Uludağ University Journal of The Faculty of Engineering, Vol. 23, No. 2, 2018 RESEARCH

DOI: 10.17482/uumfd.352833

STREAMFLOW AND SEDIMENT LOAD PREDICTION USING LINEAR GENETIC PROGRAMMING

Ali DANANNDEH MEHR [*]
Ali Ünal ŞORMAN [**]

Received: 14.11.2017; revised: 20.06.2018; accepted: 17.07.2018

Abstract: Daily flow and suspended sediment discharge are two major hydrological variables that affect rivers' morphology and ecosystem, particularly during flood events. Artificial neural networks (ANNs) have been successfully used to model and predict these variables in recent studies. However, these are implicit and cannot be simply used in practice. In this paper, linear genetic programming (LGP) approach has been suggested to develop explicit models to predict these variables in two rivers in Iran. The explicit relationships (prediction rules) evolved by LGP take the form of equations or program codes, which can be checked for its physical consistency. The results showed that the LGP outperforms ANNs to get global maximum and minimum discharges providing lowest root mean squared error and higher coefficient of efficiency both for training and validation periods.

Keywords: Daily discharge, sediment, prediction, linear genetic programming, artificial neural networks

Akım ve Sediment Yük Öngörümü İçin Doğrusal Genetik Programlamanın Uygulanması

Öz: Nehirlerin morfolojisini, ekosistemi ve özellikle taşkın olaylarını etkileyen iki ana değişken askıdaki sediment ve günlük akımlardır. Yapay sinir ağları (YSA), bu değişkenleri modellemek ve tahmin etmek için yakın zamanda yapılmış çalışmalarda başarıyla kullanılmıştır. Bununla birlikte, bunlar kapalı yöntemlerdir ve pratik uygulamalarda kolaylıkla kullanılamazlar. Bu makalede, İran'daki iki nehirde bu değişkenleri tahmin etmek üzere açık modeller geliştirmek için doğrusal genetik programlama (DGP) yaklaşımı önerilmiştir. DGP tarafından geliştirilen açık ilişkiler (tahmin kuralları), fiziksel tutarlılığı açısından kontrol edilebilen denklemler veya program kodları şeklindedir. Sonuçlar, global maksimum ve minimum akımları elde etme noktasında, DGP'nin YSA'ya göre daha başarılı olduğunu gerek kalibrasyon gerekse doğrulama aşamalarında hataların karelerinin ortalamasının karekökünün en düşük, verimlilik katsayısının ise daha yüksek olmasını sağlayarak göstermiştir.

Anahtar Kelimeler: Günlük akım, Sediment, Öngörüm, Doğrusal Genetik Pogramlama, Yapay sinir ağları

1. INTRODUCTION

Accurate prediction of hydrological variables such as daily streamflow and suspended sediment discharge plays an important role in floodplain management and river engineering. Many of the activities associated with the planning and operation of river systems require accurate prediction of flow characteristics. It is generally accepted that river flow variables, especially daily flow and sediment discharge have nonlinear behavior. Thus, accurate prediction of such variables can be a challenging task, especially during high flow periods. Several linear

[*] Department of Civil Engineering, Antalya Bilim University, Antalya, Turkey.
[**] Faculty of Civil Engineering, Middle East Technical University, Ankara, Turkey.
Corresponding Author: Ali Danandeh Mehr (ali.danandeh@antalya.edu.tr)

and nonlinear methods have been applied in the prediction of discharge and sediment transport in rivers and successful results have been reported. Most of the earlier studies have focused on the prediction of discharge based on stage-discharge, rainfall-discharge or time-series of discharge relationships, using either conventional methods or soft computing techniques such as artificial neural networks (ANNs), genetic programming (GP), and fuzzy logic (FL) (e.g., Kisi and Cigizoglu 2007; Aytek and Kisi 2008; Guven 2009; Danandeh Mehr et al. 2013; Danandeh Mehr and Demirel 2016; Danandeh Mehr and Kahya 2017).

