Influence of Nitrogen-di-Oxide, Temperature and Relative Humidity on Surface Ozone Modeling Process Using Multigene Symbolic Regression Genetic Programming

Alaa F. Sheta
Software Engineering Department
Zarqa University
Zarqa, Jordan

Hossam Faris
Business Information Technology Department
The University of Jordan
Amman, Jordan

***Abstract*—Automatic monitoring, data collection, analysis and prediction of environmental changes is essential for all living things. Understanding future climate changes does not only helps in measuring the influence on people life, habits, agricultural and health but also helps in avoiding disasters. Giving the high emission of chemicals on air, scientist discovered the growing depletion in ozone layer. This causes a serious environmental problem. Modeling and observing changes in the Ozone layer have been studied in the past. Understanding the dynamics of the pollutants features that influence Ozone is explored in this article. A short term prediction model for surface Ozone is offered using Multigene Symbolic Regression Genetic Programming (GP). The proposed model customs Nitrogen-di-Oxide, Temperature and Relative Humidity as the main features to predict the Ozone level. Moreover, a comparison between GP and Artificial Neural Network (ANN) in modeling Ozone is presented. The developed results show that GP outperform the ANN.**

Keywords: Air pollution; Surface Ozone; Multigene Symbolic Regression; Genetic Programming; Multilayer perceptron neural network; Prediction.

I. Introduction

Tropospheric ozone is an air pollution which causes serious human health problems. The insufficient adherence to the international standard air quality trends, growth of industrialized activities and the emitting of various types of gasses such as carbon monoxide (CO), nitrogen oxides (NO_x), Sulphur dioxide (SO_2), and Particle Pollution (PM_{10}) and ($PM_{2.5}$) in the air without any concern about the impact on human health became a common problem worldwide. These behaviors cause a rise to the earth temperature and affect many meteorological variables [1], [2].

The role of stratospheric ozone in the air is to filter out the greatest portion of the sun possibly harmful shortwave the ultraviolet (UV) radiation. This means that the depletion of ozone allows more UV emissions to touch the earths surface. Many studies proved that these UV emissions could have severe impacts on human beings, animals and plants [3]. In [4] authors explored the dramatic effects of UV radiation on the eye and the skin. Higher temperatures associated climate change possibly will lead, among numerous other effects, to increasing rate of skin cancer. The influence of ambient ozone on human health was studied for fifty US cities for five summers was presented in [5]. Countries such as New Zealand developed many studies on air quality to estimate the likely health problem which may be encountered and decide where emissions should be condensed to improve air quality. In [6], a published report studied the influence of CO, nitrogen dioxide (NO_2), SO_2, O_3, and benzene and benzo(a)pyrene (BaP) in air.

In the past, researchers proposed different types of models to forecast the concentrations of pollutants. Some of these models are statistical based like Autoregressive-moving-average (ARMA) models and linear regression models [7]–[10]. Recently, a more attention was given to machine learning techniques based models such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) [2], [11]–[15] for developing forecasting models.

In this work, Multigene Symbolic Regression GP is used to develop short term prediction model of surface Ozone. The proposed model can predict the mean surface Ozone based on limited number of attributes. They are the Nitrogen-di-oxide, temperature and relative humidity. The Multigene GP has some advantages over other techniques like ANNs such as; producing compact mathematical models that have explanation power and easy to evaluate. A complete comparison between both techniques on solving the modeling problem is presented.

This paper is organized as follows. An overview of the ANN technique is presented in Section II. GP as an evolutionary computation technique is presented in Section III. The evaluation criterion adopted to check the performance of the developed models are presented in Section IV. The area of study considered with detailed information about data collection is discussed in V. Section VI provides the experimental setup and results of the two developed models of the Ozone based ANN and Multigene Symbolic Regression GP.

II. Multilayer Perceptron ANN

ANN was first defined as an information-processing system. This system has large number of simple processing units called ”neurons”. These neurons interconnect by sending and receiving signals which activate the neurons connected to it. A huge number of these neurons constitute a neural network. ANN is distinguished by certain performance characteristics such as its architecture, its training algorithm and the activation function. In this work, we investigate the application of multilayer feedforward neural network which is one of the most common types of neural networks applied for function approximation and prediction [13], [16], [17]. In MLP-ANN, neurons are arranged in layers (input, hidden and output layers). The information in feedfoward MLP-ANN flows in only one forward direction, from the input layer, through the hidden layers to the output layer [18]. Figure 1 depicts an example of a feedfoward MLP designed for Ozone prediction. In this example, the MLP has four neurons in a single hidden layer.

