International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:11, 2013

An Improved Prediction Model of Ozone Concentration Time Series Based On Chaotic Approach

N. Z. A. Hamid, M. S. M. Noorani

Abstract**—**This study is focused on the development of prediction models of the Ozone concentration time series. Prediction model is built based on chaotic approach. Firstly, the chaotic nature of the time series is detected by means of phase space plot and the Cao method. Then, the prediction model is built and the local linear approximation method is used for the forecasting purposes. Traditional prediction of autoregressive linear model is also built. Moreover, an improvement in local linear approximation method is also performed. Prediction models are applied to the hourly Ozone time series observed at the benchmark station in Malaysia. Comparison of all models through the calculation of mean absolute error, root mean squared error and correlation coefficient shows that the one with improved prediction method is the best. Thus, chaotic approach is a good approach to be used to develop a prediction model for the Ozone concentration time series.

Keywords**—**Chaotic approach, phase space, Cao method, local linear approximation method.

I. Introduction

BREATHE and inhale the Ozone (O_3) in the air can cause dangerous reactions in the respiratory system [1]. Recent studies by [2] and [3] reported that O_3 pollution increased death rate because it leads to various respiratory diseases and cardiovascular. Hence, the development of prediction models of the O_3 concentration time series is important.

The nature of the time series can be classified into deterministic or random. Deterministic time series is predictable and random time series is not predictable. Chaotic nature is in between the deterministic and random nature [4]. Chaotic time series is predictable, however, due to the sensitive dependence upon initial conditions, then, for the chaotic time series, only short-term prediction is allowed [5]. There are various approaches that have been used by previous studies to test whether O_3 time series is chaotic or not. Using the method of correlation dimension, entropy and Lyapunov exponent [6] found that O_3 concentrations are chaotic. Recently, [7] using the correlation integral method for detecting the chaotic nature of O_3 time series at different temporal scale. Phase space plot and Cao method [8] are also able to classify the nature of the time series. However, this method has never been used on O_3 time series although both have been proven effective over time series such as suspended sediment concentration, traffic flow and earthquakes [9]–[11]. Therefore, in this study, the phase space plot and Cao method are used on O_3 time series.

In previous studies, O_3 time series is often predicted using neural network and multiple linear regressions [12], [13]. Prediction process through both methods are dependent on meteorological factors such as water temperature, humidity, solar radiation and wind speed and gaseous factors such as the precursor gases of methane, carbon monoxide (CO) and nitrogen oxide (NO_x). However, if the information of those factors is not sufficient, an alternative method is needed to run the prediction. Therefore, in this study, local approximation method, a method based on chaotic approach is used. This method has its own advantages as O_3 prediction process is done simply by using data from O_3 time series only, without involving data from other factors. Local approximation method has been used by [14] to predict hourly O_3 time series and [15] to predict the daily average of O_3 time series. Both studies yielded very satisfactory prediction. Therefore, in this study, O_3 prediction is also carried out using local approximation method. There is various sub method of local approximation method. However, the latest and most commonly used is the local linear approximation method. Thus, this method will be used in this study.

The contributions of this study are to introduce the phase space plot and Cao method for detecting the presence of chaotic nature and moreover, for the first time, local linear approximation method is adapted to the time series of O_3 in Malaysia. Therefore, we chose to conduct the study on the time series of O_3 concentration observed at the benchmark station. There are three prediction models to be developed. The first is a model based on the traditional methods of autoregressive linear. The second is the chaotic approach model and third model is an improvement of the second model. Performance of the model is reflected in the calculation of mean absolute error (mae), root mean squared error ($rmse$) and correlation coefficient (cc).

II. Time Series Data

O_3 time series is observed at the benchmark stations located in Jerantut, a wide area in Pahang, a state located at East Malaysia (Fig. 1). Jerantut is one of the main settlements populated districts and the largest district in the Pahang state. This study is the first study in Malaysia using chaotic approach for analyzing the O_3 concentration time series. Therefore, the time series observed at the benchmark stations

N. Z.A. Hamid is with the Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Perak, Malaysia (e-mail: nor_zila@yahoo.com).

M. S. M. Noorani is with the Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia (e-mail: msn@ukm.my).

are used.

