

Development of GEP-based functional relationship for sediment transport in tropical rivers

Aminuddin Ab. Ghani · H. Md. Azamathulla

Received: 16 July 2012 / Accepted: 28 September 2012 / Published online: 19 October 2012
© Springer-Verlag London 2012

Abstract This study presents gene expression programming (GEP), which is an extension to genetic programming (GP), as an alternative approach to modeling the functional relationships for the River Kurau, River Langat, and River Muda of the Malaysia. A functional relation has been developed using GEP with non-dimensional variables. The development of a GEP non-dimensional model is described. This paper compares current prediction equation with the existing GEP model for the same rivers (Zakaria et al. in *Sci Total Environ* 408:5078–5085, (2010)). The presented model in this study is a less input GEP model and that predicts good performance. The proposed GEP approach gives satisfactory results compared to existing predictors.

Keywords Malaysian Rivers · Gene expression programming · Sediment transport · Regression analysis

1 Introduction

Information regarding sediment mobility is significant in the aggradation and degradation of rivers. The estimation of river sediment load constitutes an important issue in river engineering. The sediment can increase the elevation of channel beds with excess sand and gravel for tens to hundreds of kilometers downstream. Such aggradation

promotes the lateral migration of channels and may cause serious flooding during rainstorms due to the loss of channel capacity necessary to convey floodwaters [15]. Currently, there are various sediment transport equations that have been developed based on different approaches to predict the total load transport rates. Conventional approaches used in most modeling efforts begin with an assumed form of an empirical or analytical equation and follow with a regression analysis or curve fitting using experimental data to determine the unknown model coefficients Sasal and Isik [18].

Although a number of successful attempts have been recorded by [4–6, 11–14], a wider application of theoretical models is restrained by their heavy demand in terms of computing capacity and time. Alternatively, soft computing techniques, such as artificial neural networks (ANNs), evolutionary computation (EC), fuzzy logic (FL), and genetic programming, have been successfully applied in water engineering problems since last the two decades (Nagy et al. [17], Yang et al. [19]). Bhattacharya et al. [6] used machine learning to model sediment transport.

In recent years, rapid development in Malaysia has led to an increased demand for river sand as a source of construction material, which has resulted in a mushrooming of river sand mining activities that have given rise to various problems that require urgent action by the authorities. These include riverbank erosion, riverbed degradation, river buffer zone encroachment, and deterioration of river water quality. Very often, over-mining occurs, which jeopardizes the health of the river and the environment in general. The present study summarizes the results based on field data collected at three river catchments in Malaysia, that is, the River Muda, the River Langat, and the River Kurau. Fieldwork on selected sites for the three rivers was performed to assess the capacity of the river to convey both

A. Ab. Ghani (✉) · H. Md. Azamathulla
River Engineering and Urban Drainage Research Centre
(REDAC), Engineering Campus, Universiti Sains Malaysia,
Seri Ampangan, 14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: ceamin@eng.usm.my

H. Md. Azamathulla
e-mail: redacazamath@eng.usm.my; mdazmath@gmail.com

water and sediment. Data collection on the bed material was used to characterize the physical characteristics of the sediment responsible for sediment transport, which determines the river response in terms of erosion and deposition. The three rivers clearly have bed material sizes in the sand-gravel range based on the collected data in the present study [7]. This study shows that the sediment mobility can be estimated accurately for Malaysian rivers using the gene expression programming (GEP) technique.

1.1 Overview of GEP

GEP, which is an extension to genetic programming (GP) [16], is a search technique that evolves computer programs (mathematical expressions, decision trees, polynomial constructs, logical expressions, and so on). The computer programs of GEP are all encoded in linear chromosomes, which are then expressed or translated into expression trees (ETs). ETs are sophisticated computer programs that are usually evolved to solve a particular problem and are selected according to their fitness at solving that problem. Thanks to genetic modification, population of ETs will discover traits and therefore will adapt to the particular problem that they are employed to solve. This means that, within enough time and setting the stage correctly, a good solution to the problem will be discovered [8, 9].

