Rectangular side weirs discharge coefficient estimation in circular channels using linear genetic programming approach

Ali Uyumaz, Ali Danandeh Mehr, Ercan Kahya and Hilal Erdem

ABSTRACT

Side weirs are diversion structures extensively used in irrigation, flood protection and combined sewer systems. Accurate estimation of the discharge coefficient (C_d) of side weirs is essential to compute the water surface profile over the weirs and to determine the lateral outflow rate from the system. In this paper, we have utilized a linear genetic programming (LGP) technique to develop new empirical formulas for the estimation of C_d of sharp-edged rectangular side weirs located in circular channels. For this aim, we have employed a total of 1,686 laboratory experimental observations in both sub- and supercritical flow regimes in order to train and validate the proposed models. The performance of the LGP-based models was also compared with those of different multilinear and nonlinear regression models in terms of root mean squared errors, mean absolute errors, and determination coefficient. The results indicated that an explicit LGP-based model using only mathematical functions could be employed successfully in C_d estimation in both sub- and supercritical flow conditions. Genetic-based sensitivity analysis among the input parameters demonstrated that Froude number at upstream of the weir has the most impact on the C_d estimation.

Key words | circular channels, discharge measurement, linear genetic programming, side weirs

Ali Uyumaz
Ali Danandeh Mehr (corresponding author)
Ercan Kahya
Hilal Erdem
Civil Engineering Department,
Istanbul Technical University,
Hydraulics Division, Maslak,
34469 Istanbul,
Turkey
E-mail: *danandeh@itu.edu.tr*

INTRODUCTION

Side weirs, also known as lateral weirs, are hydraulic structures widely used in open channels in order to divert excess water from a main channel into a side channel. They are commonly in rectangular, triangular, circular, and trapezoidal shapes, which have been utilized extensively in land irrigation, urban runoff drainage, and flood protection systems (Uyumaz 1997). Lateral outflow takes place when the water surface in the main channel rises above the side weir edge. Diverted flow over a side weir is a typical case of spatially varied flow which is one of the most complex flows to simulate in one dimensional flow analysis (Vatankhah 2013).

Accurate estimation of discharge coefficient (C_d) is an essential ingredient in design and operation of side weirs. A review of relevant literature indicates a number of investigations about spatially varied flow in different types of side weirs and channels (Subramanya & Awasthy 1972; El-Khashab & Smith 1976; Hager 1987; Uyumaz & Smith 1991; Uyumaz 1992; Borghei *et al.* 1999; Muslu 2001; Yüksel 2004; Vatankhah 2012, 2013). It is important to note that a distinctive approach for C_d estimation is not well known and no agreement is outward in the relevant literature about its value. Khorchani & Blanpain (2005) pointed out that C_d is the major source of uncertainty and lack of standardization in discharge estimation. A number of laboratory researches are still being carried out to modulate or optimize C_d estimation values (Bilhan *et al.* 2010; Emiroglu *et al.* 2010, 2011; Kisi *et al.* 2012; Dursun *et al.* 2012; Granata *et al.* 2013).

In recent years, data-driven techniques such as artificial neural network (ANN), genetic programming (GP) and fuzzy logic have been marked as an expedient tool for discharge

doi: 10.2166/hydro.2014.112

coefficients estimation. Khorchani & Blanpain (2005) used multilayer perceptron trained with back propagation algorithm in order to develop an ANN-based model for a side weir discharge estimation located at the outfall of the Lille urban catchment (North France). Bilhan *et al.* (2010) obtained C_d of sharp-crested rectangular side weirs in straight channels using feed forward (FF) and radial basis neural networks (RBNN). They used 843 experimental data for their simulations and indicated that the FF results in slightly better performance than the RBNN. Bilhan *et al.* (2011) reported the successful application of the ANN method to estimate the C_d of triangular labyrinth side weirs in curved channels. Kisi *et al.* (2012) applied the RBNN and generalized regression neural network techniques to predict C_d of triangular labyrinth side weirs. They demonstrated that the neural computing could be employed successfully in C_d estimation. An adaptive-neuro fuzzy inference system has been employed for C_d estimation at semi-elliptical side weirs and successful results have been reported by Dursun *et al.* (2012).

Despite providing sufficient estimation accuracy, all the aforementioned ANN-based models are implicit that may produce huge matrix of weights and biases. Thus, the necessity for further studies in order to develop not only explicit but also precise models is still receiving serious attention. In recent years, different variants/advancements of GP have been pronounced as robust explicit methods to solve a wide range of modelling problems in water resources engineering such as rainfall-runoff modelling (Whigham & Crapper 2001; Dorado *et al.* 2003; Rodríguez-Vázquez *et al.* 2012; Nourani *et al.* 2012, 2013) precipitation forecasting (Kisi & Shiri 2011), Chézy resistance coefficient determination (Giustolisi 2004), streamflow prediction (Danandeh Mehr *et al.* 2013), unit hydrograph determination (Rabuñal *et al.* 2007), sediment transport (Aytek & Kisi 2008), sea level forecasting (Ghorbani *et al.* 2010), evaporation modelling (Kisi & Guven 2010), bridge pier scour prediction (Azamathulla *et al.* 2010), hydrograph routing (Sivapragasam *et al.* 2008; Fallah-Mehdipour *et al.* 2013a), reservoir operation (Fallah-Mehdipour *et al.* 2013b), critical depth in open channels (Sharifi *et al.* 2011) and others. It was observed that only one study existed in the relevant literature related to the application of any variant/advancement of GP in C_d estimation. Kisi *et al.* (2012) developed an explicit gene-expression programming (GEP) model for this aim at triangular labyrinth side weirs. The results indicated that the GEP can be successfully used to modulate the nonlinear features of C_d in subcritical flow condition. GEP is an advancement of GP, which is also used in other hydraulic problems such as longitudinal dispersion prediction in pipelines (Sattar 2013a), pier scour depth prediction (Khan *et al.* 2012), dam breach parameters estimation (Sattar 2013b) and others.