A. Learning algorithm

To adjust ANN weights such that the learning process achieved its goal by modeling the relationship between the inputs and output we need a learning algorithm. One of the very famous learning algorithms is the backpropagation (BP) learning algorithms. BP works by adjusting a cost function to minimize the error difference between the actual output and the ANN output. This function could be simply the sum of the error square. The learning process can be split into number of phases as below:

1) **Hidden layer**:
Assume we have a set of input-output measurements in the form of x_i, y_i. The inputs x_i are always presented to the input layer, then pass to the hidden layer weighted by the weights w_{ij}. The hidden layer always have a nonlinear function known as sigmoid function (see Equation 1). The output of each neuron in the hidden layer is the summation function presented in Equation 2.

$$\begin{aligned} y_j &= \phi(S_j) \\ \psi(x) &= \frac{1}{1+e^{-x}} \end{aligned} \tag{1}$$

$$S_i = w_0 + \sum_{i=1}^{n} w_{ij} x_i \tag{2}$$

where $i = 1, 2 \ldots, n$ and $j = 1, 2, \ldots, m$. ψ and y_j are the activation function and output of the j^{th} node in the hidden layer, respectively.

2) **Output layer**:
After the computation of each output from the neurons in the hidden layer, the information is processed to the output layer. The output layer also has number of neurons which most likely less than the number of neurons in the hidden layer. Neurons in this layer most likely to have linear sigmoid function. The computed output for neurons in the output layer is presented in Equation 3.

$$Y = \varphi(\sum_{j=1}^{k} W_j y_j) \tag{3}$$

k is the number of neurons in the output layers. φ is the linear activation function. Y is the neural network output from the single neuron in the output layer as in our case study.

The learning process continues till we minimize a cost function. In our case, the cost function minimizes the difference between the actual and the result of the network as described in Equation 4. It is defined as the Root Mean Square (RMSE). RMSE can be described by Equation 4.

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{4}$$

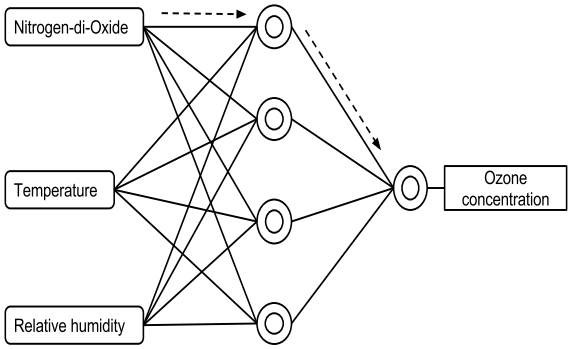

Fig. 1. Feedforward neural network for Ozone prediction

III. Genetic Programming

Genetic Programming is an evolutionary process which was successfully used to solve diversity of problem in system identification and control [19], [20]. GP was inspired from idea of nature selection and evolution introduced by Darwain. GP uses the concept of survival of the fitness to develop solutions that more likely fits to a problem. It is a population based approach. In GP, the population comes in a form of tree structure not a chromosome such as in the case of Genetic Algorithms (GAS) [21]–[24]. A block diagram which shows the GP evolutionary process is presented in Figure 2.

A. Population Initialization and Tree Representation

The initial population for any evolutionary process is produced most likely randomly. In GP a random population P_0 of tress is generated. Each tree represents a solution of a given problem. GP evolves tree structures which is composed of a set of functions and terminals sets provided by the user.

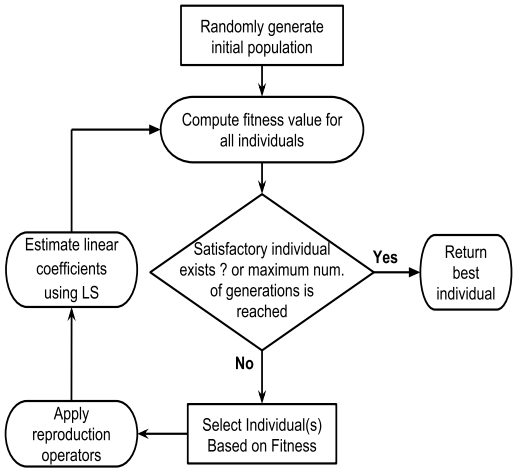

Fig. 2. Flow chart of the GP technique

The fitness of the initial population is computed according to a given fitness function.