GEP is a full-fledged genotype/phenotype system, with the genotype totally separated from the phenotype, while in GP, genotype and phenotype are one entangled mess or more formally, a simple replicator system. As a consequence of this, the full-fledged genotype/phenotype system of GEP surpasses the old GP system by a factor of 100–60,000 [8, 9].

Initially, the chromosomes of each individual of the population are randomly generated. Then, the chromosomes are expressed, and each individual is evaluated based on a fitness function and selected to reproduce with modification, leaving progeny with new traits. The individuals of new generation are, in their turn, subjected to some developmental processes as, expression of the genomes, confrontation of the selection environment, and reproduction with modification. These processes are repeated for predefined number of generations or until a solution is achieved [8, 9] and more details can be found [10].

2 Study area and data used

The present study covers six sites at each of three rivers, that is, River Muda, River Langat, and River Kurau that have different level of sand mining activities. Sungai Muda has a long history of sand mining activity along the upper

reach. Sungai Langat recently has been a major source of sand for construction with the development of Putrajaya. Fewer activities of sand mining are on-going in Sungai Kurau at the upstream of Bukit Merah reservoir.

The surveyed cross sections for the River Muda and the River Langat are single-thread channels with the top width ranging between 22.5 and 134.0 m, representing medium-sized rivers, and the top width for River Kurau ranges between 25.8 and 41.0 m, representing a small-medium river. The slopes are between 0.00008 and 0.0021, indicating that the cross sections are still natural. The details of the morphological and hydrological descriptors and range of field data are given in Table 1. The data collection includes flow velocity (V), flow depth (y_o), flow discharge (Q), suspended load (T_s), bed load (T_b), water surface slope (S_o), and width (B). The corresponding values of flow area (A) and hydraulic radius (R) are also given in Table 1. In addition, the bed elevation, water surface, and thalweg measurement (the minimum bed elevation for a cross section) were also determined at the selected cross sections. The total bed material load (T_j) is composed of the suspended load and bed load. The total bed material load must be specified for sediment transport, scour, and deposition analysis. Details of the measurement methodology are given in [3] and also different factors *affecting sediment transport* such as

$$F_m = \frac{V}{\sqrt{gd_{50}(S_s - 1)}}$$

for mobility, C_v for transport and D_{gr} , d_{50}/D , S_s for sediment, and R/d_{50} , y_o/d_{50} for conveyance shape (Fig. 1).

2.1 Multiple linear regression

Ab. Ghani [1] shows that good prediction of sediment transport in pipes could be obtained from simple regression equations. It is therefore decided to keep the form of the equation as simple and as easy to use as possible.

Based on dimensional analysis from previous studies [2, 17], the proposed function is given as follows:

$$C_v = f\left(\frac{V}{\sqrt{gd_{50}(S_s - 1)}}, \frac{R}{d_{50}}, \frac{B}{y_o}\right) \quad (1)$$

where R = Hydraulic radius and B = water surface width. Utilizing all data from the three rivers in the present study, the best equation is given as follows:

$$C_v = 2.42 \times 10^{-5} \times \left(\frac{V}{\sqrt{2g(d_{50})(S_s - 1)}}\right)^{0.022} \left(\frac{R}{d_{50}}\right)^{-0.2016} \left(\frac{B}{y_o}\right)^{0.104} \quad (2)$$

Table 1 Range of field data for three rivers [2]

Parameters	Study area		
	River Langat	River Muda	River Kurau
Flow discharge, Q (m ³ /s)	2.75–120.76	2.59–343.71	0.63–28.94
Mean flow velocity, V (m/s)	0.23–1.01	0.14–1.45	0.27–1.12
Water surface width, B (m)	16.4–37.6	9.0–90.0	6.30–26.00
Mean flow depth, y_o (m)	0.64–5.77	0.73–6.90	0.36–1.91
Flow area, A (m ²)	8.17–153.57	5.12–278.34	1.43–33.45
Hydraulic radius, R (m)	0.45–3.68	0.55–3.90	0.177–1.349
Water surface slope, S_o	0.00065–0.00185	0.00008–0.000235	0.00050–0.00210
Bed load, T_b (kg/s)	0.027–0.363	0–0.191	0.080–0.488
Suspended load, T_t (kg/s)	0.2860–99.351	0.024–15.614	0.001–2.660
Total bed material load, T_j (kg/s)	0.525–99.398	0.099–15.644	0.089–2.970
Mean sediment size, d_{50} (mm)	0.31–3.00	0.29–2.10	0.41–1.90
Manning n	0.034–0.195	0.021–0.108	0.014–0.066