Fig. 1 Location of Pahang State in East Malaysia

The time series of hourly O_3 concentration is observed for six months starting July 1, 2009 until December 31, 2009. Entire time series period is 184 days (4416 hours). The time series is recorded in ppb (part per billion) units and written in the scalar form of one-dimensional vector X with

$$X = \{x_1, x_2, x_3 ..., x_N\} \quad (1)$$

N is the total number of hours and in this study $N = 4416$. O_3 concentration time series is divided into two parts. The first part is a training set while the other is a test set to see the performance of prediction models. Training set is the time series of 153 days

$$X_{train} = \{x_1, x_2, x_3 ..., x_{3672}\} \quad (2)$$

and the remaining 31 days,

$$X_{test} = \{x_{3673}, x_{3674}, x_{3675} ..., x_{4416}\} \quad (3)$$

is the time series of test set. The overall hourly O_3 time series (training and test set) observed in Jerantut station is as shown in Fig. 2 and the statistic description of the time series is as listed in Table I.

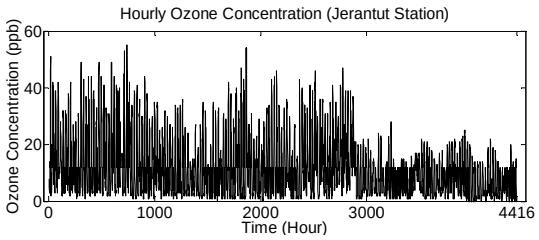

Fig. 2 Hourly O_3 concentration time series

TABLE I
STATISTICS DESCRIPTION

Statistics	Value
Mean	12
Median	9
Mode	2
Minimum	0
Maximum	55
Standard Deviation	10.7686
Variance	115.9634
Kurtosis	0.9225
Skewness	1.2279

III. CHAOTIC NATURE

A. Phase Space Plot

The training set of observed hourly O_3 concentration was recorded as data in one-dimensional vector, $X_{train} = \{x_1, x_2, x_3 ..., x_{3672}\}$ (2). With x_t is an O_3 concentration time series at t hour, a graph in two-dimensions are plotted in the plane of $\{x_t, x_{t+\tau}\}$. Hence, the parameter of delay time, τ need to be determined first. τ is the time interval value to reflect the phase space structure of the time series. τ can be determined through various method. Among them are average mutual information method and autocorrelation function. However, several studies (e.g. [9], [16]) using $\tau = 1$ and their prediction results are excellent. Since this is the first time where local linear approximation method is adapted to the time series of O_3 in Malaysia, hence, $\tau = 1$ is used.

To $\tau = 1$ that had been set, the phase space plot $\{x_t, x_{t+1}\}$ is built. The existence of a well defined attractor shows that the nature of the time series is chaotic ([9], [11]). Fig. 3 is the phase space plot of a time series of (2) with $\tau = 1$. It can be seen that there exists an attractor where most of the points converge towards it. Thus, the observed hourly O_3 time series with $\tau = 1$ is chaotic in nature. However, there is a point away from the attractor. These points are known as outliers that may result from the noise disturbance.

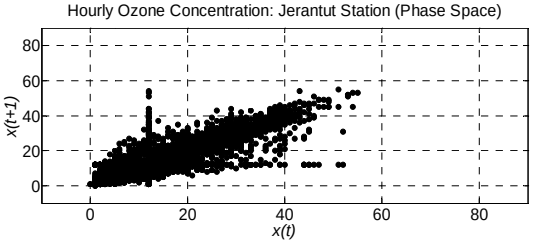

Fig. 3 Phase space plot

B. Cao Method

Time series of X_{train} is reconstructed into an m-dimensional phase space

$$\mathbf{Y}_j^m = \left(x_j, x_{j+\tau}, x_{j+2\tau}, \quad ..., x_{j+(m-1)\tau}\right) \tag{4}$$

Delay time parameter, τ was set as 1 and minimum embedding dimension m is determined using the Cao method. m is the minimum number of variables required to describe the nature of the time series [17]. This means that there are at least m variables that influence the studied time series. m is calculated using Cao method [8]. This method is chosen because apart from calculating m, the method is also able to distinguish between chaotic and random nature of the time series [8].

Cao method (for more calculation method, see [8]) involves the calculation of two parameters, namely $E1(d)$ and $E2(d)$ where d is the variation of embedding dimension. If $E1(d)$ stops changing when the value of d is greater than the value of d_0, then $d_0 + 1$ is the minimum embedding dimension, m. For a random time series data, the value of $E1(d)$ will not reach saturation with increasing d. Therefore, the graph of d against $E1(d)$ can be used to distinguish whether the nature of the time series is chaotic or random.

