Int. J. Hum. Capital Urban Manage., 2(1): 39-48, Winter 2017
DOI: 10.22034/ijhcum.2017.02.01.004

ORIGINAL RESEARCH PAPER

Application of gene expression programming to water dissolved oxygen concentration prediction

V. Mehdipour[1,] *, M. Memarianfard[1], F. Homayounfar[2]

[1]Department of Civil Engineering, Khaje Nasir Toosi University of Technology, Tehran, Iran
[2]Department of Civil Engineering, Amir Kabir University of Technology, Tehran, Iran

Received 15 October 2016; revised 8 November 2016; accepted 28 December 2016; available online 1 January 2017

ABSTRACT: This research based on record and collected data from four stations at Eymir Lake, Turkey, which are monitored daily in seven months. Water quality monitoring using former methods are time-needed and expensive, while the application of gene expression programming is more understandable, rapid, and reliable which is used in this article to provide a prediction for dissolved oxygen. The concentration of oxygen is one of the most important factors of water quality identification, which shows if water has proper ability for aquatic life, agriculture, sanitary and drink, or not. Therefore, the concentration of oxygen is one of the most important parameters, which cannot be calculated by mathematical analyses directly. Phosphor, nitrate, phosphate, dissolved nitrogen, water alkalinity, water temperature, dissolved chlorophyll, electrical conductivity, precipitation rate, wind velocity and environment temperature are parameters which used as correlated factors to better prediction of dissolved oxygen in this paper. In the best model determination coefficient and root mean square error values respectively, were found to be 0.8031 and 0.0937. Finally, the assessment of forecasted data showed that the proposed approach produces satisfactory results.

KEYWORDS: *Dissolved oxygen (DO); Electrical conductivity (EC); Eymir Lake; Gene expression programming (GEP); Water quality*

INTRODUCTION

High quality water deficiency is turning into a great problem for humankind and it gets worse every year. During the last century, water consumption grew at twice the rate of population increase (Salami and Ehteshami, 2015; Akilandeswari and Adline Mahiba, 2013; Kiºi, 2008) albeit the management of water salinity and consumption is not growing well. Knowing water resources, ability to predict the supply amount and demanded quality are vital factors for regional planning which any mistake consequences irreversible knots. Water management decisions are increasingly based on model studies (Razaq *et al*., 2016; Lacombe *et al*., 2014; Scholten *et al*., 2007) while modeling tools are becoming progressively more sophisticated (McKnight *et al*., 2010; Palani *et al*., 2009), Nevertheless, modeling for water quality studies are cheaper than former methods and is time saving process. The dissolved oxygen (DO) is an important quality index of water resources. However, it is difûcult to simulate the DO concentration by traditional mathematical methods due to the effects of different factors on different waters (Lihua *et al*., 2008). Oxygen amount effects on salubrity, aquatic life quality, agriculture, algal bloom and etcetera where in normal pressure and zero centigrade temperature this amount is 14.6 mg/L and it descend to 9.2mg/L in 20 centigrade.

Using gene expression programming (GEP), due to prediction of DO, is not common so far, which is makes

*Corresponding Author Email: *Vahid.mehdipour1992@gmail.com*
Tel.: +9891 4128 3041; Fax: +9841 4422 6566

this article an innovative effort. Numerous investigations have been carried out around the water quality. Among others, Rounds have worked on the DO concentration in Tualatin River (in northwest Oregon, Oswego Dam). Firstly, he explained the importance of research on DO before showing an ANN model using a feed-forward algorithm. The information about air temperature, solar radiation, rainfall, and stream ûow were considered as input, while DO concentration was the output of the model. The whole data collected between 1991 and 2001 (Rounds, 2002). Ding in a research used artificial neural network for prediction of water quality indexes and the result assessed acceptable (Ding *et al*., 2014). Liu *et al*. (2013) by combining support vector machine and artificial neural network tried to forecast water quality for aquaculture where they evaluated this method as a useful method. Alte and Sadgir studied the sodium absorption rate (SAR) prediction by ANN while using calcium, magnesium, electoral conductivity (EC), Alkalinity and sodium as input data (Alte and Sadgir, 2015).