The main goals of the current study are: (i) to evaluate the linear genetic programming (LGP) ability to model rectangular side weirs C_d variation in a circular channel for the first time; and (ii) to develop an explicit C_d estimation model applicable in both subcritical and supercritical approach flow conditions. The latter goal also indicates an effort to optimize the empirical relations developed by Uyumaz & Muslu (1985) to estimate C_d values at the same hydraulic conditions. Following this, the dimensionless experimental data sets taken from the study conducted by Uyumaz (1982) and reported by Uyumaz & Muslu (1985) are utilized in this study. To evaluate the capability of LGP for our modelling cases, we firstly put forward two different subsets of LGP-based models containing conditional-based LGP (CLGP) and mathematical-based LGP (MLGP). Then, these models were performed for each of the sub- and supercritical flow conditions separately. After that, we compared the estimation results of each model, at both the training and validation steps, with the corresponding observations as well as the results of a classic multilinear regression (MLR) and those of Uyumaz & Muslu (1985) models. In order to develop a single explicit C_d estimation function, at first, we distinguished more effective variables in C_d formation via a sensitivity analysis among the input parameters. Then, a single MLGP model was developed using more effective variables of both sub- and supercritical flow conditions. Ultimately, the proposed model was validated based upon all experimental observations and some experimental models acknowledged in the technical literature.

RECTANGULAR SIDE WEIR IN CIRCULAR CHANNEL

According to Figure 1, the general expression for the surface profile along the side weirs in the circular channel (Equation (1)) was derived by Uyumaz & Muslu (1985) using the conventional weir equation for discharge per unit length (El-Khashab & Smith 1976) and assumptions including

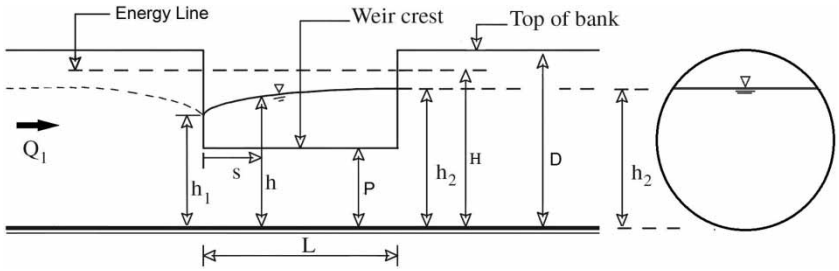

Figure 1 | Definition sketch for circular channel with lateral flow over rectangular side weir.

(i) the constant specific energy, (ii) two-dimensional flow and (iii) hydrostatic pressure along the main channel

$$\frac{da}{ds}\sqrt{2g(H-h)} - a\frac{\sqrt{2g}}{2\sqrt{(H-h)}}\frac{dh}{ds} = -C_d\sqrt{2g(h-P)}(h-P) \quad (1)$$

where a is cross-sectional area of flow, C_d is discharge coefficient, P is the weir height, h is depth of flow, H is constant specific energy, g is gravitational constant and s is the distance measured along the channel. A dimensionless form of Equation (1) resulted from dividing the variables by the channel diameter, D, and is presented in Equation (2):

$$\frac{ds}{D} = -\frac{\pi}{8C_d}F(t)dt \quad (2)$$

where t equals h/D and $F(t)$ is a function of dimensionless geometric characteristics of the weir and flow. C_d is mainly a function of the non-constant side weir head (h) along the weir's length and is influenced by the following parameters (Uyumaz & Muslu 1985):

$$C_d = f(v, h, P, L, D, g, S_0, a) \quad (3)$$

where v is mean velocity of flow, L is side weir length, and S_0 is the slope of the main channel. Using the Buckingham theorem, non-dimensional equations in functional forms can be obtained as below:

$$C_d = f(F_r, P/D, L/D) \quad (4)$$

This functional relationship has been worked out in the present study, in which F_r is main channel Froude number at the beginning of the side weir. Since v is a function of both S_0 and a and F_r is a function of v and g, consequently this relationship covers all parameters considered in Equation (3).

EXPERIMENTS AND OBSERVATION DATA

The data used in this study were taken from the experimental study reported by Uyumaz & Muslu (1985) which was performed in the hydraulics laboratory of Istanbul Technical University, Turkey. As shown in Figure 2, these experiments were performed in a circular concrete channel with 0.25 m diameter and 10.9 m length. The incoming water from the upper tank enters the initial part of the channel under gravity force. The discharge was controlled by a 90° triangular weir and was calmed by various means such as submerged baffles and floating plates before entering the channel. As the water enters the side weir location, at the mid-point of the channel approximately, a part spills over and the remainder flows towards the lower water tank. An adjustable gate located at the end of the channel had been used for water level control. The side weir discharge was measured by a 90° triangular weir and the corresponding C_d was calculated with a conventional weir equation using the average head along the weir (Uyumaz & Muslu 1985).

Experiments were carried out for both sub- and supercritical flow condition in the channel and free overflow conditions in the side weir. The experiments were conducted for different lengths of the weir, L (0.15, 0.25, 0.40, 0.50, 0.60, 0.75, and 0.85 m), weir crest heights P (0.06, 0.08, 0.10, 0.12, and 0.14 m) and channel slope varying between 0.00 and 0.02. A total of 1,686 combinations were examined for various values of the Froude number, different P/D ratio (0.24, 0.32, 0.40, 0.48, and 0.56), and different L/D