$$C_v = \left(\frac{Q_b}{Q}\right) \tag{3}$$

and

$$T_j = Q_b * \rho_s \tag{4}$$

Figures 2, 3, 4 show the sediment rating curves for three rivers using Eq. 2 (Ab. Ghani et al. [2]).

3 Development of sediment model using GEP

In this section, the sediment load is modeled using GEP approach. Initially, the “training set” is selected from the whole data and the rest is used as the “testing set”. Once the training set is selected, one could say that the learning environment of the system is defined. The further part of modeling consists of five major steps in preparing to use gene expression programming. The first is to choose the fitness function. For this problem, the fitness, f_i , of an individual program, i , is measured by

$$f_i = \sum_{j=1}^{C_r} (M - |C_{(i,j)} - T_j|) \tag{5}$$

where M is the range of selection, $C_{(i,j)}$ is the value returned by the individual chromosome i for fitness case j (out of C_r fitness cases), and T_j is the target value for fitness case j . If $|C_{(i,j)} - T_j|$ (the precision) is less than or equal to 0.01, then the precision is equal to zero, and $f_i = f_{max} = C_r M$. In this case, $M = 100$ was used; therefore, $f_{max} = 1,000$. The advantage of this kind of fitness functions is that the system can find the optimal solution by itself.

Secondly, the set of terminals T and the set of functions F are chosen to create the chromosomes. In this problem, the terminal set consists obviously of three independent

variables, that is, $T = \left(\frac{V}{\sqrt{gd_{50}(S_s-1)}}, \frac{R}{d_{50}}, \frac{B}{y_o}\right)$. The choice of the appropriate function set is not so obvious; however, a good guess can always be helpful in order to include all the necessary functions. In this study, four basic arithmetic operators (+, −, *, /) and some basic mathematical functions ($\sqrt{\quad}$, e , power) were utilized.

The third major step is to choose the chromosomal architecture, that is, the length of the head and the number of genes. We initially used single gene and 2 length of heads, increased the number of genes and heads, one after another during each run, and monitored the training and testing performance of each model. We observed that number of genes more than 2 and length of heads more than 8 did not significantly increase the training and testing performance of GEP models. Thus, length of the head, $l_h = 8$, and two genes per chromosome were employed for each GEP model in this study. The fourth major step is to choose the linking function. In this study, we tried addition and multiplication as linking functions and observed that linking the sub-ETs by addition gave better fitness (Eq. 5) values. Finally, the fifth major step is to choose the set of genetic operators that cause variation and their rates. A combination of all genetic operators (mutation, transposition, and crossover) was used for this purpose (see Table 2).

The calibration of the GEP model is performed based on 214 input-target pairs of collected data. Among the 214 data sets, 54 (25 %) is reserved for validation, 160 sets for the calibration purpose, and the remaining were used for testing, or validating, the GP model.

The best of generation individual, chromosomes 30, has fitness 687.5 for sediment load T_j . The explicit formulations of GEP for Sediment load T_j , as a function of $\left(\frac{V}{\sqrt{gd_{50}(S_s-1)}}, \frac{R}{d_{50}}, \frac{B}{y_o}\right)$, were obtained for 3 rivers as

