, 2017). FB ranges between -2 to 2 where the boundary levels indicate extreme underestimation and overestimation respectively. These four performance indicators are evaluated using Eqs. 7 to 10, where P denotes forecasted value and O indicates observed value. Both $\overline{P}$ and $\overline{O}$ are mean values of forecasted and observed values, respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(O_i - P_i\right)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|O_i - P_i\right| \tag{8}$$

$$IA = 1 - \frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{\sum_{i=1}^{n}\left(\left|O_i - \overline{O}\right| + \left|P_i - \overline{O}\right|\right)^2} \tag{9}$$

$$FB = \frac{2\left(\overline{P} - \overline{O}\right)}{\overline{P} + \overline{O}} \tag{10}$$

RESULTS AND DISCUSSION

Principal component analysis (PCA)

Eigenvalue is plotted as a function of PC in Scree plot as shown in Fig. 4. Red dashed line represents Kaiser's criterion, implying that only PCs with eigenvalues above this line are selected. Based on Fig. 4, the first three PCs were selected as their eigenvalues are above 1. Other principal components are neglected because of their eigenvalues below 1, implying redundancy with less important factors (Azid *et al.*, 2014).

Selected PCs then undergo Varimax rotation, in which factor loadings are converted into VF (Table 5). VFs with magnitude exceeding 0.75 and corresponding variables are highlighted in bold. The variables that

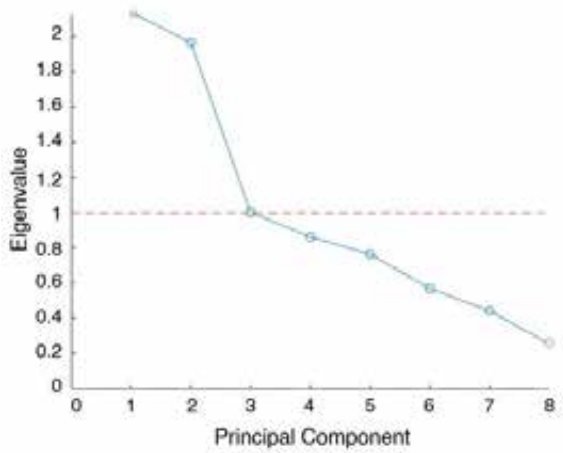


Fig. 4: Scree plot for PCA in CA0030

have VF above 0.75 in one of the PCs are WDI, RH, Temp, NO_2 and O_3, which becomes inputs for S models in neural network model. VFs with higher value imply that these variables have strong contribution towards variability of PM_{10} concentration for CA0030.

The first PC (PC_1) account for 26.70% of the total variation of PM_{10} concentration. It can be seen from Table 5 that O_3 concentration and ambient temperature (Temp in Table 5) shows strong positive contribution while relative humidity shows strong negative contribution towards PM_{10} concentration. This is closely correlated with meteorological condition in Kota Kinabalu. PM_{10} absorbs water vapour and becomes too heavy to stay suspended at high humidity level (Munir *et al.*, 2017), which is a normal condition at Kota Kinabalu. PM_{10} gains kinetic energy from heat to stay airborne, which explains the strong positive contribution in ambient temperature (Temp). Furthermore, abundance in solar radiation is received by Kota Kinabalu, ranging between 278.52 W/m^2 and 407.89 W/m^2 (Teong *et al.*, 2017). This induces generation of ground level O_3 (Xie *et al.*, 2015). High concentration of both O_3 and PM_{10} occur simultaneously during hot climate, which leads to strong positive contribution in O_3 concentration. PC_2 is related to motor vehicle emission and account for 24.57% of the total variation for PM_{10} concentration. In this PC_2, NO_2 concentration shows strong positive contribution while CO concentration shows moderate positive contribution, with VF of 0.648. This is because PM_{10} often accompanies NO_2 and CO as by-product of incomplete combustion by motor vehicles (Xie *et al.*,

Table 5: Rotated component matrix in CA0030

Variables	PC_1	PC_2	PC_3
WS	0.281	-0.695	-0.006
WDI	-0.014	-0.023	**0.987**
RH	**-0.870**	0.040	0.011
Temp	**0.830**	0.190	-0.038
CO	0.143	0.648	0.083
NO_2	0.040	**0.832**	0.022
O_3	**0.767**	-0.103	0.026
SO_2	0.025	0.566	-0.143
Variance explained (%)	26.70	24.57	12.52