For the purpose of strengthening, [8] also introduced the calculation of $E2(d)$. For random time series, $E2(d)$ will be equal to 1 for any d. However, for chaotic time series, there will always be some $E2(d)$ where $E2(d) \neq 1$. Therefore, if there exist $E2(d) \neq 1$, then, the observed time series is chaotic.

The results are as shown in Fig. 4. It is observed that after the value of $d_0 = 5$, $E1(d)$ started saturate within the value between 0.9 and 1.0. Thus, the value of m is 6. At $d = 1$ and $d = 2$, $E2(d) \neq 1$. Due to the existence of $E2(d) \neq 1$, then, according to [8], the studied time series is chaotic in nature. This further strengthens results of $E1(d)$ and the phase space plot. Thus, the prediction model based on chaotic approach is expected to perform well since the nature of the time series is chaotic.

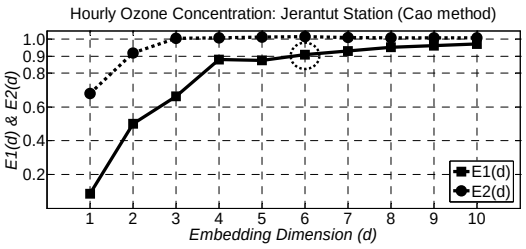

Fig. 4 *E1(d)* and *E2(d)* from Cao method

IV. Prediction Models

In this paper, three prediction models are developed. The first is the traditional prediction model using autoregressive linear method. The second model is based on the chaotic approach and the third model is an improvement of the second model. Performance of the model is reflected in the calculation of mean absolute error (mae), root mean squared error ($rmse$) and correlation coefficient (cc).

A. Autoregressive Linear Model

Through this model, a linear equation $x_{P+1} = A * x_P + B$ is fitted to the training set of (2). The prediction of x_{P+1} is obtained by inserting the value of x_P. The value of coefficients A and B is calculated through the least square method. The linear equation is

$$x_{P+1} = 0.8601 * x_P + 1.7585 \tag{5}$$

To predict x_{3673}, the x_{3672} is used. To predict x_{3674}, x_{3673} is used and so on. The prediction result of this first model and comparison with X_{test} is as shown in Fig. 5.

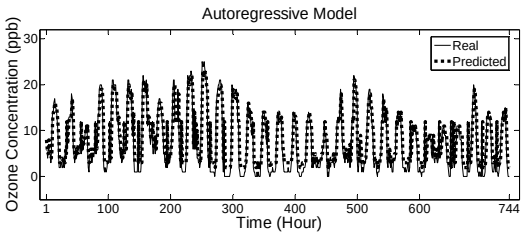

Fig. 5 Prediction result from autoregressive linear model

B. Chaotic Approach Model

For the prediction model based on chaotic approach, the local linear approximation method is used. Phase space of (4) is built with $\tau = 1$ and $m = 6$. Thus, the reconstructed phase spaces are

$$\mathbf{Y}_j^6 = \left(x_j, x_{j+1}, x_{j+2}, x_{j+3,} x_{j+4}, x_{j+5}\right) \tag{6}$$

with $j = 1,2,3,...,N-5$. Since $N = 3672$, the final phase space is

$$\mathbf{Y}_{3667}^6 = \left(x_{3667}, x_{3668}, x_{3669}, x_{3670,} x_{3671}, x_{3672}\right) \tag{7}$$

Nearest neighbor to the final phase space is sought by calculating the minimum Euclidean distance $\left\|\mathbf{Y}_{3672}^6 - \mathbf{Y}_w^6\right\|$ where $w < 3672$. k nearest neighbors are searched and labeled as

$$\mathbf{Y}_{P_i} = \left(\mathbf{Y}_{P_1}, \mathbf{Y}_{P_2}, \mathbf{Y}_{P_3}, ..., \mathbf{Y}_{P_k}\right) \quad (8)$$