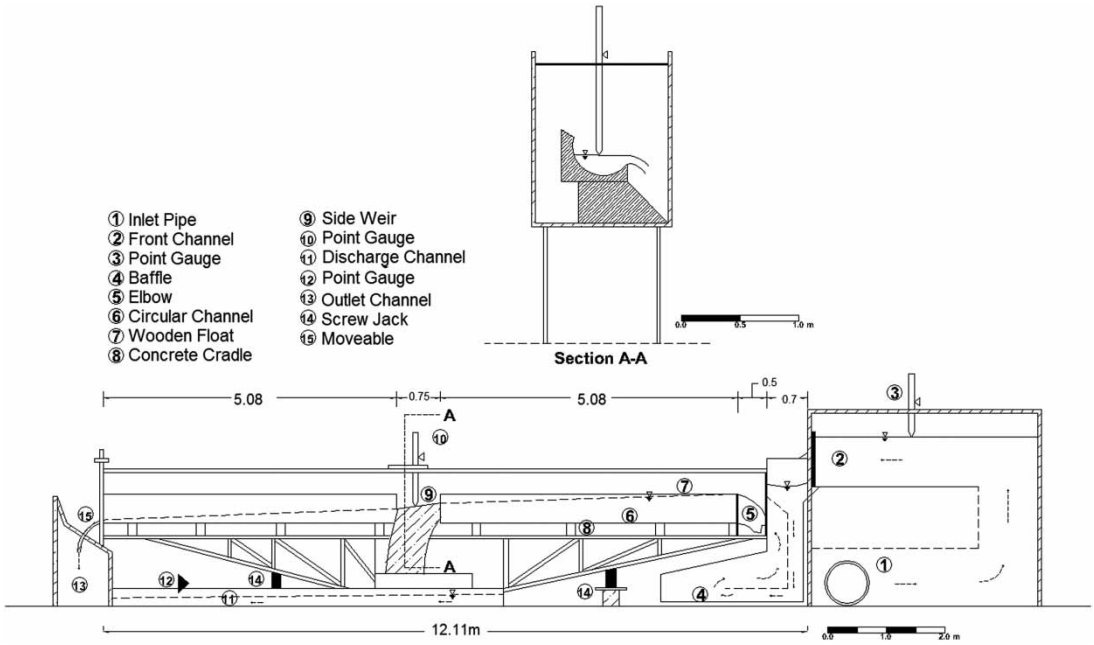

Figure 2 | Experimental arrangement, longitudinal and cross-sections.

(0.6, 1.0, 1.6, 2.0, 2.4, 3.0, and 3.4) ratio in order to map the comprehensive variation of C_d. A range of test variables are given in Table 1 and a three-dimensional scatterplot of the entire experiment is depicted in Figure 3. Each point in the figure belongs to a constant P/D ratio. The figure shows that the C_d increases when L/D ratio increases and/or F_r decreases. The figure also indicated that C_d of the side weir has a nonlinear and linear variation with respect to the F_r for sub- and supercritical regimes, respectively. It is the reason why we firstly generated and assessed different CLGP and MLGP models for each kind of approach flow regime. Based upon graphical correlation among the experimental data, Uyumaz & Muslu (1985) have suggested Equation (5) (hereafter UM5) and Equation (6) (hereafter UM6) to estimate C_d of rectangular side weirs in sub- and supercritical condition, respectively

$$C_d = \left\{0.21 + 0.094\sqrt{1.75(L/D) - 1}\right\} + \left\{0.22 - 0.08\sqrt{1.68(L/D) - 1}\right\}\sqrt{1 - F_r} \quad (5)$$

$$C_d = -\left\{0.046 + 0.0054\sqrt{1.67(L/D) - 1}\right\}F_r + \left\{0.24 + 0.021\sqrt{1 + 35.3(L/D)}\right\} \quad (6)$$

Table 1 | Range of variables used in experiments

P (m)	L (m)	P/D[a]	L/D	F_r	Q(m^3/s)	Number of experiments
0.06	0.15–0.85	0.24	0.6–3.4	0.15–2.0	0.008–0.032	395
0.08	0.15–0.85	0.32	0.6–3.4	0.1–2.0	0.008–0.032	433
0.1	0.15–0.85	0.4	0.6–3.4	0.1–1.8	0.008–0.032	385
0.12	0.15–0.85	0.48	0.6–3.4	0.08–1.2	0.008–0.032	263
0.14	0.15–0.85	0.56	0.6–3.4	0.06–0.95	0.008–0.032	210

[a]$D = 0.25$ m.

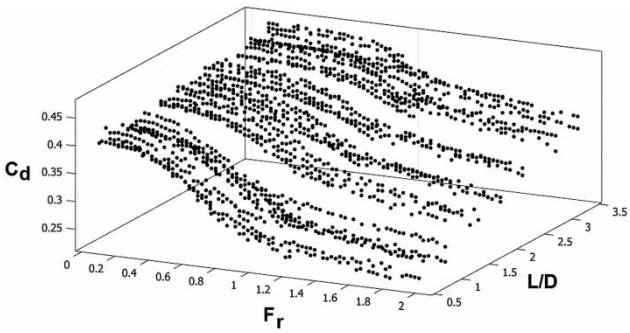

Figure 3 | Scatter plot of experimental *Cd* versus F_r and *L/D*.

Linear genetic programming

At the most brief level LGP is a steady-state, evolutionary algorithm using a fitness-based tournament selection to continuously improve a population of machine-code functions (Francone 2009). Canonical genetic programming (GP) holds candidate solutions (programs) in a tree-based genome and the transformation operators (crossover and mutation) act on tree-based genomes (Koza 1992). LGP is distinct from canonical GP systems in that the transformation operators act on a linear (not tree-based) genome (Banzhaf *et al.* 1998). A number of researchers have reported successful application of the linear genome in GP (e.g. Brameier & Banzhaf 2007; Danandeh Mehr *et al.* 2013).

Generally, LGP solves any problem through the following six steps: (i) generation of an initial population (machine-code functions) randomly by the user defined functions and terminals; (ii) selection of two functions from the population randomly, comparison of the outputs and designation of the function that is more fit as winner_1 and less fit as loser_1; (iii) selection of two other functions from the initial population randomly and designation of the winner_2 and loser_2; (iv) application of transformation (crossover and mutation) operators to winner_1 and winner_2 to create two similar but different evolved programs (i.e. offspring) as modified winners; (v) replace the loser_1 and loser_2 in the population with modified winners; and (vi) repetition of steps (i)–(v) until the predefined run termination criterion. The result of such effort is usually a program (Code), which requires additional effort of the modeller to derive an explicit equation from the program. However, there is no guarantee that such a favourable equation will be achieved.

In this study, we have considered two different types of LGP-based models (i.e. CLGP and MLGP) in order to be able to convert the codes to explicit formulations. The CLGP and MLGP models are differed based upon the allowable functions used for generating of initial populations (potential solutions). In CLGP, the initial populations (potential solutions) can be generated by a combination of different Boolean logic, conditional, transfer and arithmetic functions as well as common basic addition, subtraction and multiplication functions, whereas in MLGP the initial populations (potential solutions) are confined to be a combination of only mathematical functions. Due to this constraint, MLGP results in such programs that can be represented in explicit formulation as a potential solution, whereas CLGP may lead to a complex code. The allowable functions that we used for each kind of LGP-based models are tabulated in Table 2.