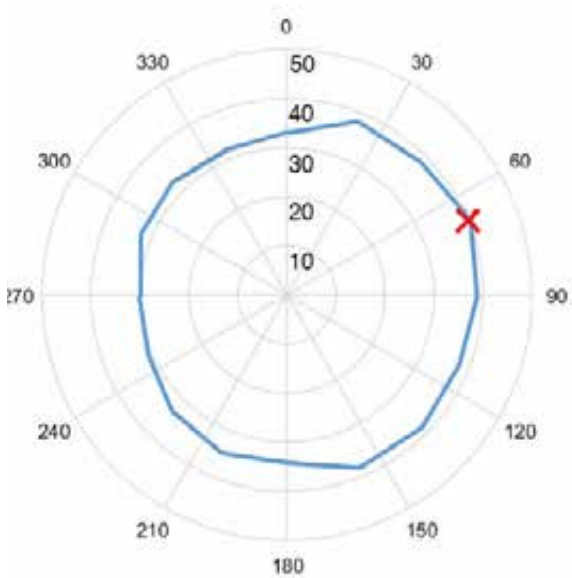
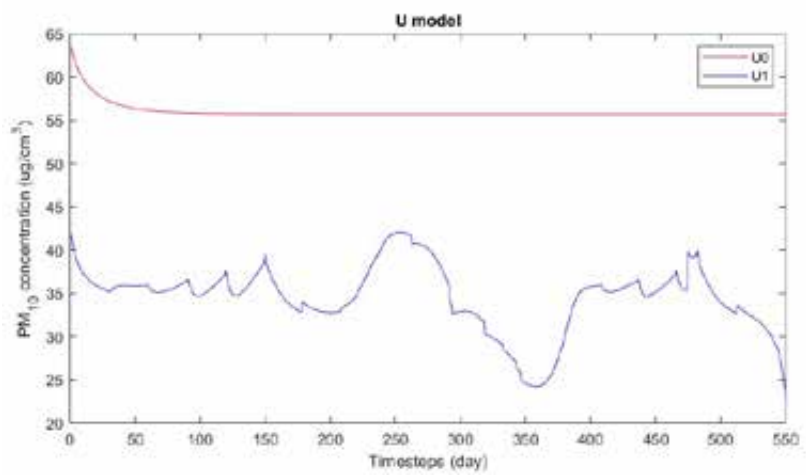
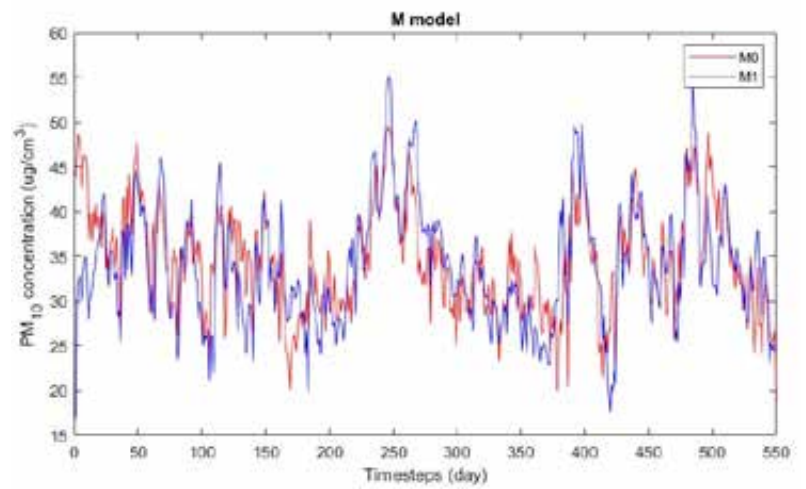


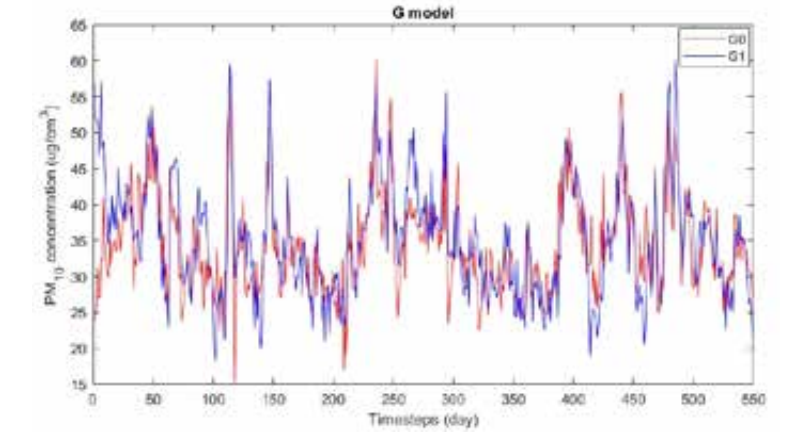
Fig. 5: Polar plot of PM_{10} concentration (radial axis) against WD (angular axis) for CA0030



(a)

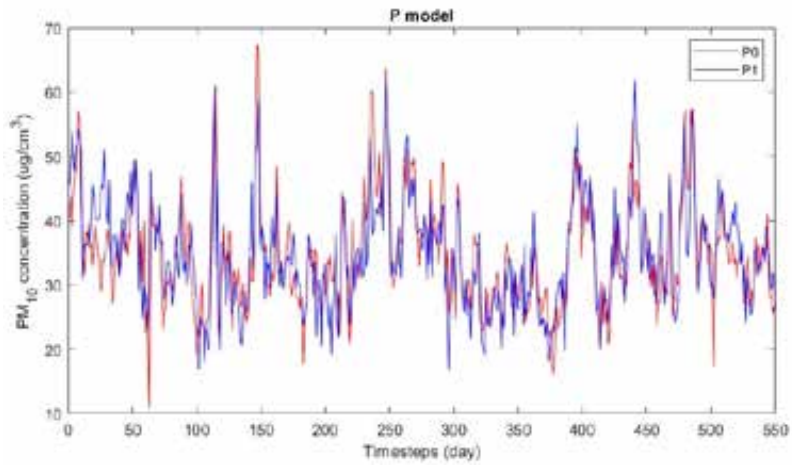


(b)

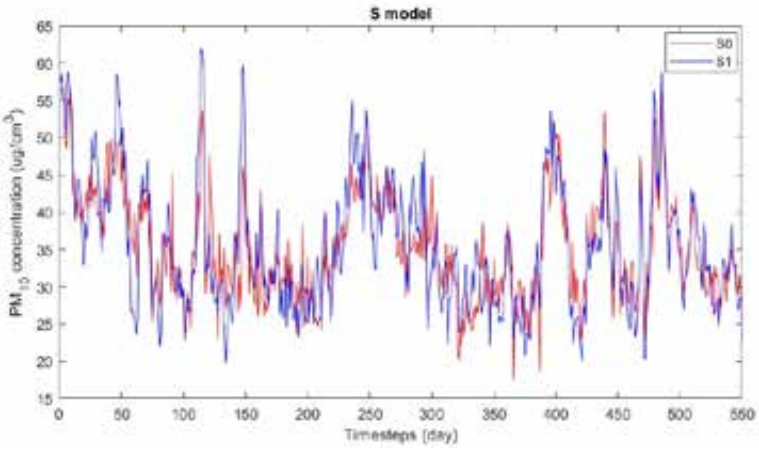


(c)