Chu *et al*. (2013) represented an ANN model that could estimate the quality of the surface water parameters using some given parameters. The results showed that the factor analysis technique was introduced to identify important water quality parameters. Results revealed that biochemical oxygen demand, permanganate index, ammonia nitrogen, nitrogen, Cu, Zn, and Pb were the most important parameters in assessing water quality variations in the study area. This project, based on GB3838-2002 "Environmental quality standard for surface water." The model is a one-layer network using the algorithm of Hopûeld Neural Network and created by the MATLAB.

Sarkar and Pandey created a model for dissolved oxygen estimation in a river by using ANN and The correlation coefficient between laboratory and predicted data is 0.9 (Sarkar and Pandey, 2015). Singh *et al*. (2011) in a research with the title of the SVM application in water quality management tried to estimate BOD as an expensive parameter and he evaluated his work as acceptable. Azamathulla *et al*. (2011) studied on Gene-Expression Programming for the Development of a Stage-Discharge Curve of the Pahang River, Malesia. They tried to compare Genetic programming with gene expression programming where they deduced GEP more relatively successful than conventional methods like GP. The overall results confirm the use of GEP as an effective tool for forecasting and the estimation of daily discharge data. These results support the use of GEP in forecasting daily discharge values and in forecasting flood events. In a significant research, Martí *et al*. (2013) compared artificial neural networks, gene expression programming and lastly multi linear regression as approaches for estimation of outlet dissolved oxygen in micro-irrigation, and the result depicted the GEP's more usefulness in a comparison with ANN and MLR. Kisi *et al*. (2013) analyzed the data from the South Platte River at Englewood, Colorado by ANN, ANFIS and GEP for dissolved oxygen best estimation which the optimal GEP's model with respect to correlation coefficient, root mean square error and mean absolute relative error criteria, answered better and more reliable models.

MATERIALS AND METHODS

Eymir Lake occurs 20 kilometers south of the capital city, Ankara, Turkey (Yagbasan and Yazicigil, 2009) and it is one of the few natural lakes in Turkey. Lake Eymir owes its origins to the alluvial damming of the Imrahor River Valley, which led to the formation of two lakes, the upstream Lake Mogan and the downstream Lake Eymir. Lake Mogan empties into Lake Eymir at the southwest corner, forming the main inflow of Lake Eymir (Beklioglu *et al*., 2003). The semi-arid dry climatic condition of Central Anatolian region influences Lake Eymir. The climate typically includes cold winters and hot summers. Fifteen-year average (1984–1999) of the annual precipitation measured 390 ± 76 mm with the maximum precipitation recorded in December and May, and the minimum recorded in August. For the first time in 1990 this lake gone under the protection of Turkey environment agency where the excessive eutrophication in warm seasons and dissolved oxygen deficiency are the greatest problems of this lake (Altinbilek *et al*., 1995). The lake is relatively shallow (Z: 3.1m) and large (125 ha) with a long hydraulic retention time (1.8–23years). Data used in this study are taken at seven months daily from four different points where it is illustrated on below pictures. Fig. 1 shows the Lake Eymir approximately place in Turkey in Fig. 2 four different points where they are shown data collecting stations.