An example of a LGP-based model, which is a part of the C code of the MLGP model developed in this study for supercritical flow condition, is illustrated as follows:

```
L0: f [0] += 0.1387641429901123 f;
L1: f [0]/ = Input001;
L2: f [0]/ = Input000;
L3: f [0]/ = −0.6102392673492432 f;
L5: f [0]* = f [0];
L6: f [0]* = f [0];
L7: f [0]/ = Input000;
L8: f [0]* = 0.03275442123413086 f;
L9: f [0] = sin (f [0]);
```

Table 2 | Function sets used for CLGP and MLGP modellings (1: allowed, 0: non-allowed)

Function set	Tasks	CLGP	MLGP
Basic	Addition, subtraction, multiplication	1	1
Arithmetic	Absolute value, change sign, scaling[a], square root	1	1
Comparison	Application of less than (<) command	1	0
Conditional	Application of If, Then, Else and Go to commands	1	0
Data transfer	Moving values around without changing the values	1	0
Division	Dividing, calculating the reminder	1	1
Exponential	$x^2 - 1$	0	1
Trigonometric	Cosine, Sine	0	1

[a]This task replace a temporary variable, *x*, with two raised to the power.

where f[0] represents the temporary computation variable created in the program by LGP. LGP uses such temporary computation variables to store values while performing calculations. The variable f[0] is initialized to zero in this program and the output is the value remaining in f[0] in the last line of the code.

To apply LGP for any problem, besides function selection, several decisions such as determination of input/output variables, fitness function, and crossover and mutation rates are required to be made (Danandeh Mehr *et al.* 2013). The fitness function value is used to rank the randomly generated initial programs and then new programs are created by using both crossover and mutation operators (Francone 2009). In our problem, the learning task is to fit a function for experimental values of C_d (target outputs). Therefore, fitness function should clarify how closely the estimated C_d of the evolved program matches the target outputs in the training/validation sets. Consequently, the mean square error (MSE) was utilized as the fitness function in our LGP system settings. The lower the MSE, the more the evolved program fits. The error for each training/validation set is the difference between the output of an evolved program and the corresponding observed C_d. According to Equation (4), we also utilized dimensionless P/D, L/D and F_r sets as input variables to the LGP system.

In addition to parameter selection, one of the main concerns of LGP modelling as well as other artificial intelligence models is the overfitting (overtraining) problem. It means that the model works very well on the training data, but not so well on the validation set (which is not used in training). This is more likely to happen when a small data set or large number of generations is used for LGP runs. Considering the plausible number of training and validation data, in order to prevent the overfitting in this research we firstly confined the maximum number of generations at both CLGP and MLGP runs to 1,000 generation and also the maximum size of the program was limited to 512 bytes (see Table 3). Then we simultaneously monitored the training and validation errors in the LGP runs to stop each run whenever the error on the validation data began to rise. This avoidance overfitting technique is a combination of methods that have already been suggested by Giustolisi (2004) and Nourani *et al.* (2013). The other parameters adopted for both CLGP and MLGP setting in this study are listed in Table 3.

Table 3 | Parameter settings for the LGP algorithm

Parameter	Value
Initial populations (programs)	500
Mutation frequency	95%
Crossover frequency	% 50
Initial program size	80 (Byte)
Maximum program size	512 (Byte)
Random constants	(−1,1)
Generation without improvement	300
Generation since start	1,000

The initial population parameter that the modeller uses to begin the evolutionary process sets the number of programs in the population. There is no upper limit for population size but it is often considered in the range of 100–1,000 for a variety of hydraulic estimations. Further details on LGP parameters can be found in Brameier & Banzhaf (2007), Poli *et al.* (2008) and Francone (2010).

Efficiency criteria

Among different models developed in this study and selected from the literature, the model that yields the best results in terms of determination coefficient (R^2), root mean squared errors (RMSE), and mean absolute errors (MAE) on both training and validation steps are selected as the most efficient models in this study

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(X_i^{\text{obs}} - X_i^{\text{pre}}\right)^2}{\sum_{i=1}^{n} \left(X_i^{\text{obs}} - X_{\text{mean}}^{\text{obs}}\right)^2} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i^{\text{obs}} - X_i^{\text{pre}}\right)^2}{n}} \tag{8}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left|X_i^{\text{obs}} - X_i^{\text{pre}}\right|}{n} \tag{9}$$

where X_i^{obs} = observed value of X, X_i^{pre} = estimated value X_{mean}^{obs} = mean value of observed data and n = number of observed data.

The RMSE sizes the goodness of the fit related to high C_d values, whereas the MAE measures a more balanced perspective of the goodness of the fit at moderate ones (Dursun *et al.* 2012). Implementation of RMSE alone does not provide a clear interpretation of the model average error. RMSE tends to become increasingly larger than MAE as the distribution of error magnitudes becomes more variable (Willmott & Matsuura 2005). These statistics can be used together to diagnose the variation in the errors in a set of forecasts. The greater difference between them, the greater the variance in the individual errors in the sample. Obviously, a high value for R^2 (up to one) and small values for RMSE and MAE indicate high efficiency of the model.

RESULTS AND DISCUSSION

As mentioned previously, in order to estimate the C_d of rectangular side weirs using LGP-based models, we carried out this study in three different simulation phases. In the first phase, different CLGP and MLGP models were developed, evaluated and assigned for each of the subcritical and supercritical flow conditions separately with respect to the Froude number at the beginning of the side weir. In the second phase, we generated and evaluated new CLGP and MLGP models disregarding upstream flow condition. This effort provided a general C_d estimator valid at any approach flow condition. Ultimately, the genetic-based sensitivity analysis has been performed among the input variables and a new MLGP model was developed discarding the least effective input variable. The performances of LGP-based models in all phases were also compared with those of the conventional MLR and nonlinear models selected from the literature. We applied Discipulus®, the LGP software package developed by Francone (2010), to establish all proposed CLGP and MLGP models.

Subcritical and supercritical approach flow

Prior to the generation of any program by CLGP or MLGP, we divided the entire experimental data into two sub- and supercritical sets with respect to the corresponding measured Froude number. Then, 70 and 30% of each set were randomly selected as training and validation subsets of CLGP or MLGP models, respectively. Table 4 shows the statistical features of each subset. Random splitting is commonly used in machine learning to improve the robustness of results.