Fig. 6: Time series plot of forecasted data calculated using (a) U models, (b) M models, (c) G models, (d) P models, and (c) S models



(d)



(c)

Fig. 6: Time series plot of forecasted data calculated using (a) U models, (b) M models, (c) G models, (d) P models, and (c) S models

2015). VF for NO_2 concentration is higher compared to CO, suggesting that oxygen gas react with ambient nitrogen gas. As for PC_3, it accounts for 12.52% of the total variation. It can be seen from Table 5 that WDI shows strong positive contribution in this PC. This suggests that PM_{10} concentration is highest when wind is blown from direction of WD_{max}, which is determined to be 67.5° based on red cross mark in polar plot as shown in Fig. 5. Examining the location of the monitoring station CA0030 and its vicinity shown in the map in Fig. 2, the direction of WD_{max} is indicated by the area in blue with 22.5° wide. It can be observed that night markets and highway are located inside the area. Wind direction blowing from these buildings and infrastructures may also contribute to higher PM_{10} concentration.

Artificial neural network (ANN)

Ten neural network models were developed to forecast PM_{10} concentration in Kota Kinabalu from 2011 to 2012. The result for each forecasting model is plotted in time series graph as shown in Fig. 6. It can be seen that the U model clearly does not forecast the future trends of PM_{10} concentration in contrast to other models. U0 model for example shows a future trend of PM_{10} concentration with the value settles down to a certain constant value over time, which does not represent the actual trend of

Table 6: Number of days exceeding allowable PM_{10} concentrations as forecasted by NAR and NARX models

Model	Number of days exceeding allowable PM_{10} concentration
U0	550
U1	0
M0	0
M1	8
G0	19
G1	22
P0	26
P1	31
S0	22
S1	28
Actual	36

Table 7: Performance indicators (in terms of mean ± standard deviation) of neural network models for PM_{10} concentration in Kota Kinabalu

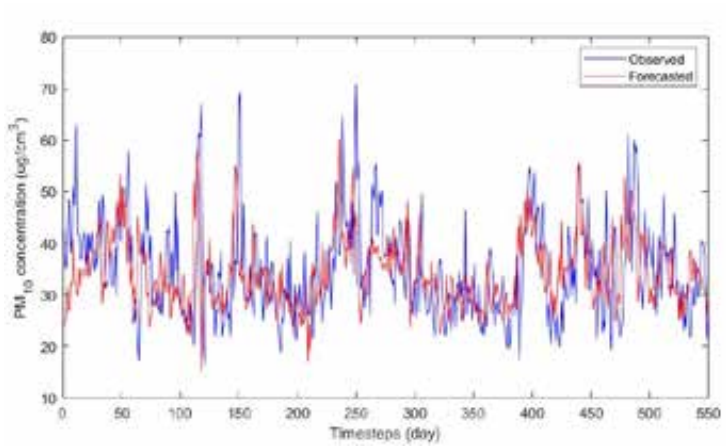
Model	RMSE (μg/m^3)	MAE (μg/m^3)	IA	FB
U0	9.292±0.072	7.343±0.113	0.145±0.046	-0.00002
U1	8.740±0.295	6.876±0.268	0.458±0.092	-0.0071
M0	7.596±0.741	5.833±0.590	0.701±0.130	0.0084
M1	7.062±0.668	5.445±0.507	0.772±0.068	-0.0015
G0	6.264±0.552	4.822±0.454	0.846±0.034	0.0056
G1	6.567±0.392	5.116±0.319	0.830±0.031	-0.0059
P0	6.236±0.526	4.850±0.421	0.844±0.034	-0.0039
P1	5.964±0.648	4.630±0.540	0.860±0.048	-0.0054
S0	7.109±0.450	5.500±0.365	0.779±0.041	0.0017
S1	7.086±0.873	5.350±0.523	0.812±0.034	-0.0156

PM_{10} concentration at all. Meanwhile, U1 model only forecasts rough trends of PM_{10} concentration. All other models including M, G, P and S show almost similar periodic pattern of PM_{10} concentration, in which the minima and maxima of PM_{10} concentration in dataset can be identified.

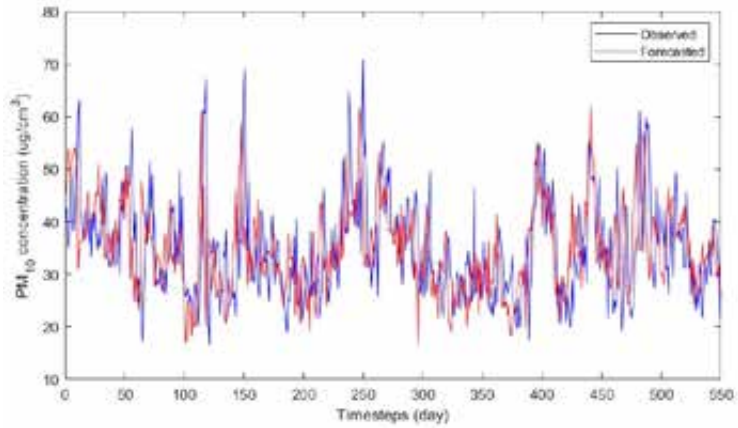
According to the Malaysian Ambient Air Quality Guidelines (MAAQG), the allowable 24-hour average PM_{10} concentration was set as 50 μg/m^3 before the year 2015. Based on Fig. 6(b) to (c), daily average PM_{10} concentration at certain timesteps exceeded the allowable limit as set in MAAQG. This is mainly due to unusually higher traffic density, more active night markets and also influence from wind direction. The trend of PM_{10} concentration events is not observed in Fig. 6(a) due to severe underfitting of U models as a result of limited number of neurons in hidden layer (Ceylan and Bulkan, 2018). Table 6 shows the number of days exceeding allowable PM_{10} concentration as forecasted by these models. Models P and S tend to forecast the number of days closer to the actual data. This shows that models P and S are able to forecast days exceeding allowable PM_{10} concentration with good accuracy.

