Water Resour Manage (2011) 25:2901–2916
DOI 10.1007/s11269-011-9845-7

Gene-Expression Programming for the Development of a Stage-Discharge Curve of the Pahang River

Hazi Mohammad Azamathulla · Aminuddin Ab. Ghani · Cheng Siang Leow · Chun Kiat Chang · Nor Azazi Zakaria

Received: 7 September 2010 / Accepted: 16 May 2011 /
Published online: 10 June 2011

Abstract This study presents Gene-Expression Programming (GEP), an extension of Genetic Programming (GP), as an alternative approach to modeling the stage-discharge relationship for the Pahang River. The results are compared to those obtained by more conventional methods, i.e., the stage rating curve (SRC) and regression techniques. Additionally, the explicit formulations of the developed GEP models are presented. The performance of the GEP model was found to be substantially superior to both GP and the conventional models.

Keywords Flooding · Pahang River · Stage-discharge · GP · GEP · Regression

1 Introduction

Malaysia is virtually free from natural disasters such as earthquakes, volcanoes, and typhoons. The most common natural disaster encountered in Malaysia is flooding. Two major types of floods occur in Malaysia, i.e., monsoon floods and flash floods. The Malaysian Department of Irrigation and Drainage (DID) has estimated that approximately 29,000 km^2, or 9%, of the total land area and more than 4.82 million

H. M. Azamathulla (✉) · A. Ab. Ghani · C. S. Leow · C. K. Chang · N. A. Zakaria
River Engineering and Urban Drainage Research Centre (REDAC), Universiti Sains Malaysia, Engineering Campus, Seri Ampangan, 14300, Nibong Tebal, Pulau Pinang, Malaysia
e-mail: redacazamath@eng.usm.my, mdazmath@gmail.com

A. Ab. Ghani
e-mail: redac02@eng.usm.my

C. S. Leow
e-mail: redac21@eng.usm.my

C. K. Chang
e-mail: redac10@eng.usm.my

N. A. Zakaria
e-mail: redac01@eng.usm.my

people (22% of the population) are potentially affected by floods annually. The yearly economic damage caused by flooding is estimated at approximately US$300 million (Chan 2005).

Both the stage and the discharge of a river vary depending on the magnitude of rainfall intensity that the river basin receives. To obtain a continuous record of discharge, the stage is recorded and the discharge is computed from a correlation of stage and measured discharge. This correlation, or training (calibration), is known as the stage-discharge relationship (Guven and Aytek 2009). Accurate information about the flow rates of rivers is important for a variety of hydrologic applications such as water and sediment bed material load estimation, water resource planning, operation and development, and hydraulic and hydrologic modeling (Guven and Aytek 2009). However, collecting data for discharge on a continuous basis is costly, especially during large flood events. An alternative approach is to convert records of water stages into discharges using a stage-discharge relationship.

Guven and Aytek (2009) noted that the stage-discharge relationship is not a simple, unique relation. Habib and Meselhe (2006) used artificial neural networks (ANNs) and nonparametric regression to develop stage-discharge relationships for coastal low-gradient streams. Based on the Jones formula, Petersen-Overleir (2006) proposed a methodology utilizing nonlinear regression as a solution for situations in which the stage-discharge relationship is affected by hysteresis due to unsteady flow; Clemmens and Wahlin (2006) evaluated the accuracy of various methods for finding stage-discharge relationships. Baiamonte and Ferro (2007) presented a new flume for measuring flow discharge in sloping channels and deduced theoretical stage-discharge relationships using dimensional analysis and the self-similarity theory; Guven and Aytek (2009) also reported that Liao and Knight (2007) proposed three analytic stage-discharge formulas for prismatic open channels that are suitable for manual calculation.

Guven and Aytek (2009) also summarized several studies found in literature related to the use of Genetic Programming (GP) and Genetic Algorithms (GA) in the area of water engineering (Massoudieh et al. 2008; Velleuxa et al. 2008). Cousin and Savic (1997), Savic et al. (1999), Drecourt (1999), Whigham and Crapper (1999, 2001), and Keijzer and Babovic (2002) applied GP to rainfall-runoff modeling. Babovic et al. (2001) applied GP to sedimentary particle settling velocity equations. Harris et al. (2003) studied velocity predictions in compound channels with vegetated floodplains using GP. Dorado et al. (2003) studied the prediction and modeling of the rainfall-runoff transformation of a typical urban basin using ANNs and GP. Giustolisi (2004) determined the Chezy resistance coefficient in corrugated channels using GP. Rabunal et al. (2007) determined the unit hydrograph of a typical urban basin using GP and ANNs. Most recently, Aytek and Kisi (2008) applied GP for suspended sediment modeling, Guven et al. (2007) applied GP for the estimation of reference evapotranspiration, Guven and Gunal (2008) predicted the depth and location of maximum scour downstream of grade-control structures, and Guven and Aytek (2009) presented stage-discharge relationships for the Schuylkill River at Berne, PA (USA) developed using GEP. Recent works by Azamathulla and Ghani (2011) on the prediction of longitudinal dispersion coefficients in streams using GP, Azamathulla et al. (2010) on bridge pier scour and Zakaria et al. (2010) on sediment transport confirm the suitability of applying GP and GEP for water resource engineering studies.

As mentioned by Guven and Aytek (2009), the body of literature contains many applications of other soft computing techniques in stage-discharge modeling. Jain and Chailsgaonkar (2000) established a stage-discharge relationship based on three-layer feed forward ANNs, Sudheer and Jain (2003) explored the effectiveness of a radial basis function (RBF), and Bhattacharya and Solomatine (2005) observed that ANNs and M5 model trees predicted the stage-discharge relationship much more accurately than the traditional rating curves. Deka and Chandramouli (2003) compared the performance of an ANN model, a modularized ANN model, a conventional curve-fitting approach, and a neuro-fuzzy model for deriving the rating curve using a case study. Overall, they found that the neuro-fuzzy NN model gave the best results. Trancoso et al. (2009) developed an advanced tool for complex river systems, and Lohani et al. (2006) presented an emerging, powerful soft computing technique, the fuzzy logic algorithm, as an alternative method for modeling stage-discharge relationships.