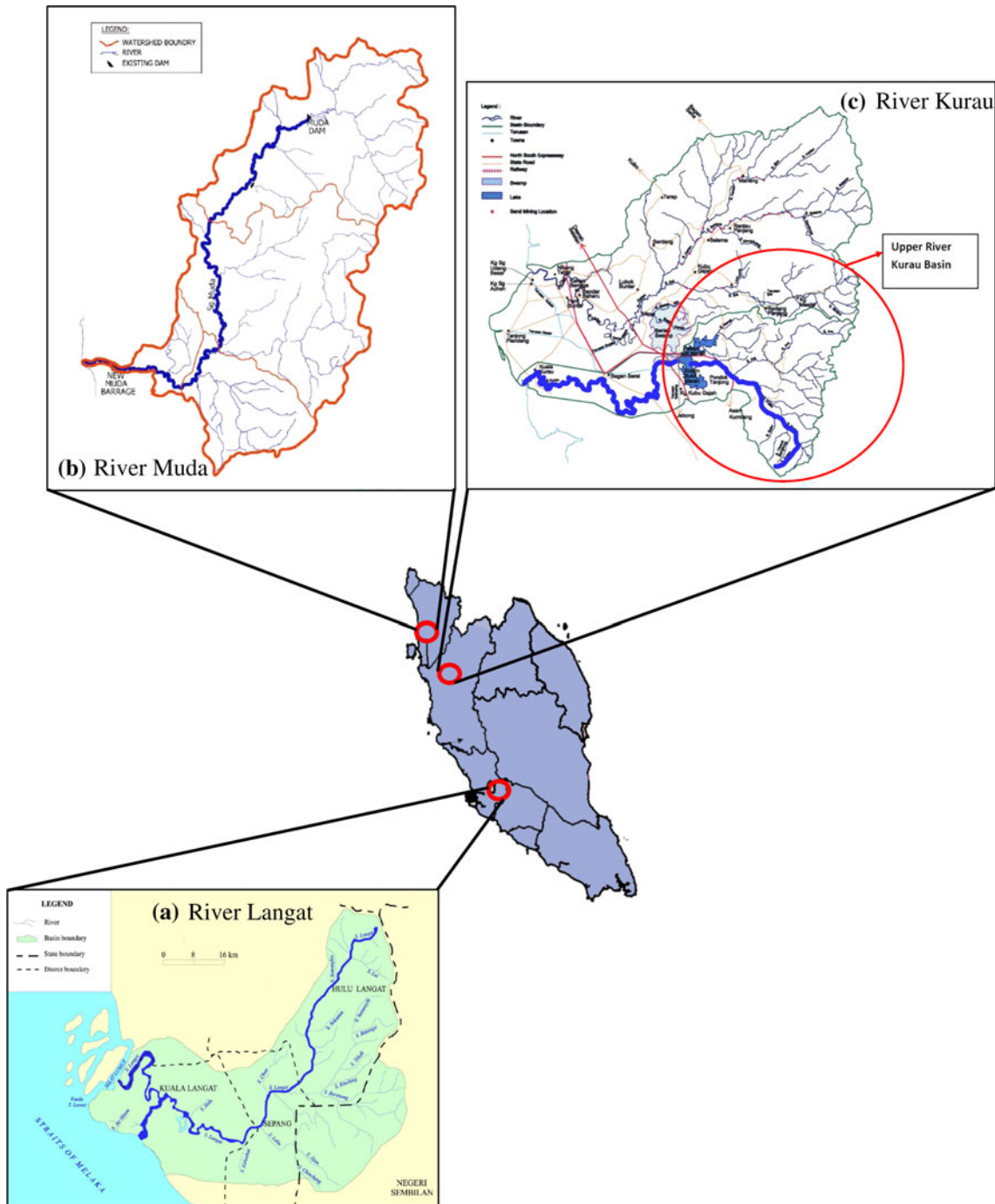


Fig. 1 Study area [2]

$$T_j = e^{\log \left[\left(3.98 + \frac{R}{d_{50}} \right) \sqrt{\frac{2.47}{\gamma_0} \frac{R}{d_{50}}} \right]} + \sqrt{\left(\sqrt{\frac{R}{d_{50}}} \right)} * \left\langle \frac{R}{d_{50}} \right\rangle^2 \left(\frac{1}{-7.05} \right) + \frac{3.05}{V / (g[S_s - 1]d_{50})^{1/2}} \quad (6)$$

Figure 5 show the expression trees of the above formulation.