A one step forward of $\mathbf{Y}_P$ was labeled as

$$\mathbf{Y}_{P_i+1} = \left(\mathbf{Y}_{P_1+1}, \mathbf{Y}_{P_2+1}, \mathbf{Y}_{P_3+1}, ..., \mathbf{Y}_{P_k+1}\right) \quad (9)$$

m-column values of each $\mathbf{Y}_{P_i}$ is searched. The corresponding m-column values of (8) is labeled as

$$x_{P_i} = \left(x_{P_1}, x_{P_2}, x_{P_3}, ..., x_{P_k}\right) \quad (10)$$

while the corresponding m-column values of (9) is wrote as

$$x_{P_i+1} = \left(x_{P_1+1}, x_{P_2+1}, x_{P_3+1}, ..., x_{P_k+1}\right) \quad (11)$$

A linear equation of $x_{P_i+1} = A * x_{P_i} + B$ is fitted to both (10) and (11) and the equation is

$$x_{P+1} = 1.1019 * x_P + 0.4662 \quad (12)$$

For forecasting purposes, (12) is used. To predict $x_{P+1} = x_{3673}$, the $x_P = x_{3672}$ is used. To predict x_{3674}, x_{3673} is used and so on. According [18], the number k is small. In this study, k is chosen by trial and error process. k in this study is chosen as $k = 200$. The prediction result of this second model and comparison with X_{test} is as shown in Fig. 6.

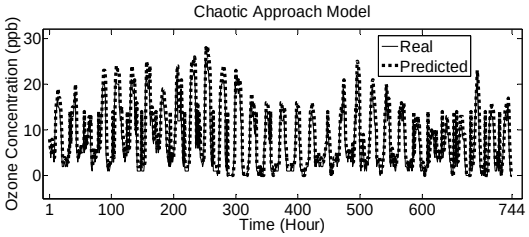

Fig. 6 Prediction result from chaotic approach model

C. Improvement of Chaotic Approach Model

In the chaotic model approach only one linear equation is derived based on the $k-$nearest neighbors to the final phase space $\mathbf{Y}_{3667}^6$. In this improved model, the final phase space is different because for each forecasting, the length of the time series is different. For example, to predict the $x_{P+1} = x_{3673}$, the time series up to x_{3672} is used to find the final phase space, $k-$nearest neighbors and the linear equation. To predict x_{3674}, the time series up to x_{3673} is used to find the final phase space, $k-$nearest neighbors and the linear equation. Each neighborhood has its own linear equations. To facilitate understanding of the development process of the improved model, Fig. 7 below can be referred. The prediction result of the improvement model and comparison with X_{test} is as shown in Fig. 8.

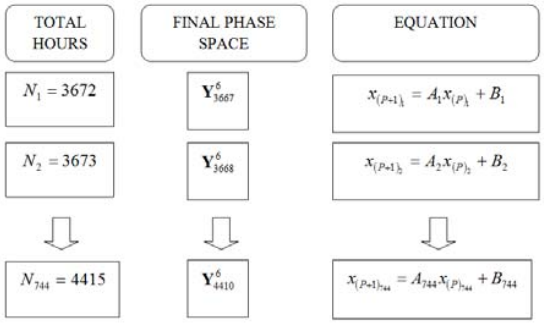

Fig. 7 The development process of the improved model

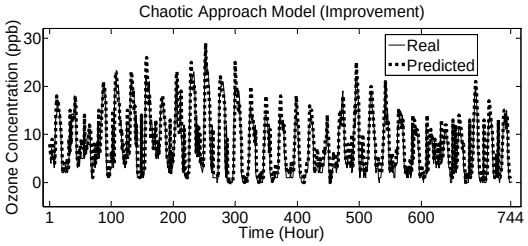

Fig. 8 Prediction result from the improved model

V. Discussion

TABLE II
COMPARISON OF THE PERFORMANCE INDICATOR

Performance Indicator	Model 1	Model 2	Model 3
Method	Autoregressive Linear	Chaotic Approach	Improved Chaotic Approach
mae	2.4059	2.6949	2.2500
rmse	3.3975	3.9101	3.2955
cc	0.8517	0.8512	0.8699

From Table II, it can be seen that all three models give good prediction results in which the cc is above 0.8 and approaches 1. According to [19], cc values exceeding 0.8 indicate that there is a strong relationship between the real and predicted time series values. However, the autoregressive linear (Model 1) method cannot predict isolated and low values (Fig. 5). In addition, this method cannot explain the nature of the time series. Thus, chaotic approach is introduced to meet the shortage of Model 1.