In the present study, we developed mathematical models for the estimation of stage-discharge relationships based on the GP and GEP techniques. To this end, the Pahang River in Malaysia was used as a case study. The performance of the GP model was compared with the stage rating curve, regression techniques, ANN and GEP.

2 Study Area and Data Used

The data set used in this study was obtained from the Malaysian Department of Irrigation and Drainage (DID); see Fig. 1. The time series of daily stage and discharge data was taken from the Temerloh station; the hydrological records were acquired from DID through their National Hydrological Network. The data for the year 2007 were chosen for the training of the proposed GEP models, and the 2004 data were chosen for the testing of the models. The daily discharge ranged from 208.12 to 5,366.55 m^3/s in 2007 and 157 to 4,308 m^3/s in 2004. The 2007 daily discharge data were chosen for training because the worst recent flood occurred in that year and the range of discharges also gave a wider spectrum. Training with a wider data spectrum can ensure a robust model to predict discharges over a wider range and better estimation of maximum flooding in extreme events.

2.1 Stage-Discharge Rating Curve

The functional relationship between stage and discharge can be established by field measurements of stage and discharge and can thereafter be expressed as a rating curve (Guven and Aytek 2009). Ideally, a rating curve describes a unique functional relationship between stage and discharge; therefore, it is obtained as a smooth and continuous curve with a reasonable degree of sensitivity. The relation can be estimated by a sufficient number of measurements suitably distributed throughout the stage range, taking into account the shape of the stage-discharge relation. The number and spacing of observations are selected to conform to the relative frequency of flow at the various stages. In other words, the number of observations in various sub-ranges is in proportion to the probable occurrence of discharge at these ranges, covering the whole range of discharge for which the relation is plotted.

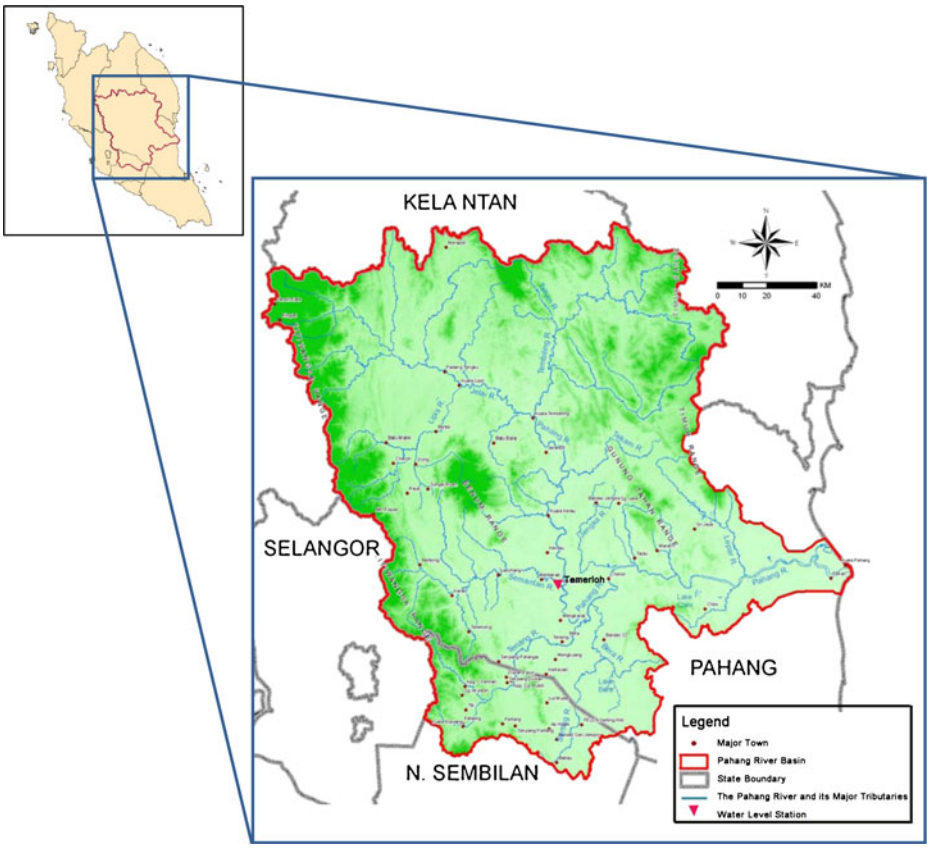

Fig. 1 Study area: the Pahang river

The stage-discharge relation may be expressed by an equation (Subramanya 1995) in the form of a parabolic equation, as in Eq. 1:

$$Q = C\,(S - a)^m \tag{1}$$

where Q is the discharge, S is the gauge height, C and m are constants, and a is the stage at zero flow (datum correction). This equation may be transformed by logarithms into Eq. 2.

$$\log \mathrm{Q} = \log \mathrm{C} + \mathrm{m}.\log\,(\mathrm{S} - \mathrm{a}) \tag{2}$$

This equation resembles the equation of a straight line given in Eq. 3.

$$\mathrm{y} = \mathrm{m}'\mathrm{x} + \mathrm{C}' \tag{3}$$

where m′ is the gradient and C' is the intersection of the line on the y axis. In other words, plotting Q against $(S - a)$ on double-logarithmic graph paper should yield a straight line. Often, two or more straight lines may be required to fit the data, and initially, it is usually possible to decide on the approximate location of the break points of each range by a careful investigation of the controls. The actual break points

may be determined by solving the two equations concerned for Q and S or by purely graphical means.