4 Results and discussion of GEP

The performance of the GEP model was compared with the traditional sediment transport equations. Overall, particularly for field measurements, the GEP models give better predictions than the existing models. The GEP model produced the least errors ($r^2 = 0.97$, MAE = 0.02122 and MSE = 0.0008) for training data and ($r^2 = 0.95$,

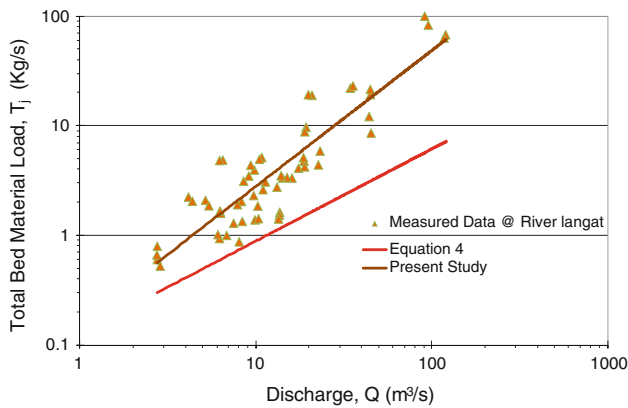


Fig. 2 Comparison of River Langat sediment rating curve for present study and Eq. 4 [2]

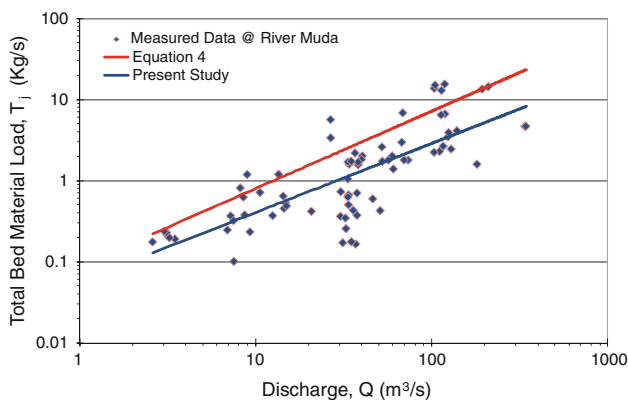


Fig. 3 Comparison of River Muda sediment rating curve for present study and Eq. 4 [2]

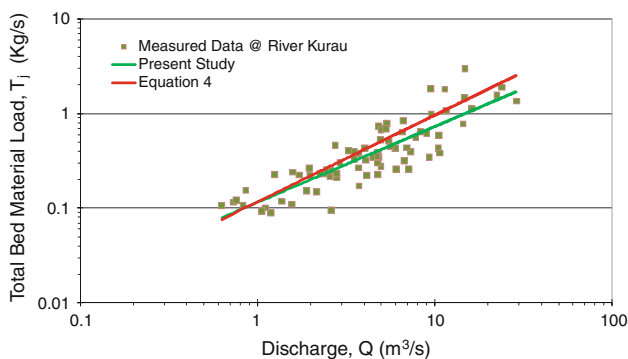


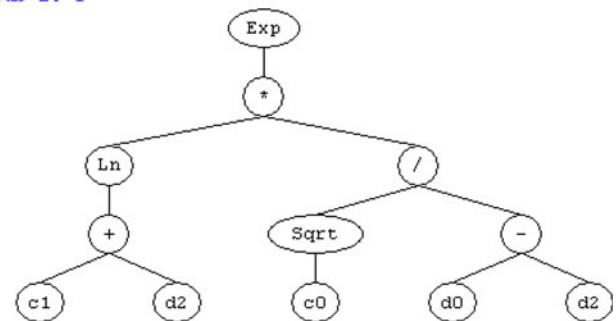
Fig. 4 Comparison of River Kurau sediment rating curve for present study and Eq. 4 [2]