Through chaotic approach, the presence of chaotic nature of the time series is detected by means of phase space plot and Cao method. By setting $\tau = 1$ and use $m = 6$ obtained from the Cao method, the phase space is built. Forecasting the phase space is done through local linear approximation method (Model 2). Next, Model 3 is developed based on the

improvement of the local linear approximation method. Models 2 and 3 are seen to have an advantage, in which apart from the prediction, the difference between the natures of the time series can also be analyzed. In addition, the prediction of isolated and low values also achieved through Model 2 and Model 3 (Figs. 6 and 8).

Furthermore, the number of variables that affect the time series can also be identified. In a series of previous studies, the relationship between O_3 time series and variables that influence O_3 has been observed. Among the variables are the variable based on the meteorological factors such as water temperature, suspended dust, relative humidity, solar radiation, solar energy, wind direction and wind speed as well as variables of gaseous pollutants such as precursor gases carbon dioxide (CO_2), methane, CO, and NO_x ([12], [20]). $m = 6$ of the $E1(d)$ plot from Cao method suggest that at least six variables influence the O_3 time series. List of variables in the past studies suggests that there are more than six variables which influence the O_3. Thus, $m = 6$ obtained is compatible with the findings of previous studies.

By comparing the performance indicator of all three models, Model 3 with the improvement of the local linear approximation method is seen more powerful than the other two models. When compared to Model 1 with autoregressive linear method, the mae is reduced 6.5%, the $rmse$ decreased 3% and the cc is increased 2%. Through comparison with Model 2, the value of mae is reduced 16.5%, $rmse$ values decreased 15.7% and the value of cc is increased 2.2%. This makes the Model 3 is the best model for prediction of O_3 time series observed in this study.

VI. Conclusion and Future Research

In this study, the chaotic nature of hourly O_3 time series is detected through phase space plot and the Cao method. Three prediction models were built. The first is based on the autoregressive linear method. The second is based on the chaotic approach and the third model is an improvement of the second model. Comparison of forecasting performance through mae, $rmse$ and cc found that the third model with improved chaotic approach method is the best model. In this study, the value of τ is set as $\tau = 1$. In the future, the method such as average mutual information and autocorrelation function is proposed to calculate the τ value. In addition, the number of the nearest neighbors k is chosen at random through the trial and error process. In the future, the method on how to choose an optimal value of k should be explored.

Acknowledgment

Thanks to Dr. Liew Juneng and Assoc. Prof. Dr. Mohd Talib Latif from the Faculty of Science and Technology, Universiti Kebangsaan Malaysia for ideas contributed and knowledge shared, to the Department of Environment, Malaysia for providing the necessary data and Universiti Pendidikan Sultan Idris as well as Ministry of Higher Education, Malaysia for sponsoring this study.