Sometimes the curve changes from a parabolic to a complex curve or vice versa, and sometimes the constants and exponents vary throughout the range. The logarithmic rating equation, therefore, may not be a single straight line or a gentle curve throughout the entire range of stages at a gauging station. In our case, with practical application in mind, we employed a single rating equation for the Temerloh Station and observed that the resulting equations (given in the next section) captured the whole range of data with reasonable accuracy.

3 Derivation of Discharge Based on Conventional Methods

The conventional models that predict stage-discharge relationships were taken into consideration in the present study, i.e., the stage-rating curve (SRC) and regression (REG). The formulations of SRC (Eq. 4) and REG (Eq. 5) were derived by computing the model constants using the least squares method, for the Temerloh station, respectively (Gross 2003).

$$\mathrm{Q} = 1.22(\mathrm{S} - 19.88)^{3.54} \tag{4}$$

$$\mathrm{Q} = 7.217\mathrm{S}^{1.285} \tag{5}$$

However, the rounded values of RMSE were 880 m^3/s (SRC) and 1580 m^3/s (REG), with a low R^2 value of 0.68 for the SRC. The REG for this station produced poor results ($R^2 = 0.345$) and is therefore not recommended to represent the stage-discharge relationship of the river. Thus, we next applied ANN, GP and GEP as alternate methods to develop this relationship.

4 Artificial Neural Network Model

Artificial neural networks provide a random mapping between an input and an output vector, typically consisting of three layers of neurons, namely, input, hidden and output, with each neuron acting as an independent computational element. The strength of neural networks is derived from the high degree of freedom associated with their architecture. Prior to application, the network is trained using observed data sets. This feeds the network with input and output pairs and determines the values of connection weights, bias or centers. The training may require the completion of many epochs (i.e., single cycles of the presentation of complete data sets to the network) until the training sum of squares error reaches a specified error goal. The concepts involved behind these training schemes are outlined in the American Society of Civil Engineers (ASCE) Task Committee (2000). The neural network toolbox contained in the MATLAB package was used in this study. The usual feed forward-type network was trained using radial basis functions (RBF). The details of the ANN model are given in Appendix.

5 Genetic Programming

Genetic programming (GP), a branch of genetic algorithms (GA) (Holland 1975), is a method for determining the most "fit" computer program by artificial evolution (Johari et al. 2006). GP initializes a population consisting of random members known as chromosomes (individual), and the fitness of each chromosome is evaluated with respect to a target value. The principle of Darwinian natural selection is used to select and reproduce "fitter" programs. GP creates computer programs of equal or unequal lengths that consist of variables (terminal) and several mathematical operators (function) sets as the solution. The function set of the system can be composed of arithmetic operations (+, −, ×, and ÷) and function calls (such as $\{e^x, x, \sin, \cos, \tan, \log, \text{sqrt}, \ln, power\}$). Each function implicitly includes an assignment to a variable, which facilitates the use of multiple program outputs in GP, whereas in a tree-based GP, these side effects must be incorporated explicitly (Brameier and Banzhaf 2001).

The present GP utilizes a two-point string crossover. Segments of random position and random length are selected in each parent and exchanged between them. If one of the resulting children exceeds the maximum length, crossover is abandoned and restarted by exchanging equalized segments (Brameier and Banzhaf 2001). An operand or an operator of an instruction is changed by mutation into another symbol over the same set. The fitness of a GP individual may be computed using Eq. 6:

$$f = \sum_{j=1}^{N} \left(\left| X_j - Y_j \right| \right), \tag{6}$$

where X_j is the value returned by a chromosome for the fitness case j, and Y_j is the expected value for the fitness case j.

In GP, the maximum size of the program is usually restricted to avoid programs growing without bounds (Brameier and Banzhaf 2001). This configuration was tested for the proposed GP stage-discharge model and was found to be sufficient. The best individual (program) of a trained GP can be converted into a functional representation by successive replacements of variables, starting with the last effective instruction (Oltean and Groşan 2003).

To the best of our knowledge, the application of GP to develop a stage-discharge curve for the Pahang River is novel. Davidson et al. (1999) determined the empirical relationships for friction in turbulent pipe flow and Babovic and Keijzer (2000a, b) found the additional resistance to flow induced by flexible vegetation. Keijzer and Babovic (2002) derived empirical equations using real-world hydraulic data, Giustolisi (2004) determined the Chezy resistance coefficient in corrugated channels, Kizhisseri et al. (2005) explored a better correlation between the temporal pattern of a flow field and sediment transport by utilizing numerical model results and field data, and Guven and Gunal (2008) predicted local scour downstream of grade-control structures. In more recent applications, Guven and Kisi (2011) managed to successfully apply GP for the estimation of suspended sediment yields in a natural river using stream flow and sediment data for the Tongue River in Montana (USA).

5.1 GP Modeling of Stage-Discharge Rating Curve

The same scenario was used to develop the GP model: the 2007 data were chosen for model training, and the 2004 data were chosen for testing. During model

Table 1 Parameters of the optimized GP model

Parameter	Description of parameter	Setting of parameter
P_1	Function set	$+, -, \times, \div, \surd$, power
P_2	Population size	250
P_3	Mutation frequency (%)	96
P_4	Crossover frequency (%)	50
P_5	Number of replication	10
P_6	Block mutation rate (%)	30
P_7	Instruction mutation rate (%)	30
P_8	Instruction data mutation rate (%)	40
P_9	Homologous crossover (%)	95
p_{10}	Program size	Initial 64, maximum 256

development in this study, the discharge (cubic meters per second) was selected as output and the stage as the input. Four basic arithmetic operators ($+, -, \times, \div$) and several basic mathematical functions ($\surd$, x^2, power) were utilized. A large number of generations (5,000) were tested. First, the maximum size of each program was specified as 256, and the program was initiated with 64 instructions. The functional set and operational parameters used in GP modeling during this study are listed in Table 1.