Because of algal high-level activity in water surface, it is argumentative to collect water samples from water surface instead of middle or bottom of the lake (Jain

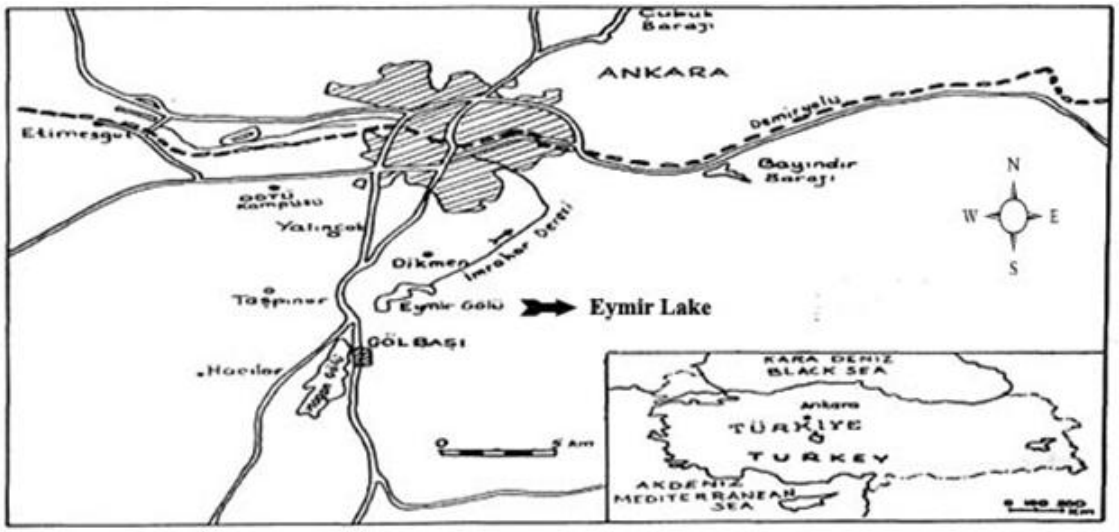

Fig. 1: Lake Eymir position in Turkey

Fig. 2: Four data collection points among the lake Eymir

and Kumar, 2007; Tayfur, 2002), so the data were collected from four stations through the seven months. To create input models once TKN, NO3, PO4 and TP assumed as central core which these parameters are effective on algal activities. Continue other parameters added one by one to this central core to create other input models. In the second type, PO4 and NO3 took as another central core and one more time other parameters added to core to create different models. About non-coral parameter, adding to core parameters the adding order planned to observe the parameters efficacies which it can be seen in the result section that the most effective and non-effective parameters are classified. Table 1 illustrates input models.

Gene expression programming (GEP) is similar to genetic algorithms (GAs) and genetic programming (GP), encouraged by the syntax of the organism's chromosomes. Selects them according to adaptability, and introduces genetic conversion using some genetic operators (Mitchell, 1996; Pour *et al*., 2014; Koza, 1992).

Table 1 : Input model

Model name	Input Parameters
Input combination A1	TKN, NO3, TP, PO4
Input combination A2	TKN, NO3, TP, PO4, alkalinity
Input combination A3	TKN, NO3, TP, PO4, alkalinity, temperature
Input combination A4	TKN, NO3, TP, PO4, alkalinity, temperature, PH
Input combination A5	TKN, NO3, TP, PO4, alkalinity, temperature, PH, chlorophyll- a
Input combination A6	TKN, NO3, TP, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity
Input combination A7	TKN, NO3, TP, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity, precipitation
Input combination A8	TKN, NO3, TP, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity, precipitation, wind
Input combination A9	TKN, NO3, TP, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity, precipitation, wind, ambient air temp
Input combination B1	NO3, PO4
Input combination B2	NO3, PO4, alkalinity
Input combination B3	NO3, PO4, alkalinity, temperature
Input combination B4	NO3, PO4, alkalinity, temperature, PH
Input combination B5	NO3, PO4, alkalinity, temperature, PH, chlorophyll- a
Input combination B6	NO3, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity
Input combination B7	NO3, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity, precipitation
Input combination B8	NO3, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity, precipitation, wind
Input combination B9	NO3, PO4, alkalinity, temperature, PH, chlorophyll- a, conductivity, precipitation, wind, ambient air temp

Gene expression programming was developed by Ferreira (2001), the base of this method is the generation and assessment of its ability for producing new generation. It uses populations of individuals, then selects the fit ones and introduces genetic variation using genetic operators.