The forecast performance (in terms of mean ± standard deviation) of all ten neural network models is tabulated in Table 7. None of the models show significant underestimation or overestimation as indicated by FB centred close to zero. U0 model shows severe underfitting, indicated by relatively low IA and constant trend over time as observed in Fig. 6. This is because U0 model only uses past values of PM_{10} concentration, which is too few variables used for forecasting, leading to failure in capturing the variability of time series dataset (Dotse *et al.*, 2018). This is also true for U1 which uses temporal variables in addition to past values of PM_{10} concentration.

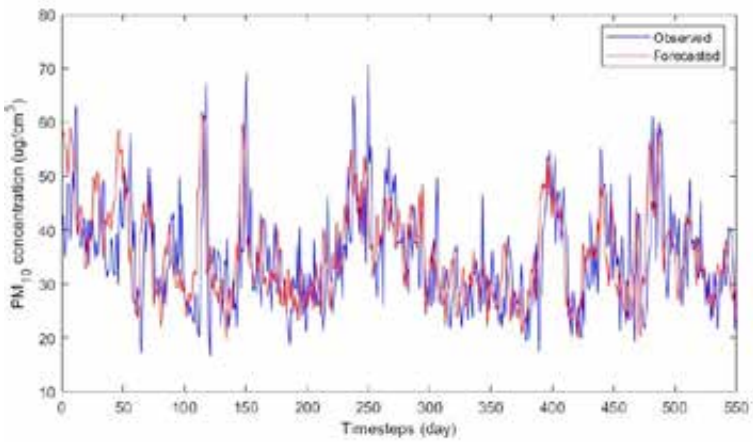
Inclusion of meteorological and gaseous variables as in M and G models significantly improves the performance of forecasting PM_{10} concentration as measured by CA0030. Neglecting FB (as no severe underestimation and overestimation occurs), G0 shows better forecasting performance (RMSE = 6.264±0.552 μg/cm^3, MAE = 4.822±0.454 μg/cm^3, IA = 0.846±0.034) compared to M models and U models. This can be seen in the verification shown in Fig. 7 that G0 model can forecast 2-year of PM10 concentration data more accurately. Slight performance degradation



(a)



(b)



(c)

Fig. 7: Time series plot of observed (blue) and forecasted (red) data by (a) G0 model, (b) P1 model, and (c) S1 model, for PM_{10} concentration in CA0030