B. Function and Terminal Sets

To develop a mathematical model which represents a relationship between input and output variables, we have to define both function and terminal sets. For a set of inputs $x_1, x_2, x3$ and x_4 to produce an output y, we may have a tree structure which produce the Equation 5. The function ϑ and terminal χ sets are given in Equation 6.

$$\begin{aligned} y &= \zeta(x_1, x_2, x_3, x_4) \\ &= a \times x_1 \times x_2 + b \times \frac{x_3}{x_4} \end{aligned} \tag{5}$$

$$\begin{aligned} \vartheta &= \{\times, +, \div\} \\ \chi &= \{x_1, x_2, x_3, x_4, \theta\} \end{aligned} \tag{6}$$

θ is defined as a random floating point number such that $\theta \in [-1, 1]$. Thus, a and b are also defined in the domain $a, b \in [-1, 1]$

C. Selection Mechanism

While population evolves, selecting individuals for both crossover and mutation depends on what is called the selection mechanism. This is essential process in the generation of new population. Many selection mechanism were presented [25]. They include roulette wheel technique, stochastic universal sampling, tournament selection and many others [26]. Number of selection mechanism used in GP were presented in [27].

D. Multigene Symbolic Regression GP

Symbolic regression method was presented by J. Koza [19]. The objective of this method is to search the space of possible mathematical expressions (i.e. equations) while minimizing some error criteria. Developing mathematical function between input variables x_i and an output y is a challenge. It is important to find the function ζ which relates the inputs and output. Symbolic regression explores both the space of models along with the space of all possible parameters simultaneously such that it can find the best model which minimize the error criterion.

E. Crossover

Crossover is the main operator in any evolutionary process. Crossover is performed between two individuals (i.e. Tree) [28]. A study of crossover operators in GP was presented in [29]. Assuming we have two parents of genes $T_1, \ldots, T_5$ and $R_1, \ldots, R_3$. In Table I, we show the crossover operation in multigene GP.

TABLE I
CROSSOVER IN MULTIGENE GP

T_1	T_2	T_3	T_4	T_5	R_1	R_2	R_3
T_1	R_2	R_3	T_4	T_5	R_1	T_2	T_3

In Multigene symbolic regression, the model output $\hat{y}$ is formed by a weighted output of each of the trees/genes in the multigene individual plus a bias term. Each tree is a function of zero or more of the n inputs variables $x_1, \ldots, x_n$. Mathematically, a Multigene regression model can be written as:

$$\hat{y} = \delta_0 + \delta_1 \times Tree_1 + \cdots + \delta_m \times Tree_m \tag{7}$$

δ_0 represents the bias or offset term while $\delta_1, \ldots, \delta_m$ are the gene weights and m is the number of genes (i.e. trees) which constitute the available individual. The values of δ coefficients can be estimated using least square estimation technique. A simple example of two multigene model is presented in Figure 3 and Equation 8.

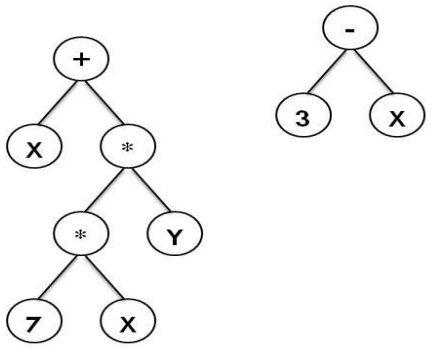

Fig. 3. Example of a Multigene Symbolic GP model

$$\delta_0 + \delta_1[X + (7 \times X) \times Y] + \delta_2[3 - X] \tag{8}$$

F. Mutation

Mutation is a relatively important operator it helps in keeping diversity in the population especially when most individual has the same fitness. Mutation helps keeping the exploration in the population. Mutation in multigene GP operates almost the same way as in standard GP.

IV. Performance Criterion

Number of performance criterion were used to evaluate the performance of the developed ANN and Multigene GP models. These evaluation criterion are presented in the following equations.

1) Euclidian distance (ED):

$$ED = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

2) Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (10)$$

3) Mean Magnitude of Relative Error (MMRE):

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (11)$$

where y and $\hat{y}$ are the actual measured Ozone level and the predicted Ozone level developed by the ANN and GP models given n measurements.