The basic difference between the Genetic Algorithms and Gene expression programming is the nature of the individuals; in GAs the individuals are linear threads of fixed length which are named chromosomes; in GP the individuals are nonlinear entities of different sizes and forms (parse trees) and in GEP the individuals are encoded as linear threads of constant length chromosomes (Ferreira, 2002).

GEP commences the solution of a particular problem with the random birth of the chromosomes of the starting population. In the next step, chromosomes are expressed and the compatibility of every individual is measured. Selected individuals reproduce with modification, leaving child with new attributes. The individuals of this new generation are in their turn, subjected to the equal progressive process: expression of the genomes, an encounter with the selection ambient and reproduction with modification. The operation is repeated for a determined number of generations or until a solution is found (Ferreira, 2001; Karimi *et al*., 2016). A brief flowchart of GEP is shown in Fig. 3. Due to evaluation applied model and scrutiny of estimated data in this paper, we used root mean square error (RMSE) which represents the sample standard deviation of the differences between computed and observed values correlation coefficient (R) and

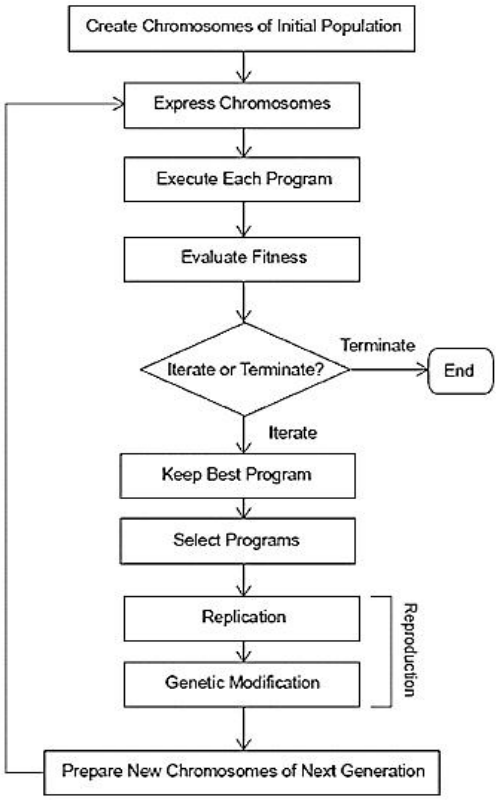

Fig. 3: The flowchart of a gene expression algorithm (Ferreira, 2006)

Table 2: Evaluation Criteria Table

Name	Equation
Root Mean Square Error	$RMSE = \sqrt{\sum_{i=\backslash}^{N} \frac{(y_m - y_p)^2}{N}}$
Correlation Coefficient	$R = \frac{\sum_{i=1}^{N}(y_m - \bar{y}_m)\times(y_p - \bar{y}_p)}{\sqrt{\sum_{i=1}^{N}(y_m - \bar{y}_m)^2} \times \sqrt{\sum_{i=1}^{N}(y_p - \bar{y}_p)^2}}$
Determination Coefficient	$DC = 1\text{-}\frac{\sum_{i=1}^{N}(y_m - y_p)^2}{\sum_{i=1}^{N}(y_m - \bar{y}_m)^2}$

y_m, y_p : *observed and predicted dissolved oxygen are respectively*

$\bar{y}_m, \bar{y}_p$: *average of observed and predicted dissolved oxygen are respectively*

N : *number of data*

determination coefficient (DC) as evaluation criteria which are shown in Table 2.

As a usual technique in the GEP overtraining solution, the early stopping technique is used. The added text in Materials and Methods section will be:

To evade the problem of overtraining datasets, the early halting technique is used. On the other hand, training datasets separated into two groups: one for learning and one for cross-validation. As the MSE of cross-validation data series was minor than prior iteration training proceeded, conversely, if mentioned error is more than before, training was completed.