The simplified analytic form of the proposed GP model is expressed as Eq. 7 below:

$$Q = -0.9347S^3 + 118.6S^2 - 3836S + 36855 \tag{7}$$

5.2 Training and Testing Results of GP Modeling

The performance of the GP model in the training and testing sets was validated in terms of the common statistical measures R^2 (coefficient of determination) and root-mean-square error (RMSE).The GP approach resulted in a highly nonlinear relationship between stage and discharge, with high accuracy and relatively low error. The testing performance of the proposed GP model revealed a high generalization capacity, with $R^2 = 0.99$ and RMSE $= 6.31$ m^3/s in training and $R^2 = 0.9512$, RMSE $= 173.5$ m^3/s in testing. Figures 2 and 3, respectively, show the performance of the GP training and testing runs.

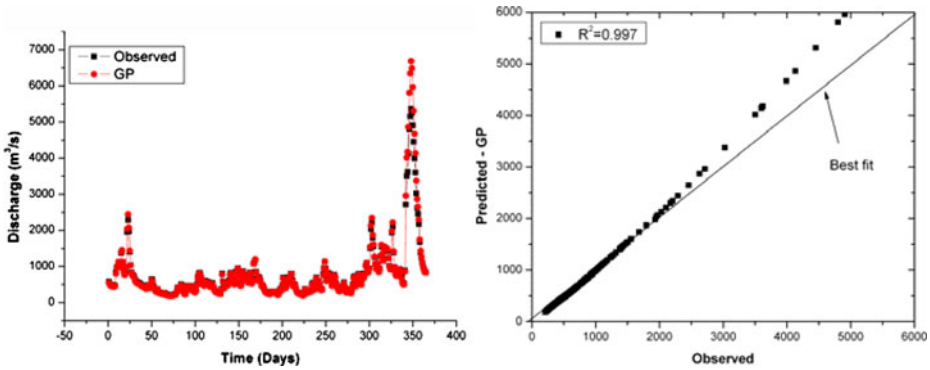

Fig. 2 Comparison between Observed and GP-based Discharge (Training) (2007 Flood)

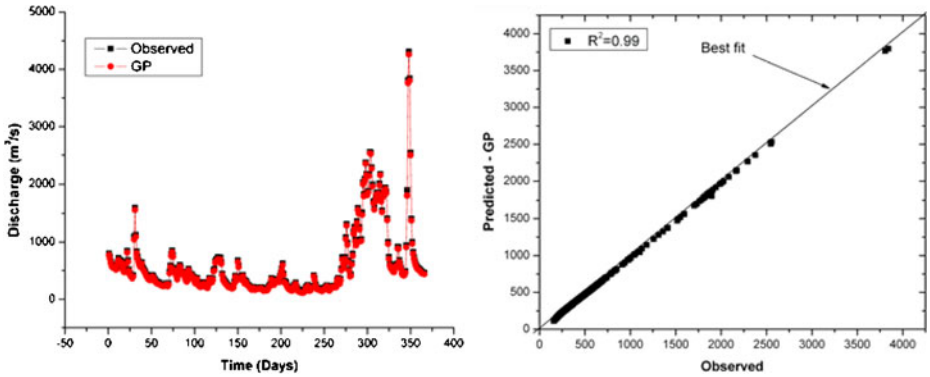

Fig. 3 Comparison between Observed and GP-based Discharge (Testing) (2004 Flood)

6 Development of the GEP Model

6.1 Overview of Gene-Expression Programming

Gene-Expression Programming (GEP) is a new evolutionary Artificial Intelligence technique developed by Ferreira (2001a). This technique is an extension of GP, developed by Koza (1992). The genome is encoded as linear chromosomes of fixed length, as in Genetic Algorithm (GA); however, in GEP the genes are then expressed as a phenotype in the form of expression trees. GEP combines the advantages of both its predecessors, GA and GP, and removes their limitations. GEP is a full-fledged genotype/phenotype system in which both are dealt with separately, whereas GP is a simple replicator system. As a consequence of this difference, the complete genotype/phenotype GEP system surpasses the older GP system by a factor of 100–60,000 (Ferreira 2001a, b).

In GEP, just like in other evolutionary methods, the process starts with the random generation of an initial population consisting of individual chromosomes of fixed length. The chromosomes may contain one or more than one genes. Each individual chromosome in the initial population is then expressed, and its fitness is evaluated using one of the fitness function equations available in the literature. These chromosomes are then selected based on their fitness values using a roulette wheel selection process. Fitter chromosomes have greater chances of selection for passage to the next generation. After selection, these are reproduced with some modifications performed by the genetic operators. In Gene Expression Programming, genetic operators such as mutation, inversion, transposition and recombination are used for these modifications. Mutation is the most efficient genetic operator, and it is sometime used as the only means of modification. The new individuals are then subjected to the same process of modification, and the process continues until the maximum number of generations is reached or the required accuracy is achieved (Ferreira 2001a, b).

Because a random numerical constant (RNC) is a crucial part of any mathematical model, it must be taken into account; however, Gene Expression Programming has the ability to handle RNCs efficiently. In GEP, an extra terminal '?' and an extra domain D_c after tail of the each gene is introduced to handle RNCs. In the GEP

system, there are several genetic operators used for the genetic modification of chromosomes (Ferreira 2006):

Mutation This is the most important and influential of all the operators. In GEP modeling, mutation can take place at any position on a genome. However, the structural organization of chromosomes must remain the same; i.e., in the heads, any function can be replaced by a terminal function, but in the tails, terminals can only be changed into other terminals, as there is no function in the tail. In this way, all the new individuals produced by mutation are structurally correct programs.

Inversion In this operator, a sequence within the head of gene is selected and inverted. The chromosome is randomly chosen, as are the gene to be modified and the initial and terminal points of the portion of head to be inverted.

Insertion Sequence Transposition Insertion sequence (IS) elements are short genomic segments having a function or terminal at the first position. This operator randomly chooses a chromosome, a gene to be modified and the start and end of the IS element, which is then transposed to the start of the gene just after the root.