In recent studies, GP has been pronounced as a robust alternative for the modelling of environmental process (Guven et al. 2008; Uyumaz et al. 2014; Roushangar and Homayounfar 2015; Danandeh Mehr and Nourani 2017). For example, Babovic and Keijzer (2002) applied GP to rainfall-runoff modeling and Giustolisi (2004) showed that GP can be successfully used to determine Chezy resistance coefficient in corrugated channels. It was observed that only few studies existed in the relevant literature related to the use of linear GP (LGP) in the field of environmental studies. For instance, Aytek and Kisi (2008) used LGP for suspended sediment modeling at two stations on the Tongue River in Montana, USA, and indicated that LGP formulation performs quite well compared to sediment rating curves and multi-linear regression models. In another study, Danandeh Mehr et al. (2014) showed that LGP can be used to model monthly streamflow between two successive stations on Çoruh River, Turkey. Tofiq and Guven (2014) explored the capability of LGP for creating quantitative relationship between large-scale climate variables (including NCEP re-analysis data and Coupled Global Climate Model CGCM3.1 outputs) and local-scale discharge flowing to Darbandikhan Dam, Iraq, as predictand variable in the statistical downscaling. The study demonstrated that transforming the discharge data through natural logarithm improves the performance of the LGP. In addition, the results showed that NCEP predictors have better correlation with the dam inflow data than the CGCM3 predictors. More recently, a wavelet–LGP integration has been used by Ravansalar et al. (2017) to model and forecast monthly streamflow in Beshar River, Iran. The authors showed that discrete wavelet decomposition of flow time series can significantly increase forecasting accuracy of LGP.

Our review showed both LGP and ANN are well enough to model variety of hydrological phenomena. However, more studies are required to compare pros and cons of these techniques. Thus, the main aim of the present research is to investigate/compare the capability of the techniques to predict daily streamflow and suspended sediment discharge. To this end, two case study applications are demonstrated in the following sections.

2. THEORETICAL CONSIDERATIONS

2.1. Linear GP (LGP)

Genetic programming (Koza 1992) is a development for genetic algorithm. The main difference between genetic programming and genetic algorithm is the representation of the solution. Genetic programming creates computer programs in the lisp or any other computer languages as the solution; whereas genetic algorithm creates a string of numbers that represents the solution (Olyaie et al. 2017). GP uses four steps to solve problems. (i) Generate an initial population of random combinations of the functions and terminals of the problem (computer programs), (ii) execute each program in the population and assign it a fitness value according to how well the program solves the problem, (iii) create new population of computer programs using genetic operators including crossover, mutation, and reproduction, and (iv) select the best computer program in the population, the best-so-far solution. To the fundamental of GP, the reader is referred to Koza (1992).

The LGP is an advancement of GP that uses fitness-based tournament selection to continuously improve a population of machine-code functions. In other words, the LGP is based on efficient GP processes using a linear genome. While the GP holds candidate solutions

(programs) in a tree structure (see Figure 1) and the genetic operators (crossover and mutation) act on tree nodes, in LGP transformation operators act on a linear (not tree-based) genome.

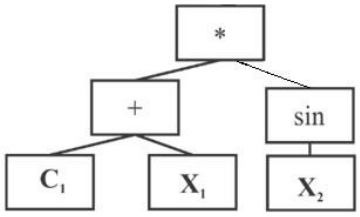

Figure 1:
Tree representation of the computer programs in GP representing $(C_1+X_1) \times Sin\ (X_2)$