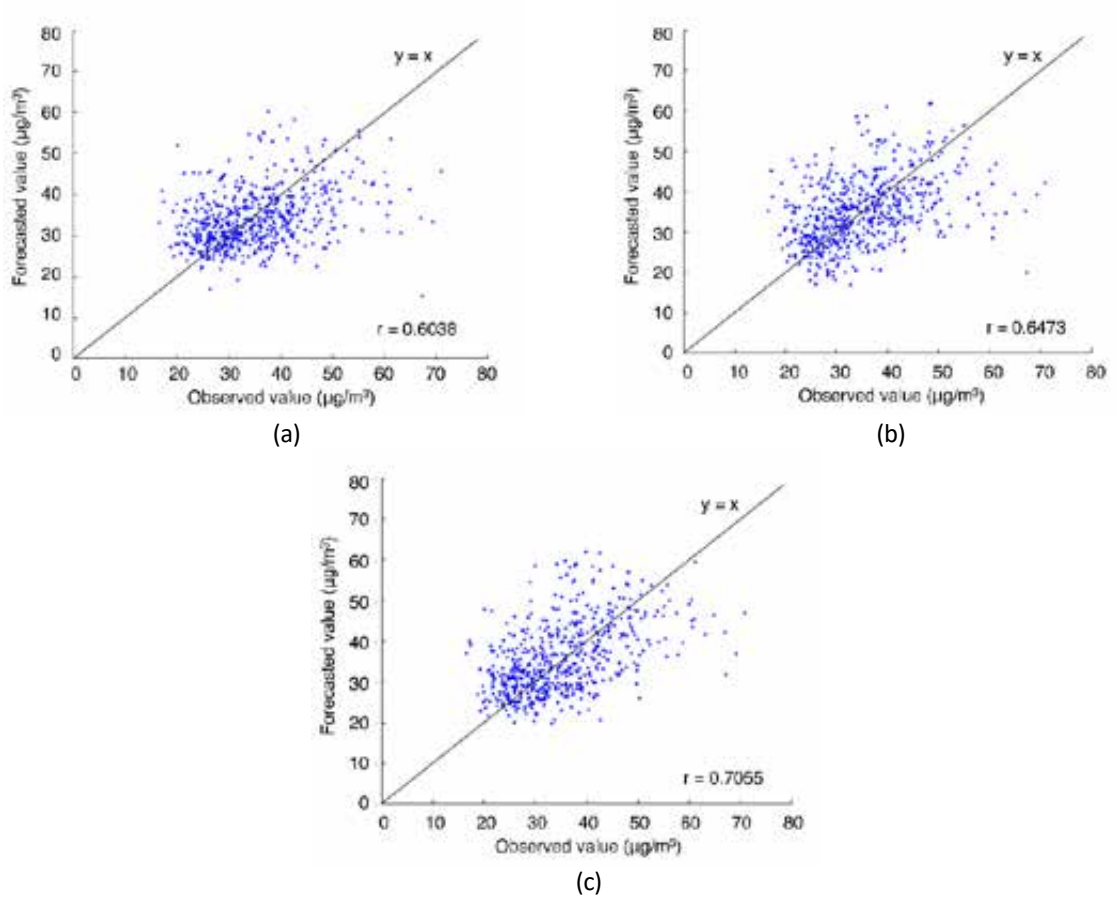


Fig. 8: Plot of forecasted value against observed value for (a) G0 model, (b) P1 model, and (c) S1 model

occurs for G1 model as indicated by higher RMSE and MAE, as well as low IA. This is attributed to overfitting of forecast model which occurs when too many input variables are used (Ceylan and Bulkan, 2018). Using PCs as input variables instead of original variables (as in P models) also improves performance of forecasting model (RMSE = 5.964±0.648 μg/m^3, MAE = 4.630±0.540 μg/m^3, IA = 0.860±0.048). It can be seen that unlike G models, inclusion of temporal variables (as in P1 model) does not cause overfitting because PCs are orthogonal and uncorrelated to each other (Ul-Saufie *et al.*, 2013). Higher standard deviation in the performance indicators for P1 model suggests that the performance of forecast model is less consistent with every forecast attempt. As for S models, the forecasting performance is not the best as compared to other models, which may be due to a smaller number of input variables used in forecasting. However, S1 model is capable of forecasting PM_{10} concentration with good performance (RMSE = 7.086±0.873 μg/m^3, MAE = 5.350±0.523 μg/m^3, IA = 0.812±0.034), as reflected by time series plot in Fig. 7(b) with relatively high IA as shown in Table 7. Fig. 8 reveals that there is moderately strong correlation between forecasted value from the three models (G0, P1 and S1) and observed value. Although the coefficient of correlation for the three models does not show strong correlation, the three models can still capture future trend of PM_{10} concentration as revealed by high IA values and time series plots in Fig. 7.

P1 model forecasts PM_{10} concentration in Kota Kinabalu with the best performance as shown in Table 7. This is reflected by its highest IA value while having the lowest RMSE and MAE compared to other models. This is possible because P1 model uses principal

components converted from all variables together with principal components. The performance indicators in RMSE and MAE and show that P1 has a relatively high standard deviation, suggesting that slight overfitting occasionally occurs during development. Overall, S1 model forecasts PM_{10} concentration with good performance as illustrated by time series plot in Fig. 7(b). Unlike G and P models, S1 does not require all input variables and PCA transformation to achieve good forecasting performance.

CONCLUSION

Based on PCA conducted in this research, rotated component matrix shows that WDI, RH, Temp, NO_2 and O_3 (from the first three PCs) strongly contributes to variation of PM10 concentration in 10 years (from 2003 to 2012). PC_1 suggests that meteorological condition and ground ozone generation strongly contributes to PM_{10} concentration variation. PC_2 concerns with motor vehicle emission, mainly reaction of oxygen gas and ambient nitrogen gas in engines. PC_3 is related to PM_{10} emission sourced from buildings and infrastructures, mainly night markets and highway located close to CA0030 monitoring station. As for forecast model developed using NAR and NARX, U models show severe underfitting, reflected by failure in forecasting PM_{10} concentration data accurately and is confirmed by low values of IA. Addition of more input variables in NARX model led to better forecasting performance, as indicated by M and G models. Among U, M and G models, the performance peaks at G0 model (RMSE = 6.264±0.552 μg/cm^3, MAE = 4.822±0.454 μg/cm^3, IA = 0.846±0.034), and overfitting occurs at G1 model as slightly too many input variables were used. Using principal components as inputs instead of original variables also show good performance as reflected by P models. Addition of temporal variables further improves the performance of forecast model, as in P1 (RMSE = 5.964±0.648 μg/m^3, MAE = 4.630±0.540 μg/m^3, IA = 0.860±0.048). While P1 model in overall shows the best forecast performance, it is not the most ideal model when used in real time application. Not only that it requires all variables in addition with variable transformation, but the standard deviation of performance indicators also indicate that the development of P1 model is not the most stable. S1 model is selected because it can forecast PM_{10} concentration with good performance (RMSE = 7.086±0.873 μg/m^3, MAE = 5.350±0.523 μg/m^3, IA = 0.812±0.034) without requiring all input variables and variable transformation. In real-time application, S1 model is preferred in forecasting PM_{10} concentration data for Kota Kinabalu as it requires fewer input variables to achieve accurate result. The method of model development is yet to be studied for other monitoring stations in Sabah. Furthermore, models with fewer input variables especially univariate (U0) models should be studied in order to achieve good forecasting results without requiring many input variables.

AUTHOR CONTRIBUTIONS

R. Muhammad Izzuddin is the main contributor to this paper. He carried out the data analysis and modelling; and drafted the manuscript. F.P. Chee participated in the design of this study and led this project. R. Muhammad Izzuddin, F.P. Chee and H.W.J. Chang conceived the original idea with many helpful suggestions from Sentian, J. Dayou and C.M. Payus participated in the study coordination and helped to review the manuscript. S.S.K. Kong assisted in the data collection from the Department of Environment.