MAE = 0.06122 and MSE = 0.0034) (Fig. 6). The presented model in this study is a less input GEP model and that predicts good performance compared to Zakaria et al.’s [20] GEP model which took longer duration to train GEP model due more inputs. The present GEP model was completed calibration (training) less than 30 min on

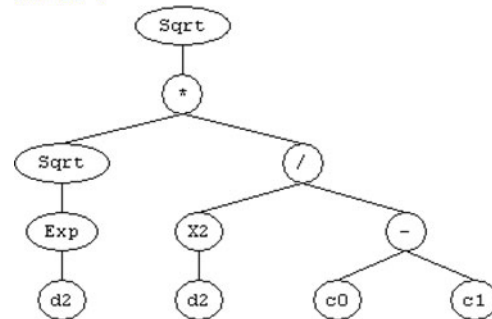
Table 2 Parameters of the optimized GEP model

Parameter	Description of parameter	Setting of parameter
P_1	Function set	+, −, *, /, √, power
P_2	Mutation rate %	30
P_3	Inversion rate %	30
P_4	One point and two point recombination rate respectively %	30, 30
P_5	Gene recombination rate	95
P_6	Gene transportation rate	0.1

Sub-ET 1



Sub-ET 2



Sub-ET 3

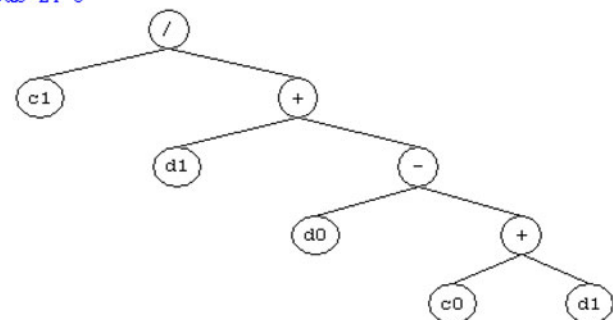


Fig. 5 Expression tree (ET) for the GEP formulation

a standard personal computer (Intel Core i7 with CPU speed of 2.19 GHz and 1.878 GB of RAM running Windows XP).

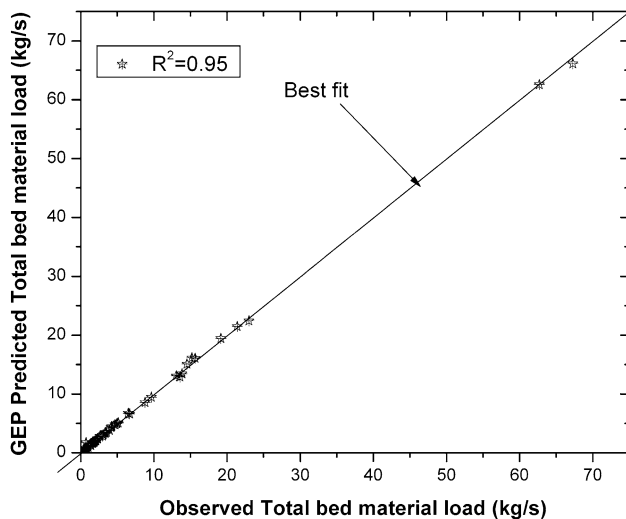


Fig. 6 Observed versus predicted sediment load by GEP for 3 rivers (testing)

The most significant advantage of the proposed GEP model compared to classical regression analysis based models (traditional equations) is that it is capable of mapping the data into a high dimensional feature space, where a variety of methods (described in the previous section) are used to find relations in the data. Since the mapping is quite general, the relations found in this way are accordingly very general.