Data preprocessing: The aim of pre-processing or normalization of data is creating comparable value of all available data (Simeonov *et al*., 2002). Generally, normalization is a change of data into limited ranges. All data in this paper have been written into [0-1] and been rated in this limit. The used equation for this pre-processing data is:

$$X = 0.1 + 0.9\frac{(X_i - X_{\min})}{(X_{\max} - X_{\min})} \quad (1)$$

RESULTS AND DISCUSSION

In this study gene expression programming applied to estimate the dissolved oxygen as an important quality indicator of water. After preprocessing data limited in range of [0-1] and then divided into two data, test and train. Next step different models of different inputs were defined and modeling process completed. Table 3 sums up the validation statistics for prediction dissolved oxygen for testing datasets.

The results of this disquisition show that DO estimation of water surface is possible by using some parameters. As can be seen from Table 3, input combination B9 with nitrogen and phosphate as coral parameters and water alkalinity, water temperature, Chlorophyll, electoral conductivity, precipitation, wind speed and air temperature as added parameters gives the most accurate simulations of DO estimation with.

According to Table 3 input combination with nitrate and phosphate (NO3 and PO4) as an input parameters lead to better performance with regard to other combinations. Comparing B3 and B4 with B1 and B2 demonstrates PH positive effects on DO estimation.

Table 3: Evaluation criteria of different models applied to estimation of dissolved oxygen (test data)

Input combination	R	DC	RMSE
Input combination A1	0.3523	0.1590	0.1482
Input combination A2	0.3626	0.1621	0.1433
Input combination A3	0.5678	0.2594	0.1291
Input combination A4	0.6196	0.4015	0.1181
Input combination A5	0.2764	0.1211	0.1724
Input combination A6	0.4845	0.2267	0.1497
Input combination A7	0.5130	0.4576	0.1324
Input combination A8	0.6488	0.5253	0.1149
Input combination A9	0.5792	0.4891	0.1237
Input combination B1	0.3715	0.1613	0.1378
Input combination B2	0.4450	0.1667	0.1351
Input combination B3	0.6104	0.3682	0.1111
Input combination B4	0.6813	0.3578	0.1120
Input combination B5	0.6825	0.3170	0.1155
Input combination B6	0.6382	0.3472	0.1135
Input combination B7	0.7028	0.3785	0.1102
Input combination B8	0.7280	0.3831	0.1087
Input combination B9	0.8031	0.5511	0.0937

Furthermore, it could be pointed that Chlorophyll, EC, precipitation and wind speed have not practical effects on DO prediction. At last, in model B9 adding water temperature impressively meliorates mentioned input model. The scatter plot of dissolved oxygen simulations for testing dataset is plotted versus observed values in Fig. 4. In addition, Fig. 5 demonstrates comparisons between observed and simulated data with GEP's best model (B9) results.

As a result of comparison spots of input combination B9 are nearer to the line between observed DO and simulated DO. The line shows the best status of models in which the observed and simulated data are accurately equal as shown in Fig. 4. Spots are around the line where in other input combinations spots are more scattered. This programming seeks the optimum stance, but the complexity of dissolved oxygen of water and its dependency of numerous parameters cause models inaccurate. Fig. 5 tries to show that the model is learning the observed data path and it simulates more accurately while data number adding during the days. Selected model in elementary days follows the observed data,