ACKNOWLEDGEMENT

The authors would like to express gratitude towards Universiti Malaysia Sabah for supporting this study by providing grants [SBK0352-2018, SGI0054-2018 and GUG0378-2018] and the Department of Environment Malaysia for providing meteorological and pollutant data of this study.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

ABBREVIATIONS

0	without temporal variables (U, M, G, P and S models)
1	with temporal variables (U, M, G, P and S models)
ANN	Artificial neural network
CAQM	Continuous air quality monitoring
CO	Carbon monoxide
d_{th}	Integer day in a year
DOE	Department of Environment
EM	Expectation maximization
FB	Fractional bias
FF	Feedforward
FFBP	Feedforward backpropagation
G	Gaseous (ANN model with meteorological and gaseous input variables)
GRNN	General regression neural network
IA	Index of agreement
I	Factor loading
km	kilometre
LM	Levenberg-Marquardt
m	Metre
m^2	Metre Square
m_{th}	Integer month in a year
M	Meteorological (ANN model with meteorological input variables only)
MAAQG	Malaysian Ambient Air Quality Guidelines
MAE	Mean absolute error
MATLAB	Matrix laboratory
max	maximum value
MLP	Multilayer perceptron
MLR	Multiple linear regression
N	Number of input variables
NAR	Nonlinear autoregressive
NARX	Nonlinear autoregressive with exogenous input
NNM	Nearest neighbour method
NO_2	Nitrogen dioxide
O	Observed value
$\overline{O}$	Average observed value
O_3	Ozone
P	Predicted value
P	Principal (ANN model with PC as input variables)
$\overline{P}$	Average predicted value
PC	Principal component
PCA	Principal component analysis
PCR	Principal component regression
PM	Particulate matter
PM_{10}	Particulate matter with aerodynamic diameter below 10 microns
RBF	Radial basis function
RC-NN	Recurrent neural network
RH	Relative humidity
RMSE	Root mean square error
S	Selected (ANN model with selected variables based on rotated component matrix)
SO_2	Sulphur dioxide
T	Number of days in a year
Temp	Ambient temperature
U	Univariate (ANN model without meteorological or air quality input variables)
VF	Varimax factor
w	Synaptic weight
WD	Wind direction
WD_{max}	Wind direction at maximum PM_{10} concentration
WDI	Wind direction index
WS	Wind speed
W/m^2	Watt per squared metre
X	Input variable
$\mu g/m^3$	Microgram per cubic metre
φ	Shift in sine function

REFERENCES

Abdulkadir, S.J.; Yong, S.P., (2014). Empirical analysis of parallel-NARX recurrent network for long-term chaotic financial forecasting. International Conference on Computer and Information Sciences (ICCOINS). 1–6 **(6 pages).**

Abdullah, S.; Ismail, M.; Ghazali, N. A..; Ahmed, A.M.A.N., (2018). Forecasting particulate matter (PM10) concentration: A radial basis function neural network approach. AIP Conference Proceedings, 1–6 **(6 pages).**

Abdullah, S.; Ismail, M.; Ghazali, N. A..; Ahmed, A.M.A.N., (2019)., Multi-Layer Perceptron Model for Air Quality Prediction. Malaysian J. Math. Sci. 13: 85–90 **(6 pages).**

Abdullah, S.; Ismail, M.; Fong, S.Y.; Ahmed, A.M.A.N., (2016). Evaluation for long term PM10 concentration forecasting using multi linear regression (MLR) and principal component regression