The efficiency results of the best generated CLGP and MLGP models for each individual set are given in Table 5 and compared with those of MLR and UM5 and UM6 models. The functional form of the best MLGP models for sub- and supercritical conditions were presented in Equations (10) and (11), respectively. Equations (12) and (13) also show the developed MLR functions for sub- and supercritical conditions, respectively

$$C_d = \sqrt{0.16 + 0.014A + 0.04B - 0.113F_r} \quad (F_r < 1) \tag{10}$$

where

$$A = \sqrt{\frac{((0.6 + 0.413F_r)/(P/D)) - 0.35F_r}{(P/D)} + (L/D)}$$

$$B = F_r(L/D) - F_r^2$$

$$C_d = 0.138 + \mathrm{Sin}\left\{\frac{1.096 + \mathrm{Sin}\,A - (P/D) + 1.096}{2B}\right\} \quad (F_r > 1) \tag{11}$$

where

$$A = \frac{0.00009}{(L/D)^4(P/D)^5}$$

$$B = 0.033F_r(L/D)$$

$$C_d = 0.446 - 0.105(P/D) + 0.027(L/D) - 0.094F_r \quad (F_r < 1) \tag{12}$$

$$C_d = 0.356 - 0.131(P/D) + 0.046(L/D) - 0.055F_r \quad (F_r > 1) \tag{13}$$

Table 4 | Statistical parameters of applied subsets

			Subcritical set ($F_r < 1$)			Supercritical set ($F_r \geq 1$)		
Parameter			**Entire data**	**Training set**	**Validation set**	**Entire data**	**Training set**	**Validation set**
Input variables	P/D	N	1,029	720	309	657	460	197
		Min	0.24	0.24	0.24	0.24	0.24	0.24
		Max	0.56	0.56	0.56	0.48	0.48	0.48
		Mean	0.405	0.408	0.398	0.327	0.325	0.330
		St. D	0.111	0.111	0.111	0.075	0.075	0.076
	L/D	N	1,029	720	309	657	460	197
		Min	0.6	0.6	0.6	0.6	0.6	0.6
		Max	3.4	3.4	3.4	3.4	3.4	3.4
		Mean	1.962	1.972	1.940	2.020	2.01	2.042
		St. D	0.951	0.945	0.964	0.952	0.953	0.952
	F_r	N	1,029	720	309	657	460	197
		Min	0.06	0.075	0.06	1.0	1.0	1.0
		Max	0.99	0.99	0.986	2.0	2.0	2.0
		Mean	0.549	0.508	0.645	1.419	1.426	1.402
		St. D	0.255	0.247	0.250	0.300	0.300	0.299
Target variable	C_d	N	1,029	720	309	657	460	197
		Min	0.26	0.26	0.272	0.223	0.223	0.227
		Max	0.471	0.471	0.468	0.42	0.414	0.42
		Mean	0.403	0.407	0.394	0.327	0.326	0.330
		St. D	0.040	0.038	0.043	0.048	0.048	0.046

N: Number of data; St. D: Standard deviation.

Table 5 | Goodness of fitness comparison of different C_d estimation models

	Subcritical						Supercritical					
	Training			Validation			Training			Validation		
Model	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**
CLGP	0.008	0.951	0.007	0.009	0.953	0.007	0.010	0.956	0.008	0.009	0.958	0.007
MLGP	0.008	0.953	0.007	0.008	0.965	0.006	0.009	0.959	0.007	0.010	0.951	0.008
MLR	0.017	0.821	0.012	0.018	0.828	0.014	0.011	0.952	0.009	0.010	0.950	0.008
UM5-UM6	0.019	0.764	0.016	0.019	0.810	0.016	0.018	0.861	0.014	0.017	0.868	0.014

Table 5 demonstrates that both CLGP and MLGP models perform better than the MLR, UM5 and UM6, particularly in the subcritical flow regime. The reason behind this may be relevant to the fact that C_d in the circular channel shows a nonlinear variation with respect to the Froude number in subcritical regimes (see Figure 3). Despite the mentioned nonlinear behaviour, the linear MLR's Equation (12) validation performance (RMSE = 0.018, $R^2 = 0.828$, MAE = 0.014) is observed to be highly close to those of nonlinear UM5 (RMSE = 0.019, $R^2 = 0.810$, MAE = 0.016). The reason behind this can be due to the effect of P/D on C_d that has been neglected in UM5. At supercritical flow conditions, the linear MLR's Equation (13) performance (RMSE = 0.010, $R^2 = 0.950$, MAE = 0.008) is also observed to be highly close to those of nonlinear MLGP's Equation (11) (RMSE = 0.010, $R^2 = 0.951$, MAE = 0.008). This implies the linearity of the relationship among investigated parameters at supercritical flow conditions. The MLR estimation results at both sub- and supercritical conditions are very promising. Since the MLR technique serves a

simpler formula, it can be offered as an alternative to UM6 at supercritical conditions; however, the most accurate results are provided by LGP-based models. In spite of confined function sets in the MLGP model, it provides higher performance than the CLGP in the subcritical flow regime. This result indicates that the best LGP program is not necessarily produced when a variety of functions are employed at the initial population generation. It is consistent with the results of Nourani *et al.* (2013) who developed more effective GP-based rainfall-runoff models using only basic arithmetic functions.

General C_d estimator

High performance of the proposed LGP-based models in both sub- and supercritical flow conditions motivated us to suppose that a single LGP-based model may properly estimate the value of C_d at both flow conditions. Therefore, in the second phase of the study, all 1,686 experimental data were subjected to develop a general estimation model using MLGP and CLGP approaches while neglecting the approach flow condition. For this aim, first, 70 and 30% of the entire data were randomly selected as training and validation sets, respectively. Then, the new MLGP and CLGP models were established using the similar functional and parameter sets of the first phase of the study. Statistical characteristics of training and validation sets of general formulation of the C_d are listed in Table 6.

The efficiency results of the developed general CLGP and MLGP estimation models are shown in Table 7. By comparing the performance of the general LGP-based models with the previously developed individual models given in Table 6 it can be concluded that the general models produced acceptable outcomes as accurate as individual models of each approach flow condition. Hence, neglecting the approach flow condition seems to be a plausible assumption. Therefore, in contrast with formulas given by Uyumaz & Muslu (1985), there is no obligation to consider the different flow conditions for C_d estimation. In order to obtain more reliable evaluation of performance of the general CLGP and MLGP models, MLR analysis was also performed for the entire data and the corresponding efficiency results were compared with those of CLGP and MLGP (see Table 7). The functional form of the general

Table 6 | Statistical parameters of general estimator subsets

Parameter			Entire data	Training set	Validation set
Input variables	P/D	N	1,686	1,180	506
		Min	0.24	0.24	0.24
		Max	0.56	0.56	0.56
		Mean	0.374	0.373	0.377
		St. D	0.106	0.107	0.103
	L/D	N	1,686	1,180	506
		Min	0.6	0.6	0.6
		Max	3.4	3.4	3.4
		Mean	1.985	1.964	2.03
		St. D	0.951	0.938	0.982
	F_r	N	1,686	1,180	506
		Min	0.06	0.06	0.07
		Max	2.00	2.00	2.00
		Mean	0.888	0.896	0.868
		St. D	0.505	0.506	0.503
Target variable	C_d	N	1,686	1,180	506
		Min	0.223	0.223	0.231
		Max	0.471	0.471	0.468
		Mean	0.374	0.373	0.376
		St. D	0.057	0.058	0.056

N: Number of data; St. D: Standard deviation.