V. Site characterization and data

The study area under study is Chenbagaramanputhur. It is a rural place in Kanyakumari district and is about 12 km from Nagercoil town. In the North and North East of the city, you can find the Tirunelveli district. Kerala State is located in the North West and sea in the west and south of Chenbagaramanputhur (See Figure 4).

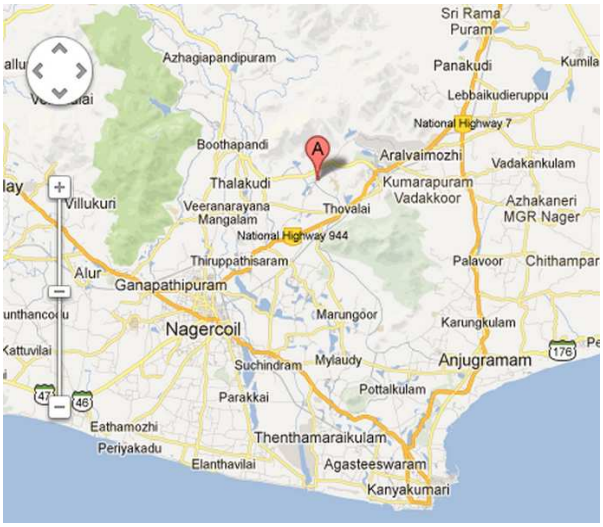

Fig. 4. Location of the area of study at Chenbagaramanputhur

The data used in this study were reported in [13]. Authors in [13] mentioned that the measurements were collected using a portable Aeroqual series S200. The Aeroqual series 200 can measure various ozone levels. Measurements were taken every 3 hours intervals for a period of 3 months during May 2009 to July 2009. Figure 5 shows the inputs and output of the proposed models. The variables used as inputs and output are presented in Table II.

TABLE II
Inputs and Output model variables

Inputs	Nitrogen dioxide concentration	x_1
	Mean temperature	x_2
	Prevailing % Relative Humidity	x_3
Output	Mean surface ozone concentration O_3	y

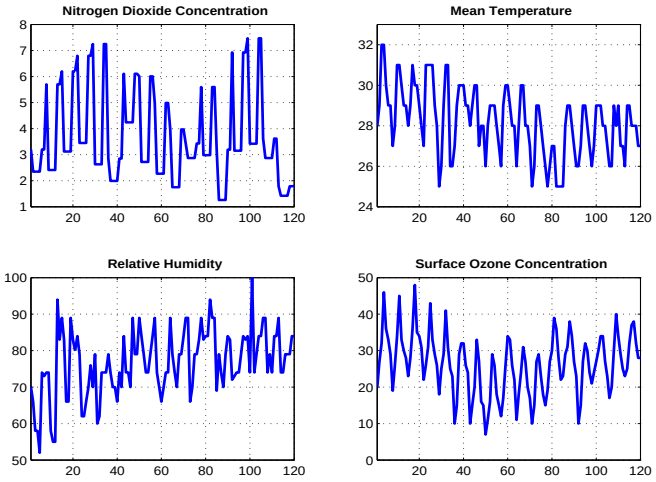

Fig. 5. Training, Testing and Validation data set [13]

VI. Experimental Setup and Results

A. Developed MLP-ANN Model

We developed a MLP-ANN model using the input-output data presented in Table II to model the surface Ozone using the parameters given in Table III. Various number of neurons in the hidden layer were explored during the learning process. The best number of neurons found was four. Figure 6, shows that the MLP training process had fast convergence to the minimum training error after only nine cycles (epchs). Figure 6 shows the actual and estimated Ozone surface values based the final developed MLP model.

B. Developed Multigene GP Model

To develop the genetic programming model, GPTIPS MATLAB Toolbox developed in [28] is used. GPTIPS is a powerful genetic programming software tool which can be used of modeling of dynamical nonlinear systems. The tool can be configured to evolve multigene tree structure. The Multigene approach often develops simpler models than evolving models consisting of one monolithic GP tree.