Root Insertion Sequence Transposition These are short genomic fragments similar to IS elements; the only difference is that here, the starting point is always a function. The chromosome, the gene to be modified and the starting and ending points of the root insertion sequence (RIS) element are randomly selected, and it is transposed to the starting point of the gene.

Gene transposition: In gene transposition, an entire gene acts as a transposon and transposes itself to the beginning of the chromosome. In contrast to the other forms of transposition, in gene transposition, the transposon (the gene) is deleted at the place of origin.

Single or double crossover/recombination: In single crossover, the parent chromosomes are paired at the same selected point. The portion of the gene downstream of the crossover point is subsequently exchanged between the two chromosomes. In double crossover, two parent chromosomes are paired and two points are randomly chosen as crossover points. The material between the crossover points is then exchanged between the parent chromosomes, forming two new daughter chromosomes (Guven and Aytek 2009).

Gene crossover In gene crossover, entire genes are exchanged between two parent chromosomes, forming two daughter chromosomes containing genes from both parents. The exchanged genes are randomly chosen and occupy exactly the same positions as in the parent chromosomes.

Because Gene Expression Programming combines the advantages of GA and GP, it has proven to be an efficient modeling tool for solving complex real world problem. Recent applications of GEP in various fields include studies by Ghani and Azamathulla (2010) on sediment transport in a sewer pipe system, Fernando et al. (2009) on rainfall-runoff model development, Guven and Aytek (2009) the stage-discharge relationships, Eldrandaly and Negm (2008) on hydraulic data prediction, Bărbulescu and Băutu (2009) on time series modeling, Gempeler (2004) on image compression and Dehuri and Cho (2008) on multi-objective classification rule mining.

The functionality of each genetic operator included in the GEP system has been described by Guven and Aytek (2009); this report also provides an application for improving stage-discharge relationships. The functional set and operational parameters used for GEP modeling in this study are listed in Table 2.

6.2 Derivation of Discharge Based on GEP

The discharge (Q) is modeled in terms of the stage (S) using a GEP approach. Initially, the "training set" is selected from the entire data set, 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 modeling also includes five major steps to prepare the GEP for use. The first is to choose the fitness function. For this problem, the fitness, f_i, of an individual program, i, is measured by:

$$f_i = \sum_{j=1}^{C_t} \left(M - \left| C_{(i,j)} - T_j \right| \right) \tag{8}$$

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_t fitness cases) and T_j is the target value for fitness case j. If $|C_{(i,j)} - T_j|$ (the precision) $\leqq 0.01$, then the precision is 0, and $f_i = f_{max} = C_t M$. In this case, $M = 100$ is used; therefore, $f_{max} = 1{,}000$. The advantage of this kind of fitness function 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 of a single independent variable, i.e., $T = \{h\}$. The choice of the appropriate function set is not obvious; however, a good guess is helpful if it includes all the necessary functions. In this study, four basic arithmetic operators ($+, -, \times, \div$) and some basic mathematical functions ($\sqrt{}$) were utilized.

The third major step is to choose the chromosomal architecture, i.e., the length of the head and the number of genes. We initially used a single gene and two head lengths and increased the number of genes and heads one at a time during each run while we monitored the training and testing performances of each model. We observed that more than two genes and a head length greater than 8 did not significantly improve the training and testing performance of GEP models. Thus, the head length, $l_h = 8$, 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, addition and multiplication operators were used as linking functions, and we observed that linking

Table 2 Genetic operators used in GEP modeling

Parameter	Definition	Value
P_1	Mutation rate	0.044
P_2	Inversion rate	0.1
P_3	One-point recombination rate	30%
P_4	Two-point recombination rate	30%
P_5	Gene recombination rate	0.1
P_6	Gene transposition rate	0.1

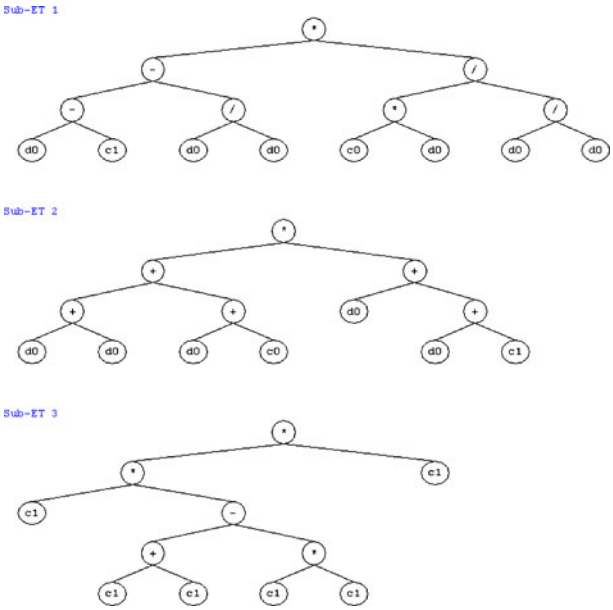

Fig. 4 Expression Tree (ET) for the GEP formulation for the Temerloh Station, where $d_o = S$

the sub-expression trees (ETs) by addition yielded the best fitness (Eq. 8) values. The fifth and final 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) is used for this purpose (Table 2).

The best individual of the first generation, chromosome 39, had fitness 528.75 for the Temerloh station. The explicit formulations of the GEP for discharge (Q) as a function of stage (S) were obtained for the Temerloh station as:

$$Q = \left(\left((S - 3.519623) - \left(\frac{S}{S}\right)\right) * \left(\frac{3.844 * S}{\frac{S}{S}}\right)\right) + \left[((S + S) + (S - 9.249)) * (S + (S - 9.546))\right] + \left[((-7.556 - 7.556) - (-7.556 * -7.556)) * -7.556 * -7.556\right] \quad (9)$$

The simplified form of Eq. 9 is as follows:

$$Q = 9.84 * S^2 - 64.391 * S - 4033.296 \quad (10)$$

Figure 4 shows the expression trees for the above formulation. It should be noted that the proposed GEP formulations in Eq. 9 are valid for parameters ranging between the minimum and maximum values given in Table 2.