Table 7 | Comparison of the performance of the estimation models

Model	Training			Validation		
	RMSE	R^2	MAE	RMSE	R^2	MAE
CLGP	0.013	0.949	0.010	0.013	0.950	0.010
MLGP Equation (14)	0.015	0.930	0.011	0.014	0.934	0.011
MLR Equation (15)	0.018	0.902	0.014	0.018	0.891	0.015
MLGP Equation (16)	0.015	0.924	0.012	0.016	0.924	0.013
MLR Equation (17)	0.022	0.863	0.017	0.023	0.864	0.017

MLGP and MLR models predicting C_d at any kind of flow condition are given in Equations (14) and (15), respectively

$$C_d = 0.227 + 0.218|\mathrm{Cos}(0.514A(P/D) - B)| \quad \text{MLGP} \qquad (14)$$

where

$$A = 2(0.9F_r - (P/D))^2 - (P/D) - 0.56$$

$$B = 0.347 + 0.27(L/D) - F_r$$

$$C_d = 0.424 - 0.105(P/D) + 0.036(L/D) - 0.093F_r \quad \text{MLR} \tag{15}$$

Sensitivity analysis

As shown in Equations (5) and (6), the UM5 and UM6 models estimate C_d using only L/D and F_r parameters and neglecting P/D. It is also inevitable that more input variables in evolutionary computing methods may lead to more complex formulations (Nourani *et al*. 2012). Therefore, in the last phase of the study, firstly, an attempt was made to distinguish the most effective variables in C_d estimation by sensitivity analysis. Then, using the two more dominant variables, the MLGP's Equation (14) and MLR's Equation (15) were reanalysed for the same training and validation sets in order to generate two-variable estimation models which will serve more fair comparison with UM5 and UM6.

To perform sensitivity analysis, the importance of input variables in creating the best 30 programs, corresponding to the MLGP's Equation (14), were considered. The input variable impacts for the 30 best MLGP models in terms of frequency, average and maximum impact are listed in Table 8. The frequency values show what percentage of the best 30 programs from the model contained the referenced input. The average and maximum impact columns show the average and the maximum effect of removing the referenced input from each of the best programs and replace it with a permuted version of that input, respectively.

According to Table 8, it can be concluded that although all input variables were selected to model the best 30 programs, P/D has the least removal impact in the value of C_d. The greatest impacts belong to the F_r and L/D parameters, respectively. This consequence is consistent with the results of graphic based sensitivity analysis implemented by Uyumaz & Muslu (1985) for the same input variables. Using the distinguished dominant parameters (i.e. F_r and L/D) and similar training and validation sets of general C_d estimator models, the new two-variable MLGP and MLR models were generated for C_d estimation as given in Equations (16) and (17), respectively. Performance results of MLGP's Equation (16) and MLR's Equation (17) are also given in Table 7

$$C_d = 0.264 + 0.176\text{Cos}(2^A - 1) \tag{16}$$

where

$$A = 2^{\sqrt{2^B - 1}} - 1$$

$$B = F_r\{1.5 + 0.5\cos[2.352(L/D) + 0.352\cos(3.163F_r)]\}$$

$$C_d = 0.379 + 0.035(L/D) - 0.084F_r \tag{17}$$

At the final stage, in order to validate effectiveness and physical consistency of the proposed two-variable MLGP model (i.e. Equation (16)), the model was applied for entire sub- and supercritical experimental data sets (Table 4, columns 4 and 7) and its efficiency results were compared with those of two-variable MLR (i.e. Equation (17)), UM5, UM6 models in Table 9 as well as general solutions developed by Hager (1993, 1994) (Equations (18) and (19)) for circular pipes.

$$C_d = 0.40 + 0.01(L/D) - 0.185F_r^2/(L/D) \quad (F_r < 1) \tag{18}$$

$$C_d = 0.042(7 + L/D - 1.25F_r) \quad (F_r > 1) \tag{19}$$

Table 8 | Input variable impacts in MLGP modeling

Variable	Frequency (1.00 = 100%)	Average impact	Maximum impact
P/D	1.00	0.036	0.111
L/D	1.00	0.638	0.936
F_r	1.00	0.825	0.948

Table 9 | Effectiveness of two-variable estimation models

	Subcritical flow			Supercritical flow		
Model	RMSE	R^2	MAE	RMSE	R^2	MAE
MLGP Equation (16)	0.016	0.850	0.013	0.015	0.895	0.012
UM5 and UM6	0.019	0.786	0.016	0.018	0.863	0.014
MLR Equation (17)	0.021	0.725	0.017	0.021	0.809	0.016
Hager Equations ((18) and (19))	0.040	0.012	0.030	0.027	0.677	0.023