The data set described earlier was loaded then the Multigene GP was applied using GPTIPS Tool. The parameters of

TABLE III
Neural Network parameters

Parameter	Value
Architecture	Multi Layer perceptron
Number of hidden layers	1
Nodes in first hidden layer	4
Epochs	50
Training method	Scaled conjugate gradient

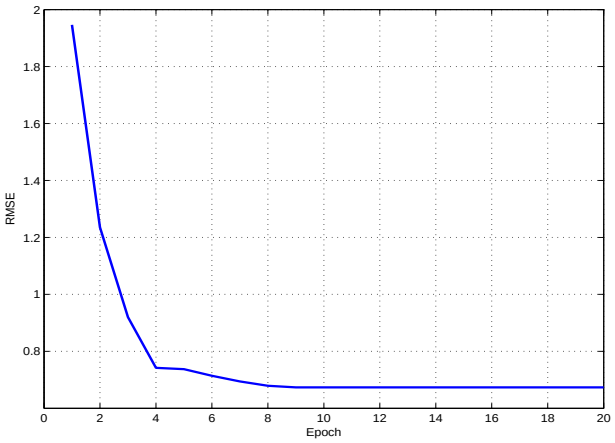

Fig. 6. Convergence of the MLP-ANN

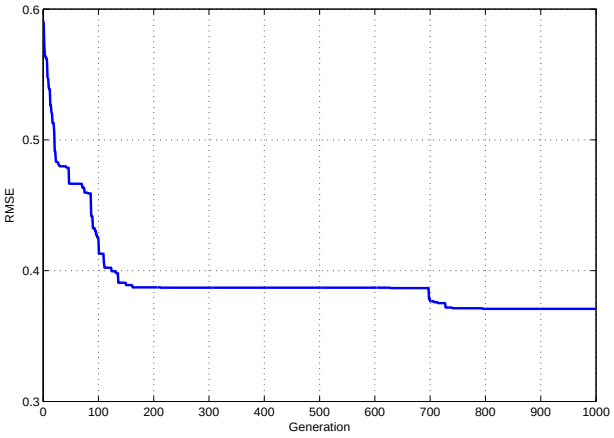

Fig. 8. Convergence of the GP evolutionary process

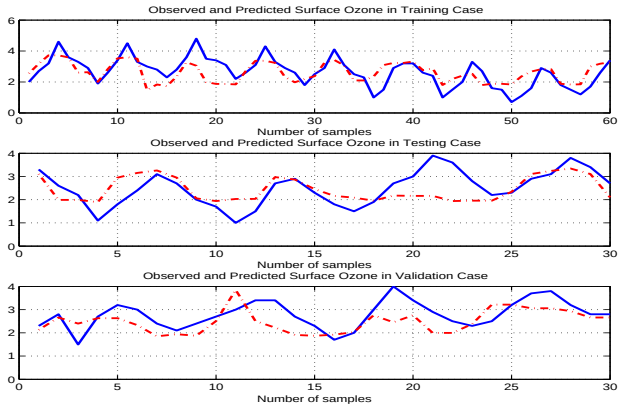

Fig. 7. Observed and Predicted O_3 using MLP-ANN model

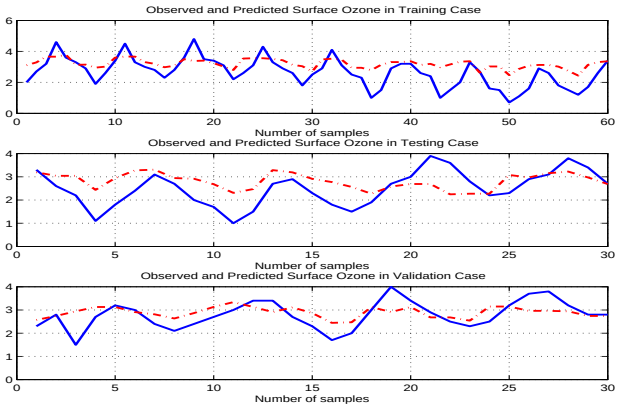

Fig. 9. Observed and Predicted O_3 using Multigene GP Model

the algorithm were tuned as listed in Table IV. In Figure 8, the convergence of GP over 100 generations is shown. The best generated Surface Ozone Multigene GP model is given in Equation 12. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. Figure 9 shows the actual and estimated surface Ozone values based the developed GP model.