References

[1] I. S. Mudway and F. J. Kelly, "Ozone and the lung: a sensitive issue," *Molecular Aspects of Medicine*, vol. 21, pp. 1-48, 2000.
[2] S. I. V. Sousa, M. C. M. Alvim-Ferraz, and F. G. Martin, "Health effects of ozone focusing on childhood asthma: What is now known - a review from an epidemiological point of view," *Chemosphere*, vol. 90, pp. 2051–2058, 2013.
[3] W. R. W. Mahiyuddin, M. Sahani, R. Aripin, M. T. Latif, T. Thach, and C. Wong, "Short-term effects of daily air pollution on mortality," *Atmospheric Environment*, vol. 65, pp. 69-79, 2013.
[4] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data.* Springer-Verlag, Inc., New York, 1996, pp. 15-16.
[5] J. C. Sprott, *Chaos and Time-Series Analysis.* Oxford University Press, 2003, pp. 104-105.
[6] V. Cuculeanu, C. Rada, and A. Lupu, "Study on the geometrical and dynamical characteristics of the Arosa ozone series attractor," *Geophys*, vol. 52-53, pp. 77-85, 2008.
[7] A. B. Chelani, "Nonlinear dynamical analysis of ground level ozone concentrations at different temporal scales," *Atmospheric Environment*, vol. 44, pp. 4318-4324, 2010.
[8] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D*, vol. 110, pp. 43-50, 1997.
[9] B. Sivakumar, "A phase-space reconstruction approach to prediction of suspended sediment concentration in rivers," *Journal of Hydrology*, vol. 258, pp. 149-162, 2002.
[10] C. Frazier, and K.M. Kockelman, "Chaos Theory and Transportation Systems: An Instructive Example," *Transportation Research*, vol. 1897, pp. 9-17, 2004.
[11] S. S. Lakshmi, and R. K. Tiwari, "Model dissection from earthquake time series: A comparative analysis using modern non-linear forecasting and artificial neural network approaches," *Computers & Geosciences* vol. 35, pp. 191–204, 2009.
[12] S. I. V. Sousa, F. G. Martins, M. C. Pereira, and M. C. M. Alvim-Ferra, "Prediction of ozone concentrations in Oporto city with statistical approaches." *Chemosphere* vol. 64, pp. 1141–1149, 2006.
[13] M. Banja, D. K. Papanastasiou, A. Poupkou, and D. Melas, "Development of a short–term ozone prediction tool in Tirana area based on meteorological variables," *Atmospheric Pollution Research* vol. 3, pp. 32-38, 2012.
[14] J. Chen, S. Islam, and P. Biswas, "Nonlinear dynamics of hourly O3 concentration: nonparametric short term prediction," *Atmospheric Environment*, vol. 32 no. 11, pp. 1839-1848, 1998.
[15] K. Kocak, L. Saylan, and O. Sen, "Nonlinear time series prediction of O_3 concentration in Istanbul," *Atmospheric Environment*, vol. 34, pp. 1267-1271, 2000.
[16] A. W. Jayawardena, "Runoff prediction using a local approximation method," *IAHS*, vol. 239, pp. 167-171, 1997.
[17] S. K. Regonda, B. Rajagopalan, U. Lali, M. Clark, and Y. I. Moon, "Local polynomial method for ensemble predict of time series," *Nonlinear Processes in Geophysics*, vol. 12, pp. 397-406, 2005.
[18] M. Casdagli, "Chaos and deterministic versus stochastic nonlinear modeling," *J. Royal Stat. Soc. B.* vol. 54, no. 2, pp. 303-328, 1991.
[19] K. Hardy, *Linear Algebra for Engineers and Scientists using Matlab.* Pearson Education, Inc., Upper Saddle River, New Jersey, 2005, pp. 238-240.
[20] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, and E. Alper, "Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach," *Advances in Environmental Research*, vol. 5, pp. 47-59, 2001.

N. Z. A. Hamid was born in Kepala Batas, Penang, Malaysia on November 14, 1983. She earned a Bachelor of Education (Honours) in Mathematics from Universiti Putra Malaysia, Selangor, Malaysia in 2006. She continued her studies at the Masters level and earned a Master Degree in Applied Mathematics from Universiti Kebangsaan Malaysia, Selangor, Malaysia in 2008. Her expertise areas are dynamical systems and chaos theory.

Currently, she is a doctor of philosophy student at the Universiti Kebangsaan Malaysia. She also served as a lecturer at the Universiti Pendidikan Sultan Idris, Perak, Malaysia.

Several papers as a result of Mrs. Hamid's doctoral research studies have been successfully produced (eg, N. Z. A. Hamid, and M. S. M. Noorani, "On prediction of Subang, Selangor daily rainfall data: An application of local approximation method," *Jurnal Sains dan Matematik*, vol. 4, no. 2, pp. 49-57, 2012, N. Z. A. Hamid, and M. S. M. Noorani, "Modeling of prediction system: An application of the nearest neighbor approach to chaotic data," *App. Math. and Comp. Intel.*, vol. 2, no. 1, pp. 137–148, 2013, N. Z. A. Hamid, M. S. M. Noorani, L. Juneng, and M. T. Latif, "Prediction of ozone concentrations using nonlinear prediction method," *AIP Conf. Proc*. vol. 1522, pp. 125-131, 2013).

M. S. M. Noorani was born in Pasir Mas, Kelantan, Malaysia on October 10, 1961. He obtained his Bachelor, Master and Doctor of Philosophy from the University of Warwick, United Kingdom. His research field covers the area of dynamical systems, ergodic theory, topology and functional analysis.

In 2012, Elsevier awarded him as one of the 20 most cited authors in the journal of Communication in Nonlinear Science and Numerical Simulation for the period of 2007-2011. Currently, he is a Professor at the Universiti Kebangsaan Malaysia.

In the mathematical community, Prof. Noorani was the President of the Malaysian Mathematical Sciences Society. He was also active as a Council Member of the Southeast Asian Mathematical Society, American Mathematical Society, Society for Industrial and Applied Mathematics USA and Mathematical Association of America.