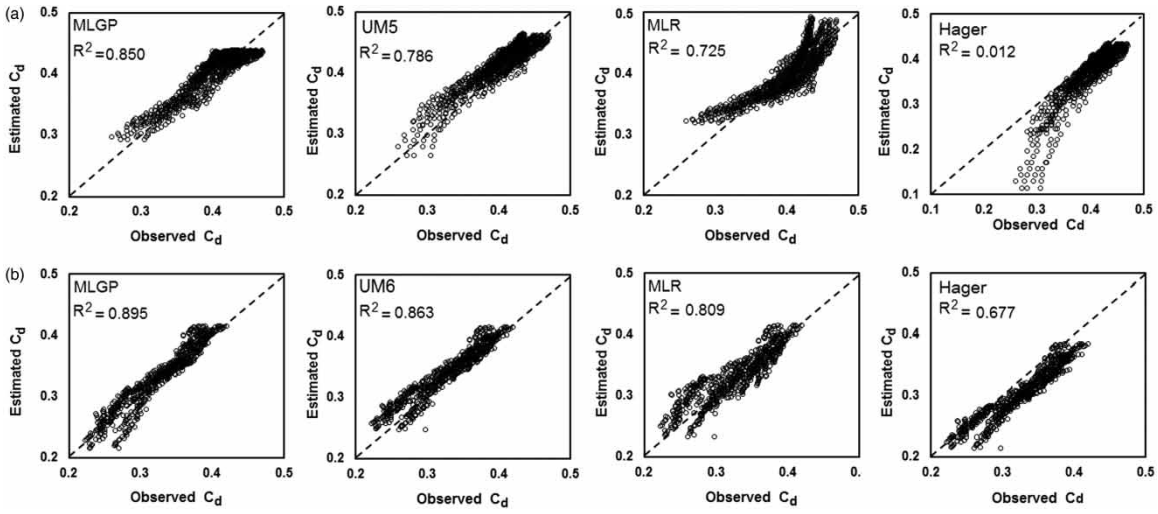

Figure 4 | Scatter plots of MLGP, UM5, UM6, MLR, and Hager models estimations for (a) subcritical and (b) supercritical approach flow.

The models of Hager (1993, 1994) have been considered suitable for the comparison because these not only relate to circular pipes but also use the dimensionless parameters identical to ours.

Table 9 indicates that the MLGP model is superior to its counterparts with respect to various performance criteria. In Figure 4, estimated C_d values obtained by each model (i.e. Equations (5), (6) and (16)–(19)) were plotted against the corresponding experimental values. It can be seen from Figure 4 that the MLGP estimates closer to the observed values with a higher R^2 value than those of the UM5, UM6, MLR and Hager models, particularly in subcritical flow condition. The MLR in subcritical condition has a highly scattered distribution that implies the nonlinearity of the investigated phenomenon in subcritical flow conditions. In contrast, the MLR seems to be better than the Hager in both sub- and supercritical flow condition; however, the latter serves nonlinear formulation. It indicates the linearity of the phenomenon in supercritical flow condition which previously had been demonstrated in Figure 3. In supercritical flow conditions (Figure 4(b)), all models apart from Hager seem to be acceptable for the estimation of discharge coefficients ($R^2 > 0.8$). However, the MLGP is still superior to its counterparts. Furthermore, it provides a fixed model for both approach flow conditions.

It is evident that for both sub- and supercritical approach flow, the UM5 and UM6 overestimate the effective C_d, whereas there is a significant tendency for the Hager to underestimate the effective C_d, particularly in the subcritical approach flow.

Monitoring the residuals between the Hager estimations and effective C_d values reveals the fact that high uncertainty zone of Hager estimations in the subcritical approach flow (i.e. estimated C_d values less than 0.3) belongs to experimental C_d values corresponding to the L/D less than 1. Therefore, Equation (18) is only valid for weirs longer than the channel diameter ($L > D$).

CONCLUSIONS

In this study, for the first time two variants of the LGP approach containing CLGP and MLGP models have been developed to estimate the discharge coefficient of sharp-edged rectangular side weirs in circular channel. A number of 1,686 experimental data sets were used for the CLGP and MLGP simulations in both sub- and supercritical flow regime. Input parameters utilized for the simulations are dimensionless weir length, dimensionless weir height, and approach flow Froude number. Using the observation data, the performance of LGP-based models were compared with those of the corresponding MLR, the nonlinear UM5 and UM6 models suggested by Uyumaz & Muslu (1985), and the nonlinear models suggested by Hager (1993, 1994). Comparing the results of different estimation models indicates that the CLGP and MLGP models perform better

than their counterparts, particularly in subcritical flow regime. Application of the developed three-variable MLR model was also offered as an alternative to UM5 due to higher performance results of the model and linearity of the investigated phenomenon in supercritical flow condition. Sensitivity analysis among the input parameters indicated that the approach flow Froude number and dimensionless weir length have the highest impact in C_d, respectively. The study recommended a single explicit two-variable MLGP formulation to estimate C_d of rectangular side weirs in circular channels with R^2 values more than 85–89% in sub- and supercritical flow conditions, respectively. This model reduces the RMSE with respect to those of UM5 and UM6 by 16% and 12% and those of Hager by 60% and 43% percentages in sub- and supercritical flow conditions, respectively.

Our review showed that there is currently too limited works available on the estimation of C_d of rectangular side weirs located in circular channels. As we only used a set of experimental data from a channel with fixed diameter and limited range of Froude number for higher P/D ratios, further studies using different experimental data sets may be required to strengthen our conclusions. As another suggestion for future research, the ability of LGP technique can be also investigated by presented methodology for C_d estimation of other kinds of side weirs.

ACKNOWLEDGEMENTS

We wish to express our sincere gratitude to the three anonymous reviewers and the associate editor of the paper whose suggestions and comments have greatly helped us to improve the quality of the manuscript. We also thank Prof. Zekai Şen and Mr Tewodros Assefa Nigussie, Istanbul Technical University, for their valuable contribution of ideas for improving the work.

REFERENCES

Aytek, A. & Kisi, O. 2008 A genetic programming approach to suspended sediment modelling. *J. Hydrol.* **351** (3–4), 288–298.

Azamathulla, H. M., Ghani, A., Zakaria, N. & Guven, A. 2010 Genetic programming to predict bridge pier scour. *J. Hydraul. Eng.* **136** (3), 165–169.

Banzhaf, W., Nordin, P., Keller, R. & Francone, F. D. 1998 *Genetic Programming – an Introduction to the Automatic Evolution of Computer Programs and its Application*. Morgan Kaufmann, Heidelberg, San Francisco.

Bilhan, O., Emiroglu, M. E. & Kisi, O. 2010 Application of two different neural network techniques to lateral outflow over rectangular side weirs located on a straight channel. *Adv. Eng. Softw.* **41** (6), 831–837.

Bilhan, O., Emiroglu, M. E. & Kisi, O. 2011 Use of artificial neural networks for prediction of discharge coefficient of triangular labyrinth side weir in curved channels. *Adv. Eng. Softw.* **42** (4), 208–214.

Borghei, S. M., Jalili, M. R. & Ghodsian, M. 1999 Discharge coefficient for sharp-crested side weir in subcritical flow. *J. Hydraul. Eng.* **125** (10), 1051–1056.