(PCR) models. Environ. Asia. 9(2): 101–110 **(10 pages).**
Abdul-Wahab, S.A.; Bakheit, C.S.; Al-Alawi, S.M., (2005) Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Model. Software. 20(10): 1263–1271 **(9 pages).**
Antanasijević, D.Z.; Pocajt, V.V.; Povrenović, D.S.; Ristić, M.D.; Perić-Grujić, A.A., (2013). PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ., 443: 511–519 **(9 pages).**
Arhami, M.; Kamali, N.; Rajabi, M.M., (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environ. Sci. Pollut. Res., 20(7): 4777–4789 **(13 pages).**
Azid, A.; Juahir, H.; Toriman, M. E.; Kamarudin, M. K. A.; Saudi, A. S. M.; Hasnam, C.N.C.; Aziz, N.A.A.; Azaman, F.; Latif, M. T.; Zainuddin, S. F. M.; Osman, M. R.; Yamin, M., (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. Water Air Soil Pollut., 225(8): 2063–2076 **(14 pages)**.
Besar, S.N.A.; Ladin, M.A.; Harith, N.S.H.; Bolong, N.; Saad, I.; Taha, N., (2020). An overview of the transportation issues in Kota Kinabalu, Sabah.IOP Conference Series: Earth and Environ. Sci., 1–9 **(9 pages).**
Biancofiore, F.; Busilacchio, M.; Verdecchia, M.; Tomassetti, B.; Aruffo, E.; Bianco, S.; Di Tomasso, S.; Colangeli, C.; Rosatelli, G.; Di Carlo, P., (2017). Recursive neural network model for analysis and forecast of PM_{10} and $PM_{2.5}$. Atmos. Pollut. Res., 8: 652–659 **(8 pages).**
Cabaneros, S.M.L.S.; Calautit, J.K.S.; Hughes, B.R., (2017). Hybrid Artificial Neural Network Models for Effective Prediction and Mitigation of Urban Roadside NO_2 Pollution. Energy Procedia. 3524–3530 **(7 pages).**
Ceylan, Z.; Bulkan, S., (2018). Forecasting PM_{10} Levels using ANN and MLR: A case study for Sakarya City. Global Nest J., 20(2): 281–290 **(10 pages).**
Chang, H.W. J.; Chee, F.P.; Kong, S.K.S.; Sentian, J., (2018). Variability of the PM_{10} concentration in the urban atmosphere of Sabah and its responses to diurnal and weekly changes of CO, NO_2, SO_2 and Ozone. Asian J. Atmos. Environ., 12(2): 109–126 **(18 pages).**
Díaz-Robles, L.A.; Ortega, J.C.; Fu, J.S.; Reed, G.D.; Chow, J.C; Watson, J.G.; Moncada-Herrera, J.A., (2008). A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmos. Environ., 42: 8331–8340 **(10 pages).**
Djamila, H.; Ming, C.C.; Kumaresan, S., (2011). Estimation of exterior vertical daylight for the humid tropic of Kota Kinabalu city in East Malaysia. Renewable Energy. 36(1): 9–15 **(7 pages).**
Dominick, D.; Juahir, H.; Latif, M.T.; Zain, S.M.; Aris, A.Z., (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmos. Environ., 60: 172–181 **(10 pages).**
Dotse, S.Q.; Petra, M.I.; Dagar, L.; De Silva, L.C., (2018). Application of computational intelligence techniques to forecast daily PM_{10} exceedances in Brunei Darussalam. Atmos. Pollut. Res., 9: 358–368 **(11 pages).**
Elangasinghe, M.A.; Singhal, N.; Dirks, K.N.; Salmond, J.A.; Samarasinghe, S., (2014). Complex time series analysis of PM_{10} and $PM_{2.5}$ for a coastal site using artificial neural network modelling and k-means clustering. Atmos. Environ., 94: 106–116 **(11 pages).**
Fan, J.; Li, Q.; Hou, J.; Feng, X.; Karimian, H.; Lin, S., (2013). A spatiotemporal prediction framework for air pollution based on deep RNN. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.: 15– 2 **(8 pages).**
Feng, X.; Li, Q.; Zhu, Y.; Hou, J.; Jin, L.; Wang, J., (2015). Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ., 107, 118–128 **(11 pages).**
Franceschi, F.; Cobo, M.; Figueredo, M., (2018). Discovering relationships and forecasting PM_{10} and $PM_{2.5}$ concentrations in Bogotá Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. Atmos. Pollut. Res., 9: 912–922 **(11 pages).**
Graham, J.W., (2009). Missing Data Analysis: Making It Work in the Real World. Annu. Rev. Psychol., 60: 549–579 **(31 pages).**
Grivas, G.; Chaloulakou, A., (2006). Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece. Atmos. Environ., 40: 1216–1229 **(14 pages).**
Gvozdić, V.; Kovač-Andrić, E.; Brana, J., (2011). Influence of Meteorological Factors NO_2, SO_2, CO and PM_{10} on the Concentration of O_3 in the Urban Atmosphere of Eastern Croatia. Environ. Model. Assess., 16(5): 491–501 **(11 pages).**
Karri, R.R.; Mohammadyan, M.; Ghoochani, M.; Mohammadpoure, R.A.; Yusup, Y.; Rafatullah, M.; Mohammadyan, M.; Sahu, J. N., (2018). Modeling airborne indoor and outdoor particulate matter using genetic programming. Sustainable Cities Soc., 43: 395–405 **(11 pages).**
Kim, K.H.; Kabir, E.; Kabir, S., (2015). A review on the human health impact of airborne particulate matter. Environ. Int., 74: 136–143 **(8 pages).**
Lou, C.; Liu, H.; Li, Y.; Peng, Y.; Wang, J.; Dai, L., (2017). Relationships of relative humidity with $PM_{2.5}$ and PM_{10} in the Yangtze River Delta, China. Environ. Monit. Assess., 74: 136–143 **(8 pages).**
Muhammad Izzuddin, R.; Chee, F.P.; Dayou, J.; Chang, H.W. J.; Kong, S.K.S.; Sentian, J., (2019). Temporal Assessment on Variation of PM_{10} Concentration in Kota Kinabalu using Principal Component Analysis and Fourier Analysis. Curr. World. Environ., 14(3): 400–410 **(11 pages).**
Muhammad Izzuddin, R.; Chee, F.P.; Dayou, J.; Chang, H.W.J.; Kong, S.K.S.; Sentian, J., (2020). Missing Value Imputation for PM_{10} Concentration in Sabah using Nearest Neighbour Method (NNM) and Expectation Maximization (EM) Algorithm. Asian J. Atmos. Environ., 14(1): 62–72 **(11 pages).**
Munir, S.; Habeebullah; T.M.; Mohammed, A.M.F.; Morsy, E.A.; Rehan, M.; Ali, K., (2017). Analysing $PM_{2.5}$ and its association with PM_{10} and meteorology in the arid climate of Makkah, Saudi Arabia. Aerosol Air Qual. Res., 17: 453–464 **(12 pages).**
Noor, H.M.; Nasrudin, N.; Foo, J., (2014). Determinants of Customer Satisfaction of Service Quality: City Bus Service in Kota Kinabalu, Malaysia. Procedia Social Behav. Sci., 153: 595–605 **(11 pages).**
Özbay, B.; Keskin, G.A.; Doğruparmak, Ş.Ç.; Ayberk, S., (2011). Multivariate methods for ground-level ozone modeling. Atmos. Res., 102: 57–65 **(9 pages).**
Paschalidou, A.K.; Karakitsios, S.; Kleanthous, S.; Kassomenos, P.A., (2011). Forecasting hourly PM_{10} concentration in Cyprus through artificial neural networks and multiple regression models: Implications to local environmental management. Environ. Sci. Pollut. Res., 18(2): 316–327 **(10 pages).**
Polat, E.; Gunay, S., (2015). The Comparison of Partial Least Squares Regression, Principal Component Regression and Ridge Regression With Multiple Linear Regression for Predicting PM_{10} Concentration Level Based on Meteorological Parameters. J. Data Sci., 13: 663–692 **(10 pages).**
Potdar, K.; Pardawala, T.S., (2017). Forecasting Ambient Air Quality in Mumbai using Neural Networks. 5th National Conference on Role of Engineers in Nation Building: 1–4 **(4 pages).**
Saxena, S.; Mathur, A.K., (2017). Prediction of Respirable Particulate Matter (PM_{10}) Concentration using Artificial Neural Network in Kota City. Asian J. of Convergence Technol., 3(3): 1–7 **(7 pages).**
Shahraiyni, H.T.; Sodoudi, S., (2016). Statistical modeling approaches for PM_{10} prediction in urban areas; A review of 21st-century studies. Atmos., 2(15): 1–24 **(24 pages).**
Shekarrizfard, M.; Karimi-Jashni, A.; Hadad, K., (2012). Wavelet transform-based artificial neural networks (WT-ANN) in PM_{10}