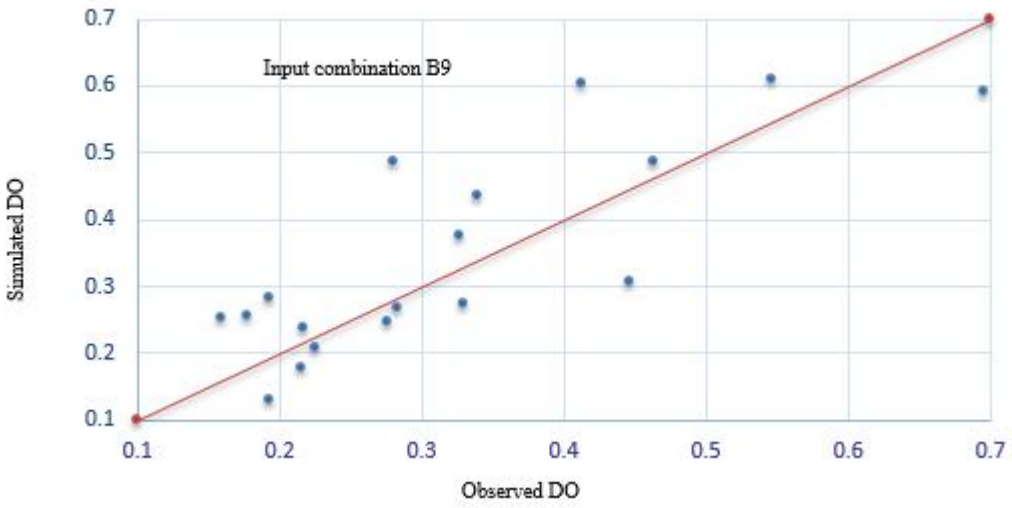

Fig. 4: linear relation between observed and forecasted data

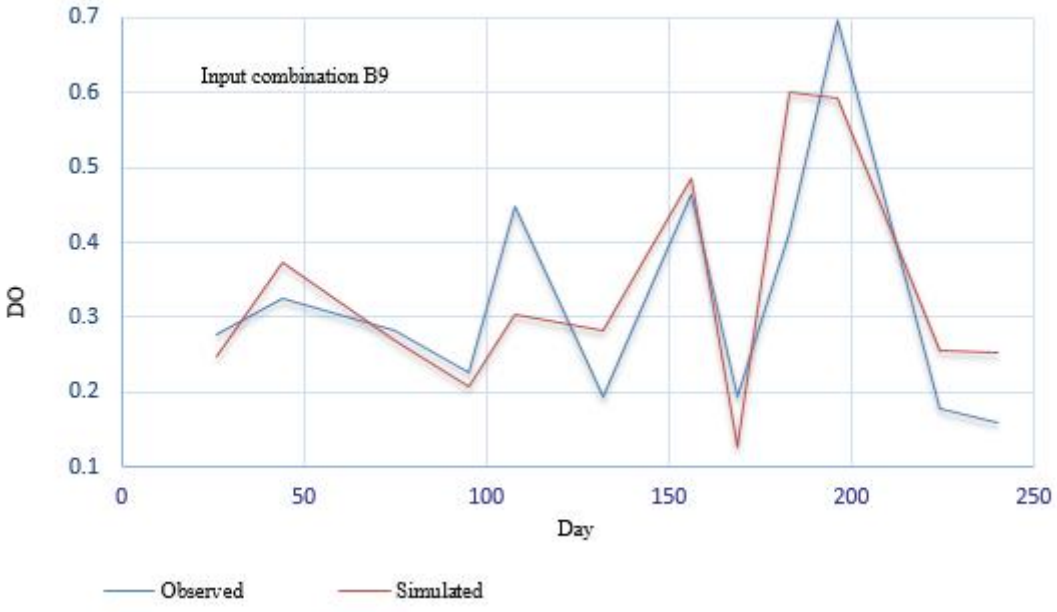

Fig. 5: Daily observed comparison of and simulated DO amount

but it is not able to predict precisely abrupt changes. Of course, data augmentation helps model to train more where it causes accuracy.

CONCLUSION

In this article, the abilities of GEP approach were investigated to predict water dissolved oxygen (DO) using Eymir Lake datasets. The data from four stations on this lake used for training and testing datasets. Different models compared with each other with respect to performance criteria such as the correlation coefficient, Nash–Sutcliffe and root mean square error. In every model, the most effective parameters are taken as coral parameters where other parameters added one by one. As the parallel effort, coral parameters changed and in next steps the number of cores changed either, which comparison of results demonstrates PH, water temperature and air temperature have the key roles on dissolved oxygen forecasting. The input combination B9 with a NO3-PO4 as a core input parameter evaluated as the most accurate model which answers the minimum amount of RMSE (=0.0937) and maximum amount of R (=0.8031) and DC (=0.5511). Finally, it can be inferred that gene expression programming (GEP) can establish a highly accurate models to predict objective function.