7 GEP Training and Testing Results

Different input sets of daily stage-discharge records were developed to calibrate the GEP models. The GEP estimates were then compared to the observed data via scatter plots for the training (Fig. 5) and testing sets (Fig. 6). As shown in Figs. 5 and 6,

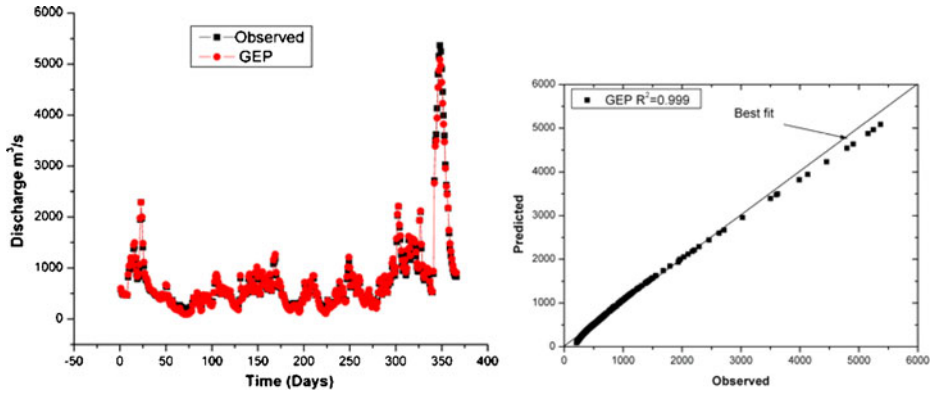

Fig. 5 The GEP-based training period for the Temerloh Station (2007 Flood)

it was evident that the proposed GEP model determined the nonlinear relationship between the input and the output variables impressively well, with an R^2 of 0.993 and an RMSE of 62.34 m^3/s. Comparing the GEP predictions with the observed data for the test stage demonstrated a high generalization capacity with $R^2 = 0.945$ and RMSE = 78.98 m^3/s for the 2004 data. Together, these findings demonstrate the acceptable performance of the GEP models for estimating Q in both the training and testing stages.

8 Results and Discussion

The performance of the tested methods was analyzed by computing the R^2 and RMSE values for the daily observed flows using the GP, GEP and other methods, as summarized in Table 3. The RMSE and R^2 values were calculated to measure the

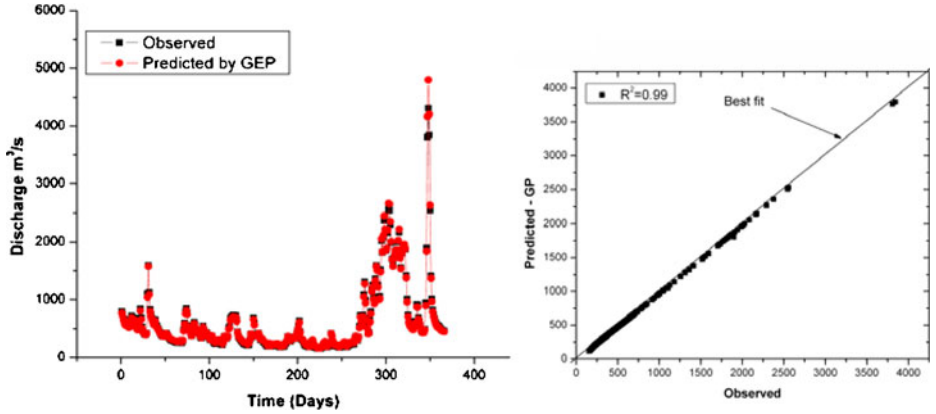

Fig. 6 The GEP-based testing period for the Temerloh Station (2004 Flood)

Table 3 Performance of various stage-discharge modeling techniques

Modeling technique	R^2	RMSE (m^3/s)
SRC	0.68	880
REG	0.345	1589
ANN	0.74	550
GP	0.951	173.5
GEP	0.945	78.98

deviation from and approximation of observed flows obtained from the Temerloh station. Here, a low RMSE value implies a good performance of the applied method. It was obvious that two models, i.e., GP and GEP, outperformed the rest in terms of the goodness-of-fit indicators. Both GP and GEP produced R^2 values very close to 1, suggesting very little discrepancy between observed and predicted discharges. GEP was particularly impressive, as the RMSE remained at a very low level (78.98 m^3/s) compared to the other models.

Further visual verification of the predicted results of the GP and GEP models verified the superiority of the GEP model. Although they shared similarly good R^2 values, Fig. 7 clearly shows that beyond the 28 m elevation mark, the GP model tended to overestimate the corresponding discharge, and this prediction error grew exponentially to the increase in stages thereafter. In contrast, the GEP showed more consistent and accurate predictions up to an elevation of 32 m, after which the model tended to slightly underestimate the discharge for a given stage. However, this prediction error was still relatively much smaller than that produced by the GP model. We therefore concluded that although GP and GEP produced almost equally excellent predictions at lower flows (or stages), GEP was more reliable in predicting the higher discharges that usually occurred during extreme flood events.

In this particular study, the search for a stage-discharge estimation model led to the testing of the SRC, REG, ANN, GP and GEP models. As ANN does not produces an analytic mathematical equation, its application is very much limited to those that possess the knowledge to utilize this modeling technique. As discussed earlier, GEP performed better than GP during higher stage and discharge levels, making it the more preferable model. Figure 8 shows plots of the observed and predicted discharges (from the GEP and SRC models) for the Pahang River in 2007. It is obvious that GEP performed much better than the conventional stage-discharge model, i.e., SRC. This is especially true during the prediction of maximum flooding, which is very critical in all flood mitigation and management activities.