pollution level estimation, based on circular variables. Environ. Sci. Pollut. Res., 19: 256–268 **(13 pages).**

Teong, K.V.; Sukarno, K.; Hian, J., Chang, W.; Chee, F.P.; Ho, C.M.; Dayou, J., (2017). The Monsoon effect on rainfall and solar radiation in Kota Kinabalu. Trans. Sci. Technol., 4(4): 460–465 **(6 pages).**

Ul-Saufie, A.Z.; Yahaya, A.S.; Ramli, N.A.; Hamid, H.A., (2011). Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM_{10} Concentration Level Based On Gaseous And Meteorological Parameters. Int. J. App. Sci. Technol., 1(4): 42–49 **(8 pages).**

Ul-Saufie, A.Z.; Yahaya, A.S.; Ramli, N.A.; Hamid, H.A., (2015). PM_{10} concentrations short term prediction using feedforward backpropagation and general regression neural network in a sub-urban area. J. Environ. Sci. Technol., 8(2): 59–73. **(15 pages).**

Ul-Saufie, A.Z.; Yahaya, A.S.; Ramli, N.A.; Rosaida, N.; Hamid, H.A., (2013). Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmos. Environ., 77: 621–630 **(10 pages).**

Vijayaraghavan, N.; Mohan, G.S., (2016). Air pollution analysis for Kannur City using artificial neural network. Int. J. Sci. Res., 5(10): 1399–1401 **(3 pages).**

Vlachogianni, A.; Kassomenos, P.; Karppinen, A.; Karakitsios, S.; Kukkonen, J., (2011). Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Sci. Total Environ., 409: 1559–1571 **(13 pages).**

Voukantsis, D.; Karatzas, K.; Kukkonen, J.; Räsänen, T.; Karppinen, A.; Kolehmainen, M., (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ., 406: 1266–1276 **(11 pages).**

Willmott, C.J.; Matsuura, K., (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res., 30: 79–82 **(4 pages).**

Wu, Z.; Fan, J.; Gao, Y.; Shang, H.; Song, H., (2019). Study on prediction model of space-time Distribution of air pollutants based on artificial neural network. Environ. Eng. Manage. J., 18(7): 1876–1890 **(15 pages).**

Xie, Y.; Bin, Z.; Lin, Z.; Rong, L., (2015). Spatiotemporal variations of PM2.5 and PM10 concentrations between 31 Chinese cities and their relationships with SO_2, NO2, CO and O3. Particuology, 20: 141–149 **(9 pages).**

Yu, H.; Wilamowski, B.M., (2016). Levenberg-marquardt training. Intell. Sys., 2: 1–16 **(16 pages).**

AUTHOR (S) BIOSKETCHES

Rumaling M.I., M.Sc. Instructor, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia.
Email: *miru9608@gmail.com*
ORCID: 0000-0003-2081-8283

Chee, F.P., Ph.D., Associate Professor, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia. Email: *fpchee06@gmail.com*
ORCID: 0000-0002-9782-5572

Chang, H.W.J., Ph.D. Candidate, Instructor, Preparatory Centre for Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia. Email: *jacksonchw@ums.edu.my*
ORCID: 0000-0002-1403-3730

Payus, C.M., Ph.D., Associate Professor, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia. Email: *melpayus@ums.edu.my*
ORCID: 0000-0003-1947-2844

Kong, S.K., Ph.D., Instructor, Department of Atmospheric Sciences, National Central University, Taoyuan, 32001, Taiwan.
Email: *kongksk@gmail.com*
ORCID: 0000-0002-7297-7393

Dayou, J., Ph.D., Associate Professor, Energy, Vibration and Sound Research Group, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia. Email: *jed@ums.edu.my*
ORCID: 0000-0002-3753-1759

Sentian, J., Ph.D., Associate Professor, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia. Email: *jsentian@ums.edu.my*
ORCID: 0000-0002-7121-2372

HOW TO CITE THIS ARTICLE

Rumaling M.I.; Chee, F.P.; Chang, H.W.J.; Payus, C.M.; Kong, S.K.; Dayou, J.; Sentian, J., (2022). Forecasting particulate matter concentration using nonlinear autoregression with exogenous input model. Global J. Environ. Sci. Manage., 8(1): 27-44.

DOI: 10.22034/gjesm.2022.01.03

url: https://www.gjesm.net/article_245403.html