ACKNOWLEDGEMENT

The authors are grateful to Department of Civil Engineering in Khajeh Nasir Toosi University of Technology and dear Professor F. Vafayi for providing software facilities.

CONFLICT OF INTEREST

The author declares that there is no conflict of interests regarding the publication of this manuscript.

REFERENCES

Akilandeswari, S.; Adline Mahiba, H., (2013). Prediction of BOD values in engineering work industrial efûuent by Anûs modeling. Int. J. Res. Pure Appl. Phys., 3(2): 7–9 (**3 pages**).

Alte, P. D.; Sadgir, P. A., (2015). Water quality prediction by using ANN. Int. J. Adv. Found. Res. In Sci. Eng., 1: 278-285 (**8 pages**).

Altinbilek, D..; Usul, N.; Yazicioglu, H.; Kutoglu, Y.; Merzi, N.; Gögüs, M.; Doyuran, V. ; Günyakti, A., (1995). Water resources and environment management plan project for Gölvesi-Mogan-Ayrir Lake. In Mogan and Eymir lakes first environment Conference, 13–21 (**9 pages**). (In Turkish)

Azamathulla, H. M.; Ghani, A. A.; Leow, C. S.; Chang, C. K.; Zakaria, N. A., (2011). Gene-expression programming for the development of a stage-discharge curve of the Pahang River. Water resour. Manage., 25(11): 2901-2916 (**16 pages**).

Beklioglu, M.; Ince, O.; Tuzun, I., (2003). Restoration of the eutrophic Lake Eymir, Turkey, by biomanipulation after a major external nutrient control I. Hydrobiologia, 490(1): 93-105 (**13 pages**).

Chu, H.B.; Lu, W.X.; Zhang, L., (2013). Application of artiûcial neural network in environmental water quality assessment. J. Agric. Sci. Technol. 15(2): 343–356 (**14 pages**)

Ding, Y.R.; Cai, Y.J.; Sun, P.D.; Chen, B., (2014). The use of combined neural network and genetic algorithms for prediction of water quality. J. Appl. Res. Technol., 12(3) : 493–499 (**7 pages**)

Ferreira, C., (2001). Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst., 13(2): 87–129 (**43 pages**).

Ferreira, C., (2002). Gene expression programming in problem solving. Part VI, In Soft computing and industry, Springer, London, 635-653 (**19 pages**).

Ferreira, C., (2006). Gene expression programming: mathematical modeling by an artificial intelligence, Vol. 21. Berlin-Heidelberg, Springer-Verlag.

Jain, A.; Kumar, A. M., (2007). Hybrid neural network models for hydrologic time series forecasting. Appl. Soft Comput., 7(2): 585-592 (**8 pages**).

Karimi, S.; Shiri, J.; Kisi, O.; Shiri, A.A., (2016). Short-term and long-term streamflow prediction by using'wavelet–gene expression'programming approach. J. Hydraul. Eng., 22(2): 148-162 (**15 pages**).

Kiºi, Ö., (2008). River flow forecasting and estimation using different artificial neural network techniques. Hydrol. Res., 39(1): 27-40 (**14 pages**).

Kisi, O.; Akbari, N.; Sanatipour, M.; Hashemi, A.; Teimourzadeh, K.; Shiri, J., (2013). Modeling of Dissolved Oxygen in River Water Using Artificial Intelligence Techniques. J. Environ. Inf., 22 (2): 92-101 (**10 pages**).

Koza, J.R., (1992). Genetic programming: on the programming of computers by means of natural selection, Vol. 1, MIT Press, Cambridge.