5 Conclusions

Sediment transport in rivers is a complex phenomenon. The nature and motivation of traditional total load models differ significantly. These approaches are normally able to make predictions within about one order of magnitude of the actual measurements. The data used covers a wide range of the pertinent parameters from the collected actual river data. To overcome the complexity and uncertainty associated with total load estimation, this research demonstrates that GEP model can be applied for accurate prediction. A GEP model with the all the inputs mentioned produced satisfactory perform adequately with less inputs compared to Zakaria et al. [20]. A GEP model that completed with trained values of requiring input of grouped parameters pertaining to mobility, transport, sediment, and conveyance shape is recommended in order to predict sediment load. The GEP model was able to successfully predict total load transport in a great variety of fluvial environments, including both sand and gravel rivers. The high value of the coefficient of determination ($r^2 = 0.95$) implies that the GEP model provides an excellent fit for the measured data. These results suggest that the proposed GEP model is a robust total sediment load predictor.

References

1. Ab. Ghani A (1993) Sediment transport in sewers. Ph.D. thesis, University of Newcastle upon Tyne, UK
2. Ab. Ghani A, Azamathulla HMD, Chang CK, Zakaria NA, Abu Hasan Z (2011) Prediction of total bed material load for rivers in Malaysia: a case study of Langat, Muda and Kurau Rivers. *J Environ Fluid Mech* 11(3):307–318
3. Ab. Ghani A, Chang CK, Abdulla R, Zakaria NA (2003) Guidelines for field data collection and analysis of river sediment, Department of Irrigation and Drainage Malaysia, Kuala Lumpur, 35 pp. ISBN: 983-3067-03-4
4. Azamathulla HMD, Ab. Ghani A, Zakaria NA, Guven A (2010) Genetic programming to predict bridge pier scour. *ASCE J Hydraul Eng* 136(3):165–169
5. Azamathulla HMD, Chang CK, Ab. Ghani AA, Zakaria NA, Ariffin J, Abu Hasan Z (2009) An ANFIS-based approach for predicting the bed load for moderately-sized rivers. *J Hydro-Environ Res* 3(1):35–44
6. Bhattacharya B, Price RK, Solomatine DP (2007) Machine learning approach to modeling sediment transport. *ASCE J Hydraul Eng* 133(4):440–450
7. DID (2009) Department of Irrigation and Drainage Malaysia or DID. Study on river sand mining capacity in Malaysia
8. Ferreira C (2001) Gene expression programming in problem solving, 6th Online World Conference on Soft Computing in Industrial Applications (invited tutorial)
9. Ferreira C (2001) Gene expression programming: a new adaptive algorithm for solving problems. *Complex Syst* 13(2):87–129
10. Guven A, Aytok A (2009) A new approach for stage-discharge relationship: gene-expression programming. *J Hydrol Eng* 14(8): 812–820
11. Guven A, Gunal M (2008) Prediction of scour downstream of grade-control structures using neural networks. *ASCE J Hydraul Eng* 134(11):1656–1660
12. Guven A, Gunal M (2008) Genetic programming approach for prediction of local scour downstream of hydraulic structures. *ASCE J Irrig Drain Eng* 134(2):241–249
13. Guven A (2009) Linear genetic programming for time-series modelling of daily flow rate. *J Earth Syst Sci* 118(2):137–146
14. Guven A, Aytok A, Yuce MI, Aksoy H (2008) Genetic programming-based empirical model for daily reference evapotranspiration model. *Clean Soil Air Water* 36(10–11):905–912
15. Kisi O (2005) Suspended sediment estimation using neuro-fuzzy and neural network approaches. *Hydrol Sci* 50(4):683–696
16. Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. The MIT Press, Cambridge
17. Nagy HM, Watanabe K, Hirano M (2002) Prediction of sediment load concentration in rivers using artificial neural network model. *ASCE, J Hydraul Eng* 128(6):588–595
18. Sasal EMD, Isik S (2005) Suspended sediment load estimation in lower Sakarya River by using soft computational methods. In: Proceeding of the international conference on computational and mathematical methods in science and engineering, CMMSE 2005, Alicante, Spain, pp 395–406
19. Yang et al (2009) Evaluation of total load sediment transport using ANN. *Int J Sediment Res* 24(3):274–286
20. Zakaria NA, Azamathulla HMD, Chang CK, Ab. Ghani A (2010) Gene expression programming for total bed material load estimation—a case study. *Sci Total Environ (STOTEN)* 408:5078–5085