An example of LGP evolved program in C language describing the flow of water (Q) through porous media, the well-known Darcy's Law: Q=K.I.A, is illustrated as follows (Hrnijca and Danandeh Mehr 2019):

```
L0:  f [0] = 0.0;

L1:  f [0] += I;

L2:  f [0] *= K;

L3:  f [0] *= A;
```

where I = pressure gradient, K = hydraulic conductivity, A = area, and f[0] is an accumulator variable representing the final output (i.e. Q) of the evolved program. LGP employs such temporary variable to store values while performing calculations (Uyumaz et al. 2014). The temporary variable equals to zero by default and the output is the value remaining in it in the last line of the program. It should be mentioned that in this program, evolving introns have been removed previously. In analogy with natural introns, deoxyribonucleic acid (DNA) parts of genes with information that are not expressed in proteins, an intron in LGP is defined as a program portion without any effect on the calculation of the output(s) for all possible inputs. A simple examples of an introns is:

```
L0:  f [0] += -1.00f;

L1:  f [0] += +1.00f;
```

2.2. Artificial Neural Networks (ANNs)

ANNs are flexible regression methods in which a modeler uses input and output data sets to figure out the system attitude. Feed-forward backpropagation (FFBP) is probably of the most popular ANNs in hydro-environmental applications (Danandeh Mehr et al. 2015), which considered as general nonlinear approximation. FFBP is a supervised learning technique, meaning that the desired outputs are known in advance. The network generates the desired outputs from the inputs by minimizing the estimation error using a set of synaptic weights. FFBP networks typically contain three parts: a) input layer comprising a number of input nodes, b) one or more hidden layers and c) a number of output layer nodes. The number of hidden layers and nodes are key design parameters of FFBP. The design issues, training mechanisms and application of FFBP in hydrological modelling have been the subject for plenty of studies in recent three decades. To avoid redundancy, we refer the readers to Sajikumara and

Thandaveswara (1999), Abrahart et al. (2012) and Danandeh Mehr et al. (2015). An FFBP network with one hidden layer is illustrated in Figure 2. It shows that a neuron connection only exist from a neuron in the input layer to other neurons in the hidden layer or from a neuron in the hidden layer to other neurons in the subsequent output layer. The letters M, N and O in the figure denote the number of neurons in input, hidden and output layers, respectively. The weights are different in the hidden and output layers, and their values are adjusted during the back propagation training process.

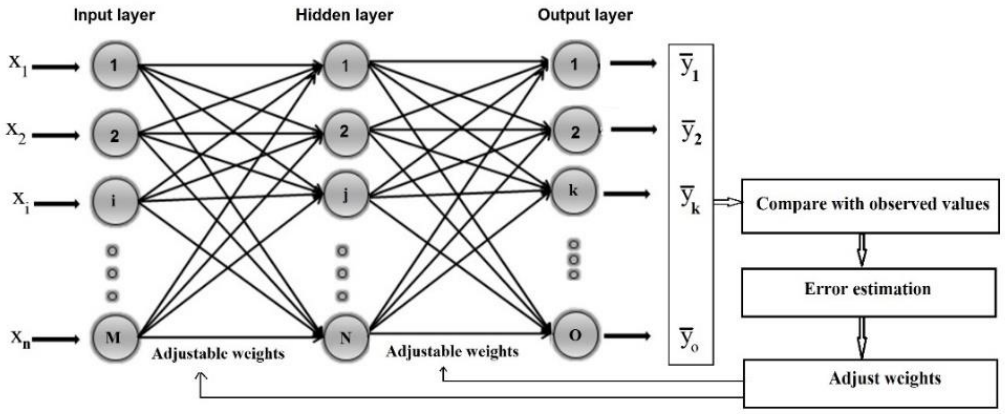

Figure 2:
A three-layered FFBP network used in the study

3. CASE STUDIES

The study area includes Lighvan Chai and Absardeh rivers located in the northwest and west part of Iran, respectively (Figure 3). The daily flow and suspended sediment discharge data of Lighvan Station (37º 55' N, 46º 22' E) on Lighvan Chai River operated by Iran Ministry of Energy (MOE) were used for suspended sediment prediction. Daily flow discharge data of Mohammad haji Station (33º 44' 13" N, 48º 45' 15" E) on Absardeh River operated by MOE was used for streamflow forecasting. The locations of these stations are shown in Figure 3. For Lighvan Station, the data from January 1998 to December 2003 (6 water years) were used for modelling. The first five years were used for model training and the last year (2003 water year) was used for validation. For Absardeh River, the data from January 2004 to December 2007 (4 water years) were utilized for modelling. The first three years were chosen for calibration and the data of last year (2007 water year) was used for validation. The statistical parameters of observed flow and sediment load at Lighvan and Mohammad haji stations are given in Table1.