TABLE IV
GP TUNING PARAMETERS

Population size	50
Number of generation	1000
Selection mechanism	Tournament
Max. tree depth	12
Probability of Crossover	0.85
Probability of Mutation	0.1
Max. No. of genes allowed in an individual	7

$$\begin{aligned} y &= 0.01442\ x_1^2 - 0.9507\ x_2 - 0.2634\ x_3 - 0.7902\ x_1 \\ &+ 0.006796\ x_2^2 + 0.000828\ \ x_3^2 + 0.03329\ x_1 x_2 \\ &- 0.0001513\ x_2^2\ (2x_1 - x_3) + 30.63 \end{aligned} \quad (12)$$

C. Comments on Results

In order to compare the performance of GP and MLP for predicting Ozone concentrations, the evaluations criteria discussed in Section IV are used to assess both developed model. The criteria measurements for the models are computed and summarized in Table V. It can be noticed that the Multigene GP model has shown better prediction results over the MLP model for training, testing and validation partitions by means of all evaluation criteria. Moreover, the final developed GP model shown in Equation 12 is considered much simpler than the complex model of the ANN approach.

VII. CONCLUSIONS AND FUTURE WORK

A comparison between genetic programming and multilayer perceptron neural networks were presented for short term prediction of surface Ozone based on limited number of measured pollutant and meteorological variables. The GP approach adopted is based on Multigene symbolic regression which generates mathematical models of linear combinations of low order non-linear transformations of the input variables. Based on this comparison, it can be concluded that the evolutionary models of the Multigene GP have promising potential for predicting surface ozone concentrations when

TABLE V
EVALUATION CRITERIA FOR THE DEVELOPED MODELS

	Multigene GP			MLP-ANN		
	Training	Testing	Validation	Training	Testing	Validation
RMSE	0.90342	0.75708	0.51826	0.74203	0.68969	0.58236
ED	6.9979	4.1467	2.8386	5.7477	3.7776	3.1897
MAE	0.71996	0.62665	0.40887	0.59362	0.53972	0.48025
MMRE	0.414	0.32339	0.16539	0.29577	0.24084	0.19439

the available number of measured pollutant and meteorological variables is limited as the case investigated in this study. The Ozone Multigene GP model was also a compact model. Future investigation on Multigene GP and other soft computing techniques on handling the environmental monitoring problems will be considered.

ACKNOWLEDGEMENT

This research is funded by the Deanship of Research at Zarqa University, Jordan.

REFERENCES

[1] 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, no. 1, pp. 47 – 59, 2001.

[2] S. Abdul-Wahab and S. Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," *Environmental Modelling & Software*, vol. 17, no. 3, pp. 219 – 228, 2002.

[3] M. Norval, R. M. Lucas, A. P. Cullen, F. R. de Gruijl, J. Longstreth, Y. Takizawa, and J. C. van der Leun, "The human health effects of ozone depletion and interactions with climate change," *Photochem. Photobiol. Sci.*, vol. 10, pp. 199–225, 2011.

[4] H. Trker and M. Yel, "Effects of ultraviolet radiation on mole rats kidney: A histopathologic and ultrastructural study," *Journal of Radiation Research and Applied Sciences*, 2014.

[5] M. Bell, R. Goldberg, C. Hogrefe, P. Kinney, K. Knowlton, B. Lynn, J. Rosenthal, C. Rosenzweig, and J. Patz, "Climate change, ambient ozone, and health in 50 us cities," *Climatic Change*, vol. 82, pp. 61–76, 2007.

[6] B. Carbon, "Monitoring of CO, NO2, SO2, ozone, benzene and benzo(a)pyrene in new zealand, air quality technical report no. 42," Tech. Rep., 2004.

[7] O. Pastor-Bárcenas, E. Soria-Olivas, J. D. Mart'ın-Guerrero, G. Camps-Valls, J. L. Carrasco-Rodr'ıguez, and S. del Valle-Tascón, "Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modelling," *Ecological Modelling*, vol. 182, no. 2, pp. 149–158, 2005.

[8] E. Agirre, A. Anta, and L. J. R. Barron, "Forecasting ozone levels using artificial neural networks," *Forecasting Models*, pp. 208–218, 2010.

[9] V. R. Prybutok, J. Yi, and D. Mitchell, "Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentrations," *European Journal of Operational Research*, vol. 122, no. 1, pp. 31 – 40, 2000.

[10] V. Gvozdic, E. Kovac-Andric, and J. Brana, "Influence of meteorological factors NO2, SO2, CO and PM10 on the concentration of O3 in the urban atmosphere of eastern croatia," *Environmental Modeling and Assessment*, vol. 16, 2011.