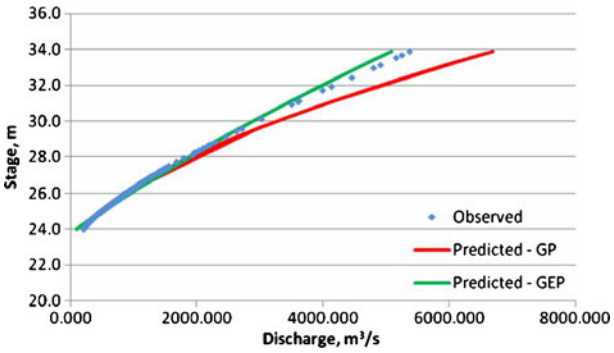

Fig. 7 Comparison of different predictions of stage-discharge rating curves for the Pahang River during the 2007 flooding period

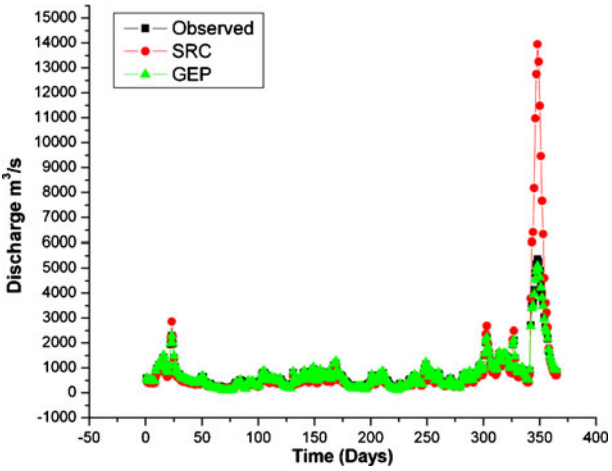

Fig. 8 Observed versus GEP and SRC scatter plots in the training period for the Temerloh Station for the year 2007

9 Conclusions

The main aim of this study was to evaluate Genetic Programming (GP) and Gene Expression Programming (GEP) as alternative tools for modeling the stage-discharge relationship for the Pahang River. Stage and discharge data from 2 years, 2004 and 2007, were used to compare the performance of the GP and GEP models against that of the more conventional SRC and REG approaches. The GEP model was found to be considerably better than the conventional SRC, REG and GP models. GEP was also relatively more successful than GP, especially in estimating large discharge values during flood events. The results of this study are highly promising and suggest that GEP modeling is a more versatile technique than GP and represents an improved alternative to the more conventional approaches for the determination of stage-discharge relationships. 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.

Acknowledgements The analysis was conducted at the River Engineering and Urban Drainage Research Center (REDAC) at the Universiti Sains Malaysia in Nibong Tebal, Malaysia. Support from the Malaysian Department of Irrigation and Drainage is gratefully acknowledged. The authors also wish to express their sincere gratitude to Universiti Sains Malaysia for funding a research university grant to conduct this ongoing research (PRE1001/PREDAC/814056) led by the second author.

Appendix

The data for the year 2007 were chosen for training the proposed GEP models, and the 2004 data were chosen for testing. As dictated by the use of the Gaussian function, all patterns were normalized within the range of (0.0, 1.0) before use. The RBF network (1 input, 20 hidden neurons and 1 output) was trained using various values of spread (α) between 0 and 1. The value of 0.01 was selected because it yielded the best performance for the training data. The ANN model predicted with

a high error (RMSE = 550 m^3/s and $R^2 = 0.74$) and hence, was not recommended in this study.

References

Aytek A, Kisi O (2008) A genetic programming approach to suspended sediment modeling. J Hydrol 351:288–298

Azamathulla HMd, Ghani AAb (2011) Genetic programming for longitudinal dispersion coefficients in streams. Water Resour Manag 25(6):1537–1544

Azamathulla HMd, Ghani AA, Zakaria NA, Guven A (2010) Genetic programming to predict bridge pier scour. ASCE J Hydraul Eng 136(3):165–169

Babovic V, Keijzer M (2000a) Rainfall-runoff modeling based on genetic programming. Nord Hydrol 33(5):331–346

Babovic V, Keijzer M (2000b) Genetic programming as a model induction engine. J Hydroinform 2(1):35–60

Babovic V, Keijzer M, Aguilera DR, Harrington J (2001) Automatic discovery of settling velocity equations. D2K Technical Report0201-1; Danish Technical Research Council (STVF). http://www.d2k.dk

Baiamonte G, Ferro V (2007) Simple flume for flow measurement in sloping channel. J Irrig Drain Eng 133(1):71–78

Bărbulescu A, Băutu E (2009) ARIMA models versus gene expression programming in precipitation modeling. Recent Advances in Evolutionary Computing, WSEAS Press, pp 112–117 (ISBN 978-960-474-067-3, ISSN 1790-5109)

Bhattacharya B, Solomatine DP (2005) Neural networks and M5 model trees in modeling water level-discharge relationship. Neurocomputing 63:381–396

Brameier M, Banzhaf W (2001) A comparison of linear genetic programming and neural networks in medical data mining. IEEE Trans Evol Comput 5:17–26

Chan NW (2005) Sustainable management of rivers in Malaysia: involving all stakeholders. Int J River Basin Manag 3(3):147–162

Clemmens AJ, Wahlin BT (2006) Accuracy of annual volume from current-meter-based discharges. J Hydrol Eng 11(5):489–501

Cousin N, Savic DA (1997) A rainfall-runoff model using genetic programming. Centre for Systems and Control Engineering. Report No. 97/03, School of Engineering, University of Exeter, Exeter, United Kingdom, p 70

Davidson JW, Savic DA, Walters GA (1999) Method for identification of explicit polynomial formulae for the friction in turbulent pipe flow. J Hydroinform 1(2):115–126