Before applying the LGP and ANN methods, all the input/target data were normalized to rescale in the range [0.1, 0.9]. The river flow and suspended sediment load were normalized by the following formula suggested by Danandehmehr et al. (2013):

$$X_n = 0.1 + 0.8 \times \left(\frac{X_O - X_{\min}}{X_{\max} - X_{\min}} \right) \quad (1)$$

where X_n = normalized data, X_{max} = maximum of the data values, X_{min} = minimum of the data values and X_o= observed data.

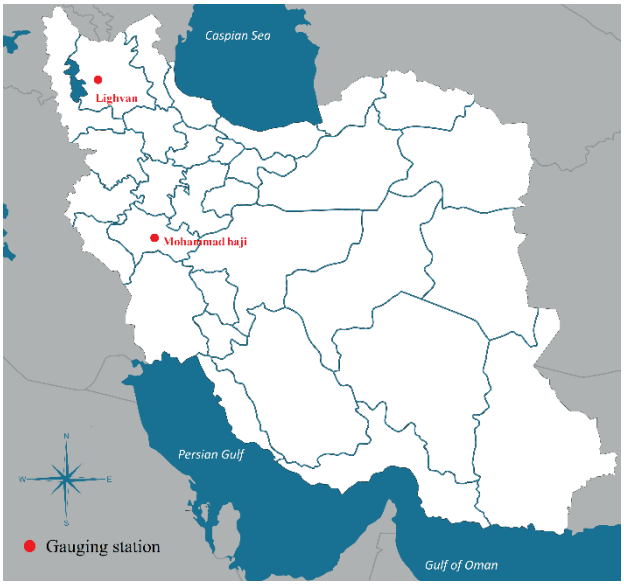

Figure 3:
Locations of the hydrometric stations used in the study

Table 1. The daily statistical parameters of observed flow and sediment data

Station	Lighvan		Mohammad haji
Description	Flow Discharge (m^3/s)	Sediment load (ton/day)	Flow Discharge (m^3/s)
Number of data	2187	2187	1460
Maximum	5.38	124.57	82.2
Minimum	0.0	0.0	0.2
Average	0.66	6.32	2.36
Variance	0.54	105.65	19.33
Standard Deviation	0.74	10.28	4.4

4. MODELS APPLICATION AND DISCUSSION OF RESULTS

The LGP commercial software, namely Discipulus (Francone 2001), was used in the present study to create both flow and discharge prediction programs. Here, 50 runs were performed to produce a wide range of models that use basic arithmetic operations together with random constants. Termination of each run was also considered as 100 generations without any improvement in fitness function. At each run, 30 best programs were selected and added to a pool of solutions. The best and ultimate solution is selected among 1500 programs (50*30=1500) available in the pool based on root mean square error (RMSE) statistic. ANN models were created using neural network toolbox of MATLAB software so that the optimum number of hidden neurons was obtained via trial and error procedure between number of neurons and the associated model accuracy as suggested by Danandeh Mehr et al. (2015). To avoid overfitting in LGP validation of the best models were done on the unseen validation dataset as was defined earlier in Section 3. In ANN runs, Different network structures were trained using Levenberg-Marquardt algorithm. At each epoch up to 1000 iterations were done and training was stopped when the validation error starts to increase.