Lacombe, G.; Douangsavanh, S.; Vogel, R. M.; McCartney, M.; Chemin, Y.; Rebelo, L. M.; Sotoukee, T., (2014). Multivariate power-law models for streamflow prediction in the Mekong Basin. J. Hydrol.: Reg. Stud., 2: 35-48 (**14 pages**).

Lihua, C.; Shengquan, M.; Li, L., (2008). A model to evaluate do of river based on artificial neural network and stylebook. J. Hainan Normal Univ. (Nat. Sci.), 21(4) : 372–376 (**5 pages**).

Liu, S.; Tai, H.; Ding, Q.; Li, D.; Xu, L.; Wei, Y., (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Math. Comput. Modell., 58(3): 458-465 (**8 pages**).

Martí, P.; Shiri, J.; Duran-Ros, M.; Arbat, G.; De Cartagena, F.R.; Puig-Bargués, J., (2013). Artificial neural networks vs. gene expression programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents. Comput. Electron. Agric., 99: 176-185 (**10 pages**).

McKnight, U. S.; Funder, S. G.; Rasmussen, J. J.; Finkel, M.; Binning, P. J.; Bjerg, P. L., (2010). An integrated model for assessing the risk of TCE groundwater contamination to human receptors and surface water ecosystems. Ecol. Eng., 36(9): 1126-1137 (**12 pages**).

Mitchell, M., (1996). An Introduction to Genetic Algorithms, MIT Press Cambridge, MA, USA.

Palani, S.; Liong, S.Y.; Tkalich, P.; Palanichamy, J., (2009). Development of a neural network model for dissolved oxygen in seawater. Indian J. Mar. Sci., 38(2): 151-159 (**9 pages**).

Pour, S. H.; Harun, S. B.; Shahid, S., (2014). Genetic programming for the downscaling of extreme rainfall events on the East Coast of Peninsular Malaysia. Atmos., 5(4): 914-936 (**23 pages**).

Razaq, S. A.; Shahid, S.; Ismail, T.; Chung, E. S.; Mohsenipour, M.; Wang, X. J., (2016). Prediction of flow duration curve in ungauged catchments using genetic expression programming. Procedia Eng., 154: 1431-1438 (**8 pages**).

Rounds, S. A., (2002). Development of a neural network model for dissolved oxygen in the Tualatin river, Oregon, In Second Federal Interagency hydrologic modeling conference, Las Vegas, Nevada, 1–13 (**13 pages**).

Salami, E. S.; Ehteshami, M., (2015). Simulation, evaluation and prediction modeling of river water quality properties (case study: Ireland Rivers). Int. J. Environ. Sci. Technol., 12(10): 3235–3242 (**8 Pages**).

Sarkar, A.; Pandey, P., (2015). River water quality modelling using artificial neural network technique. Aquatic Procedia, 4: 1070-1077 (**8 pages**).

Scholten, H.; Kassahun, A.; Refsgaard, J. C.; Kargas, T.; Gavardinas, C.; Beulens, A. J. M., (2007). A methodology to support multidisciplinary model-based water management. Environ. Modell. Software. 22(5) : 743–759 (**17 pages**).

Simeonov, V.; Einax, J.; Stanimirova, I.; Kraft, J., (2002). Environmetric modeling and interpretation of river water monitoring data. Anal. Bioanal.Chem, 374(5): 898-905 (**8 pages**).

Singh, K. P.; Basant, N.; Gupta, S., (2011). Support vector machines in water quality management. Anal. Chim. Acta, 703(2): 152-162 (**11 pages**).

Tayfur, G., (2002). Artificial neural networks for sheet sediment transport. Hydrol. Sci. J., 47(6): 879-892 (**14 pages**).

Yagbasan, O.; Yazicigil, H., (2009). Sustainable management of Mogan and Eymir Lakes in central Turkey. Environ. Geol., 56(6): 1029-1040 (**12 pages**).