[11] N. Banan, M. T. Latif, L. Juneng, and M. F. Khan, "An application of artificial neural networks for the prediction of surface ozone concentrations in malaysia," in *From Sources to Solution*. Springer, 2014, pp. 7–12.

[12] H. Faris, M. Alkasassbeh, and A. Rodan, "Artificial neural networks for surface ozone prediction: Models and analysis." *Polish Journal of Environmental Studies*, vol. 23, no. 2, 2014.

[13] R. S. Selvaraj, K. Elampari, R. GAYATHRI, and S. J. JEYAKUMAR, "A neural network model for short term prediction of surface ozone at tropical city," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5306–5312, 2010.

[14] A. Sheta, N. Ghatasheh, and H. Faris, "Forecasting global carbon dioxide emission using auto-regressive with exogenous input and evolutionary product unit neural network models," in *Information and Communication Systems (ICICS), 2015 6th International Conference on*, April 2015, pp. 182–187.

[15] M. Alkasassbeh, A. F. Sheta, H. Faris, and H. Turabieh, "Prediction of pm10 and tsp air pollution parameters using artificial neural network autoregressive, external input models: A case study in salt, jordan," *Middle-East Journal of Scientific Research*, vol. 14, no. 7, pp. 999–1009, 2013.

[16] H. Faris and A. Sheta, "Identification of the tennessee eastman chemical process reactor using genetic programming," *International Journal of Advanced Science and Technology*, vol. 50, pp. 121–140, Jan. 2013.

[17] A. Sheta and R. Hiary, "Modeling lipase production process using artificial neural networks," in *Proceedings of the 3rd IEEE International Conference on Multimedia Computing and Systems*, Tangier, Morocco, 10-12 May 2012, pp. 1158–1163.

[18] A. Sheta and M. El-Sherif, "Optimal prediction of the nile river flow using neural networks," in *Proceedings of the International Joint Conference on Neural Networks, Washington, D.C., July*, vol. 5, 1999, pp. 3438–3441.

[19] J. Koza, "Evolving a computer program to generate random numbers using the genetic programming paradigm," in *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, La Jolla,CA, 1991.

[20] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[21] K. A. De Jong, "Adaptive system design: A genetic approach," *IEEE Transaction Sys. Man. Cybern.*, vol. 10, no. 3, pp. 556–574, 1980.

[22] K. De Jong, "Are genetic algorithms function optimizers?" in *Proceedings of the Second Parallel Problem Solving From Nature Conference*. The Netherlands: Elsevier Science Press, 1992, pp. 3–14.

[23] A. Sheta and K. D. Jong, "Parameter estimation of nonlinear systems in noisy environment using genetic algorithms," in *Proceedings of the IEEE International Symposium on Intelligent Control (ISIC'96)*, 1996, pp. 360–366.

[24] A. Sheta and K. De Jong, "Time-series forecasting using GA-tuned radial basis functions," in *Information Science Journal*, 2001, pp. 221–228.

[25] B. L. Miller, B. L. Miller, D. E. Goldberg, and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex Systems*, vol. 9, pp. 193–212, 1995.

[26] S. Legg, M. Hutter, and A. Kumar, "Tournament versus fitness uniform selection," in *Proc. 2004 Congress on Evolutionary Computation (CEC-2004)*. Portland, OR: IEEE, 2004, pp. 2144–2151.

[27] E. Galvan-Lopez, B. Cody-Kenny, L. Trujillo, and A. Kattan, "Using semantics in the selection mechanism in genetic programming: A simple method for promoting semantic diversity," in *Proceedings of the 2013 IEEE Congress on Evolutionary Computation (CEC)*, June 2013, pp. 2972–2979.

[28] D. P. Searson, D. E. Leahy, and M. J. Willis, "GPTIPS : An open source genetic programming toolbox for multigene symbolic regression," in *Proceedings of the International Multi-conference of Engineers and Computer Scientists 2010 (IMECS 2010)*, vol. 1, Hong Kong, 17-19 Mar. 2010, pp. 77–80.

[29] W. Spears and V. Anand, "A study of crossover operators in genetic programming," in *Methodologies for Intelligent Systems*, ser. Lecture Notes in Computer Science, Z. Ras and M. Zemankova, Eds. Springer Berlin Heidelberg, 1991, vol. 542, pp. 409–418.