Several input combinations are tested using LGP and ANN to estimate suspended sediment load and daily discharge from the collected data about suspended sediment and daily flow discharge time series at each station. The combinations for Lighvan station are:

(i) $S_t = f(Q_t)$

(ii) $S_t = f(Q_t, S_{t-1}, Q_{t-1})$

(iii) $S_t = f(Q_t, S_{t-1}, Q_{t-1}, S_{t-2}, Q_{t-2})$

(iv) $S_t = f(Q_t, S_{t-1}, Q_{t-1}, S_{t-2}, Q_{t-2}, S_{t-3}, Q_{t-3})$

(v) $S_t = f(Q_t, S_{t-1}, Q_{t-1}, S_{t-2}, Q_{t-2}, S_{t-3}, Q_{t-3}, S_{t-4}, Q_{t-4})$

(vi) $S_t = f(Q_t, S_{t-1}, Q_{t-1}, S_{t-2}, Q_{t-2}, S_{t-3}, Q_{t-3}, S_{t-4}, Q_{t-4}, S_{t-5}, Q_{t-5})$

where Q_t and S_t denote the discharge and suspended sediment load at time t, respectively.

The dimensionless values of RMSE and correlation coefficient (R) of LGP and ANN models in training and validation periods are compared in Table 2 and Table 3 for the Lighvan Station. As shown in Table 2, Combination iii provides the best LGP model having the lowest RMSE (0.002) at validation period. In this combination, the effective inputs are the current discharge as well as the flow discharge and suspended sediment load in two antecedent days. The LGP performance for the first input combination (having only current discharge) is the worst due to the hysteresis effect between sediment load and discharge. It implies that the suspended sediment for a given level of streamflow in the rising stage of a flow hydrograph is greater than in the falling stage. It is worth to mention a high value of R in this combination (= 0.951) indicates a positive correlation between model and observed sediment load. Thus, it cannot be a signal for a perfect prediction. This is also the case for the first combination modeled by ANN.

Table 2. Efficiency results of LGP models developed for Lighvan station at training and validation period

Input model	Iteration	Training		Validation	
		RMSE	R	RMSE	R
i	50	0.045	0.878	0.044	0.951
ii	50	0.015	0.982	0.003	0.914
iii	50	0.001	0.998	0.002	0.953
iv	50	0.016	0.979	0.006	0.784
v	50	0.017	0.974	0.006	0.761
vi	50	0.018	0.973	0.007	0.758

Table 3. Efficiency results of ANN models developed for Lighvan station at training and validation period

Input model	Nodes in hidden layer	Training		Validation	
		RMSE	R	RMSE	R
i	3	0.036	0.864	0.117	0.941
ii	4	0.018	0.967	0.029	0.912
iii	5	0.016	0.974	0.023	0.892
iv	4	0.016	0.975	0.021	0.861
v	3	0.015	0.977	0.021	0.891
vi	5	0.018	0.966	0.022	0.871

Among the ANN models, the combination (V) exhibited the best performance. The time series and scatter plot of the observed versus the LGP (Combination iii) and ANN (Combination V) forecasts are illustrated in Figure 4. This figure shows that the LGP model has better accuracy than the ANN model.

Similar to the sediment load prediction combinations, five input combinations are tried to estimate daily discharge flow from daily discharge time series at Mohammad haji station. The assumed combinations are:

(i) $Q_t = f(Q_{t-1})$

(ii) $Q_t = f(Q_{t-1}, Q_{t-2})$

(iii) $Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3})$

(iv) $Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4})$

(v) $Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5})$

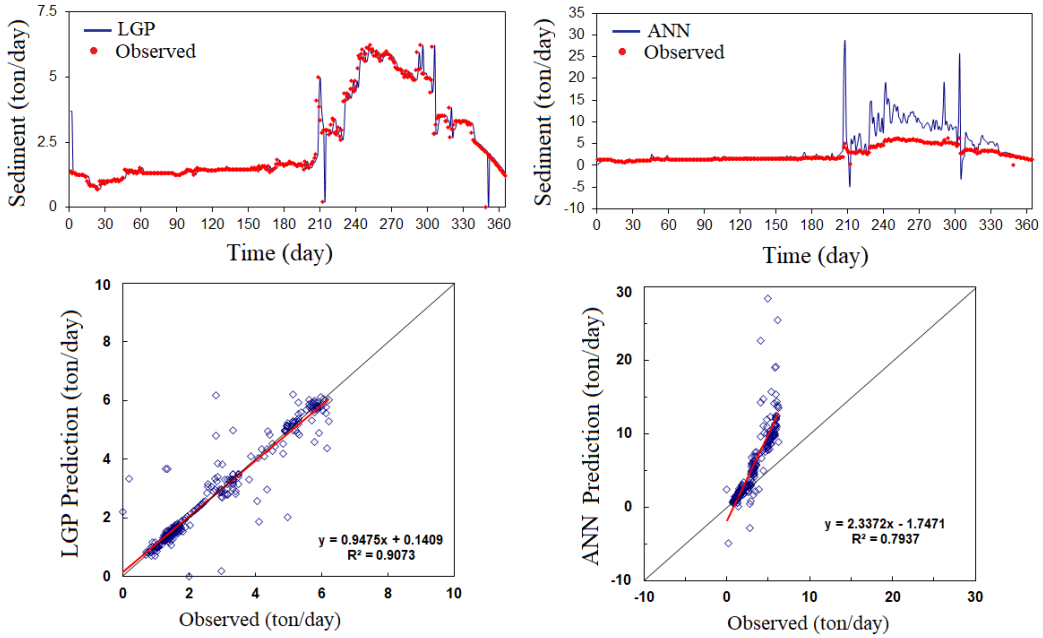

Figure4:

Observed and computed suspended sediment load by LGP and ANN model (Lighvan station)

The dimensionless values of RMSE and R of LGP and ANN models in training and validation period for Mohammad haji are given in Table 4 and Table 5, respectively. As seen from Table 4, the LGP model (Combination iii) whose inputs are current discharge and three previous discharges has the lowest RMSE (0.018) and the highest linear correlation R (0.97). This combination provides Nash–Sutcliffe efficiency coefficient around 0.94. Also from Table 5, the combination iii has the lowest RMSE (0.021) and the corresponding R = 0.95. Among the ANN models, combination iii provides Nash–Sutcliffe efficiency coefficient around 0.90. Higher accuracy of the LGP model in comparison to the ANN model is shown in Figure 5.

Table 4. Efficiency results of LGP models developed for Mohammad haji station in training and validation period

Input model	iteration	Training		Validation	
		RMSE	R	RMSE	R
i	50	0.017	0.92	0.025	0.94
ii	50	0.018	0.92	0.018	0.97
iii	50	0.014	0.96	0.018	0.97
iv	50	0.018	0.92	0.022	0.95
v	50	0.016	0.93	0.025	0.95

Table 5. Efficiency results of ANN models Mohammad haji station in training and validation period

Input model	Nodes in hidden layer	Training		Validation	
		RMSE	R	RMSE	R
i	2	0.016	0.94	0.036	0.95
ii	2	0.013	0.72	0.03	0.92
iii	2	0.016	0.94	0.021	0.95
iv	2	0.018	0.94	0.012	0.87
v	2	0.025	0.89	0.05	0.93