Dehuri S, Cho SB (2008) Multi-objective classification rule mining using gene expression programming. Third 2008 International Conference on Convergence and Hybrid Information Technology, 978-0-7695-3407-7/08 $25.00 © 2008 IEEE. doi:10.1109/ICCIT.2008.27

Deka P, Chandramouli V (2003) A fuzzy neural network model for deriving the river stage-discharge relationship. Hydrol Sci 48(2):197–209

Dorado J, Rabunal JR, Pazos A, Rivero D, Santos A, Puertas J (2003) Prediction and modeling of the rainfall-runoff transformation of a typical urban basin using ANN and GP. Appl Artif Intell 17:329–343

Drecourt JP (1999) Application of neural networks and genetic programming to rainfall-runoff modeling. D2K Technical Report 0699-1-1, Danish Hydraulic Institute, Denmark

Eldrandaly K, Negm AA (2008) Performance evaluation of gene expression programming for hydraulic data mining. Int Arab J Inf Technol 5(2):126–131

Fernando DAK, Shamseldin AY, Abrahart RJ (2009) Using gene expression programming to develop a combined runoff estimate model from conventional rainfall-runoff model outputs. 18th World IMACS / MODSIM Congress, Cairns, Australia 13–17 July 2009

Ferreira C (2001a) Gene expression programming in problem solving, 6th Online World Conference on Soft Computing in Industrial Applications (invited tutorial)

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

Ferreira C (2006) Gene-expression programming; mathematical modeling by an artificial intelligence. Springer, Heidelberg

Ghani AA, Azamathulla HMd (2010) Gene-expression programming for sediment transport in sewer pipe systems. J Pipeline Syst Eng Pract, ASCE 2(3). doi:10.1061/(ASCE)PS.1949-1204.0000076
Gempeler M (2004) Image compression using gene expression programming. (http://digital.cs.usu.edu/~xqi/Teaching/REU06/Website/Rob/RobertFinalPaper.pdf)
Giustolisi O (2004) Using genetic programming to determine Chezy resistance coefficient in corrugated channels. J Hydroinformatics 6(3):157–173
Gross J (2003) Linear regression, 3rd ed. Springer
Guven A, Aytek A (2009) A new approach for stage-discharge relationship: gene-expression programming. J Hydrol Eng 14(8):812–820
Guven A, Gunal M (2008) A genetic programming approach for prediction of local scour downstream hydraulic structures. J Irrig and Drain Eng 132(4):241–249
Guven A, Kisi O (2011) Estimation of suspended sediment yield in natural rivers using machine-coded linear genetic programming. Water Res Manag 25(2):691–704
Guven A, Aytek A, Yuce MI, Aksoy H (2007) Genetic programming-based empirical model for daily reference evapotranspiration estimation. CLEAN-Soil, Air, Water Journal 36:10–11
Habib EH, Meselhe EA (2006) Stage-discharge relations for low-gradient tidal streams using data driven models. J Hydraul Eng 132(5):482–492
Harris EL, Babovic V, Falconer RA (2003) Velocity predictions in compound channels with vegetated floodplains using genetic programming. Intl J River Basin Manage 1(2):117–123
Holland JH (1975) Adaptation in natural and artificial system. University of Michigan Press, Ann Arbor
Jain SK, Chailsgaonkar D (2000) Setting up stage-discharge relations using ANN. J Hydrol Eng 5(4):428–433
Johari A, Habibagahi G, Ghahramani A (2006) Prediction of soil-water characteristic curve using genetic programming. ASCE J Geotech Geoenviron Eng 132(5):661–665
Keijzer M, Babovic V (2002) Declarative and preferential bias in GP-based scientific discovery. Genet Program Evol M 3(1):41–79
Kizhisseri AS, Simmonds D, Rafiq Y, Borthwick M (2005) An evolutionary computation approach to sediment transport modeling. Proc., 5th Int. Conf. on Coastal Dynamics, ASCE, Barcelona, Spain
Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. MIT, Cambridge
Liao H, Knight DW (2007) Analytic stage-discharge formulas for flow in straight prismatic channels. J Hydraul Engrg 133(10):1111–1122
Lohani AK, Goel NK, Bhatia KKS (2006) Takagi-Sugeno fuzzy inference system for modeling stage-discharge relationship. J Hydrology 331:146–160
Massoudieh A, Abrishamchi A, Kayhanian M (2008) Mathematical modeling of first flush in highway storm runoff using genetic algorithm. Sci Total Environ 398:107–121
Oltean M, Groşan C (2003) A comparison of several linear genetic programming techniques. Complex Syst 14(1):1–29
Petersen-Overleir A (2006) Modelling stage-discharge relationships affected by hysteresis using the Jones formula and nonlinear regression. J Hydrological Engrg 51(3):365–388
Rabunal JR, Puertas J, Suarez J, Rivero D (2007) Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. Hydrol Process 27(4):476–485
Savic AD, Walters AG, Davidson JW (1999) A genetic programming approach to rainfall-runoff modeling. Water Resour Manag 13:219–231
Subramanya K (1995) Engineering hydrology. McGraw-Hill, Delhi, India
Sudheer KP, Jain SK (2003) Radial basis function neural network for modeling rating curves. J Hydrol Eng 8(3):161–164
Trancoso AR, Braunschweig F, Leitão PC, Obermann M, Neves R (2009) An advanced modelling tool for simulating complex river systems. Sci Total Environ 407:3004–3016
Velleuxa ML, England JF, Julien PY (2008) TREX: spatially distributed model to assess watershed contaminant transport and fate. Sci Total Environ 404:111–128
Whigham PA, Crapper PF (1999) Time series modeling using genetic programming: an application to rainfall-runoff models. In: Spector L, Langdon WB, O'Reilly U, Angeline PJ (eds) Advances in genetic programming. MIT, Cambridge, pp 89–104
Whigham PA, Crapper PF (2001) Modeling rainfall-runoff using genetic programming. Math Comput Model 33(6–7):707–721
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 408(21):5078–5085