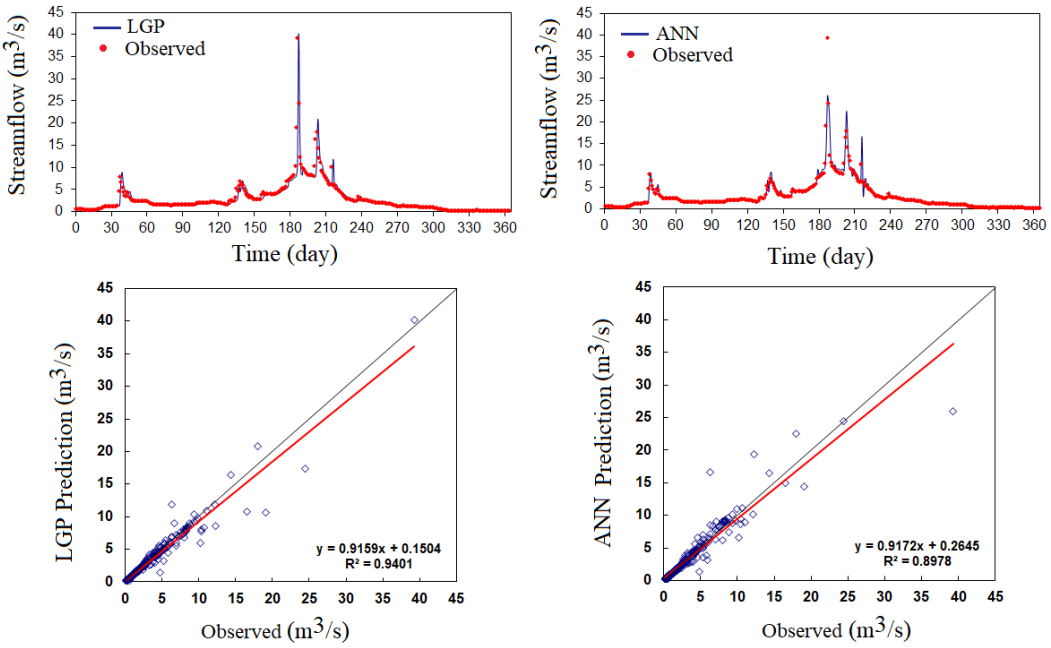

Figure5:
Observed and computed river flow by LGP and ANN model (Mohammad haji station)

5. CONCLUSIONS

This study indicates the ability and enhanced performance of linear genetic programming (LGP) technique to estimate streamflow and suspended sediment load. The LGP model is explicit and simple that can be used by anyone not necessarily being familiar with LGP. The model gives a practical way for use of LGP in environmental studies. With respect to the used statistical measures, the results obtained by LGP were more accurate than those obtained by the ANN that confirm the ability of this approach to be used as useful tool in solving forecasting problems in hydrological works. For the case of sediment load prediction, the best ANN model provides some negative predictions which are not physically acceptable. Thus, some post-processing issues should be considered in order to address this draw back of the ANN model. However, it was not the case for the LGP predictions. The results from second case study clearly showed that LGP was superior to the ANN to capture global minimum and global maximum discharge. Our study only uses data from two rivers and further work using longer and reliable data from various areas to cover temporal and spatial variability may be required to strengthen these conclusions. In the present study, LGP approach was used for the prediction of daily flow and suspended sediment–discharge. Other variants of GP, such as multigene GP or gene expression programing can be a subject of future studies.

REFERENCES

1. Abrahart, R.J., Anctil, F., Coulibaly, P., et al., (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network modelling of surface hydrology. *Progresses in Physical Geography* 36(4), 480-513. doi: 10.1177/0309133312444943

2. Aytek, A., and Kisi, O. (2008) A genetic programming approach to suspended sediment modeling, *Journal of Hydrology*, 351, 288-298. doi: 10.1016/j.jhydrol.2007.12.005

3. Babovic, V., Keijzer, M. (2002) Declarative and preferential bias in GP-based scientific discovery. *Genetic Programming and Evolvable Machines*, *3*(1), 41-79. Retrieved from https://link.springer.com/article/10.1023/A:1014596120381

4. Danandeh Mehr, A., Kahya, E. (2017) A Pareto-optimal moving average multigene genetic programming model for daily streamflow prediction, *Journal of Hydrology*,549, 603-615. doi: 10.1016/j.jhydrol.2017.04.045

5. Danandeh Mehr, A., Nourani, V. (2017) A Pareto-optimal moving average-multigene genetic programming model for rainfall-runoff modelling. *Environmental Modelling & Software*, 92, 239-251. doi: 10.1016/j.envsoft.2017.03.004

6. Danandeh Mehr, A., Demirel, M.C. (2016) On the calibration of multi-gene genetic programming to simulate low flows in the Moselle River. *Uludağ University Journal of the Faculty of Engineering*, 21 (2), 365-376. doi: 10.17482/uumfd.278107

7. Danandeh Mehr, A., Kahya E., Şahin, A. and Nazemosadat M.J. (2015) Successive-station monthly streamflow prediction using different ANN algorithms. *International Journal of Environmental Science and Technology,* 12 (7): 2191-2200. doi: 10.1007/s13762-014-0613-0

8. Danandeh Mehr, A., Kahya, E. and Yerdelen, C. (2014) Linear genetic programming application for successive-station monthly streamflow prediction. *Computers & Geosciences,* 70, 63-72.16(6), 1318-1330. doi: 10.1016/j.cageo.2014.04.015

9. Danandeh Mehr, A., Kahya E. and Olyaie E. (2013) Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *Journal of Hydrology*, 505:240–249. doi: 10.1016/j.jhydrol.2013.10.003

10. Francone, D.F. (2001) *Discipulus™ Software Owner's Manual*, Version 3.0 Register Machine Learning Technologies, Inc., Littleton, Colorado. Retrieved from https://www.cs.bham.ac.uk/~wbl/biblio/gp-html/francone_manual.html

11. Giustolisi, O. (2004) Using genetic programming to determine chezy resistance coefficient in corrugated channels, *Journal of Hydroinform*atics, 157-173. doi: 10.2166/hydro.2004.0013

12. Guven A, Aytek A, Yuce M. I . and Aksoy H. (2008) Genetic programming-based empirical model for daily reference evapotranspiration estimation. *Clean-Soil AirWater*, 36(10-11) 905-912. doi: 10.1002/clen.200800009

13. Guven, A. (2009). Linear genetic programming for time-series modeling of daily flow rate, *Journal of Earth System and Science.* 118, No. 2, 157-173. doi: 10.1007/s12040-009-0022-9

14. Hrnjica, B. and Danandeh Mehr, A. (2019) Optimized Genetic Programming Applications: Emerging Research and Opportunities, (pp. 1-310). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-6005-0

15. Kisi, O. and Cigizoglu H. K. (2007) Comparision of different ANN techniques in river flow prediction, *Civil engineering and environmental system*. vol 24(3), 211-231. doi: 10.1080/10286600600888565

16. Koza, J.R., 1992. Genetic Programming: on the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA.

17. Olyaie, E. Zare Abyaneh, H. and Danandeh Mehr, A. (2017). A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. doi: 10.1016/j.gsf.2016.04.007

18. Ravansalar, M., Rajaee, T., & Kisi, O. (2017). Wavelet-linear genetic programming: A new approach for modeling monthly streamflow. *Journal of Hydrology*, 549, 461-475. doi: 10.1016/j.jhydrol.2017.04.018

19. Roushangar, K., & Homayounfar, F. (2015). Prediction of Flow Friction Coefficient using GEP and ANN Methods. *International Journal of Artificial Intelligence and Mechatronics*, 4(2), 65-68. Retrieved from http://www.ijaim.org/vol-issues.html?view=publication&task=show&id=140

20. Sajikumar, N., & Thandaveswara, B. S. (1999). A non-linear rainfall–runoff model using an artificial neural network. *Journal of hydrology*, 216(1-2), 32-55. doi: 10.1016/S0022-1694(98)00273-X

21. Tofiq F.A., Guven, .A (2014) Prediction of design flood discharge by statistical downscaling and General Circulation Models. *Journal of Hydrology*, 517, 1145-1153. doi: 10.1016/j.jhydrol.2014.06.028

22. Uyumaz, A., Danandeh Mehr A., Kahya E. and Erdem H. (2014) Rectangular side weirs discharge coefficient estimation in circular channels using linear genetic programming approach, *Journal of Hydroinformatics*, 16(6), 1318-1330. doi: 10.2166/hydro.2014.112