Brameier, M. & Banzhaf, W. 2007 *Linear Genetic Programming*. Springer Science and Business Media, LLC, New York.

Danandeh Mehr, A., Kahya, E. & Olyaie, E. 2013 Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *J. Hydrol.* **505**, 240–249.

Dorado, J., Rabunal, J. R., 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.

Dursun, O. F., Kaya, N. & Firat, M. 2012 Estimating discharge coefficient of semi-elliptical side weir using ANFIS. *J. Hydrol.* **426–427**, 55–62.

Emiroglu, M. E., Kisi, O. & Bilhan, O. 2010 Predicting discharge capacity of triangular labyrinth side weir located on a straight channel by using an adaptive neuro-fuzzy technique. *Adv. Eng. Softw.* **41** (2), 154–160.

Emiroglu, M. E., Bilhan, O. & Kisi, O. 2011 Neural networks for estimation of discharge capacity of triangular labyrinth side-weir located on a straight channel. *Expert Syst. Appl.* **38** (1), 867–874.

El-Khashab, A. & Smith, K. V. H. 1976 Experimental investigation of flow over side weirs. *J. Hydraul. Div.* **102** (9), 1255–1268.

Fallah-Mehdipour, E., Bozorg Haddad, O., Orouji, H. & Mariño, M. A. 2013a Application of genetic programming in stage hydrograph routing of open channels. *Water Resour. Manag.* **27** (9), 3261–3272.

Fallah-Mehdipour, E., Bozorg Haddad, O. & Mariño, M. A. 2013b Developing reservoir operational decision rule by genetic programming. *J. Hydroinform.* **15** (1), 103–119.

Francone, F. D. 2009 Dynamics and Performance of a Linear Genetic Programming System. Thesis for the degree licentiate in complex systems. Chalmers University of Technology, Göteborg, Sweden.

Francone, F. D. 2010 *Discipulus™ with Notitia and Solution Analytics Owner's Manual*. Register Machine Learning Technologies, Inc., Littleton, CO, USA.

Giustolisi, O. 2004 Using genetic programming to determine Chèzy resistance coefficient in corrugated channels. *J. Hydroinform* **6** (3), 157–173.

Ghorbani, M. A., Khatibi, R., Aytek, A., Makarynskyy, O. & Shiri, J. 2010 Sea water level forecasting using genetic programming and comparing the performance with artificial neural networks. *Comput. Geosci.* **36** (5), 620–627.

Granata, F., de Marinis, G., Gargano, R. & Tricarico, C. 2013 Novel approach for side weirs in supercritical flow. *J. Irrig. Drain Eng.* **139** (8), 672–679.

Hager, W. H. 1987 Lateral outflow over side weirs. *J. Hydraul. Eng.* **113** (4), 491–504.

Hager, W. H. 1993 Streichwehre mit Kreisprofil (sideweirs with a circular profile). *Wasser/Abwasser* **134** (3), 156–163 (in German).

Hager, W. H. 1994 Supercritical flow in circular-shaped sideweirs. *J. Irrig. Drain Eng.* **120** (1), 1–12.

Khan, M., Azamathulla, H. M. d. & Tufail, M. 2012 Gene-expression programming to predict pier scour depth using laboratory data. *J. Hydroinform.* **14** (3), 628–645.

Khorchani, M. & Blanpain, O. 2005 Development of a discharge equation for side weirs using artificial neural networks. *J. Hydroinform.* **7** (1), 31–39.

Kisi, O. & Guven, A. 2010 Evapotranspiration modeling using linear genetic programming technique. *J. Irrig. Drain Eng.* **136** (10), 715–723.

Kisi, O. & Shiri, J. 2011 Precipitation forecasting using wavelet-genetic programming and wavelet-Neuro-fuzzy conjunction models. *Water Resour. Manag.* **25** (13), 3135–3152.

Kisi, O., Emiroglu, M. E., Bilhan, O. & Guven, A. 2012 Prediction of lateral outflow over triangular labyrinth side weirs under subcritical conditions using soft computing approaches. *Expert Syst. Appl.* **39** (3), 3454–3460.

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

Muslu, Y. 2001 Numerical analysis for lateral weir flow. *J. Irrig. Drain. Eng.* **127** (4), 246–253.

Nourani, V., Komasi, M. & Alami, M. T. 2012 Hybrid wavelet–genetic programming approach to optimize ANN modelling of rainfall–runoff process. *J. Hydrol. Eng.* **17** (6), 724–741.

Nourani, V., Komasi, M. & Alami, M. T. 2013 Geomorphology-based genetic programming approach for rainfall-runoff modeling. *J. Hydroinform.* **15** (2), 427–445.

Poli, R., Langdon, W. B. & McPhee, N. F. 2008 A field guide to genetic programming. www.gp-field-guide.org.uk.

Rabuñal, J. R., Puertas, J., Suárez, J. & Rivero, D. 2007 Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. *Hydrol. Process.* **21** (4), 476–485.

Rodríguez-Vázquez, K., Arganis-Juárez, M. L., Cruickshank-Villanueva, C. & Domínguez- Mora, R. 2012 Rainfall-runoff modelling using genetic programming. *J. Hydroinform.* **14** (1), 108–121.

Sattar, A. 2013a Gene expression models for the prediction of longitudinal dispersion coefficients in transitional and turbulent pipe flow. *J. Pipeline Syst. Eng. Pract.* **5** (1), 04013011.

Sattar, A. 2013b Gene expression models for prediction of dam breach parameters. *J. Hydroinform.* doi:10.2166/hydro.2013.084.

Sharifi, S., Sterling, M. & Knight, D. W. 2011 Prediction of end-depth ratio in open channels using genetic programming. *J. Hydroinform.* **13** (1), 36–48.

Sivapragasam, C., Maheswaran, R. & Venkatesh, V. 2008 Genetic programming approach for flood routing in natural channels. *Hydrol. Process.* **22** (5), 623–628.

Subramanya, K. & Awasthy, S. C. 1972 Spatially varied flow over side-weirs. *J. Hydraul. Div. Proc.* **98** (HY1), 1–10.

Uyumaz, A. 1992 Side weir in triangular channel. *J. Irrig. Drain. Eng.* **118** (6), 965–970.

Uyumaz, A. 1997 Side weir in U-shaped channels. *J. Hydraul. Eng.* **123** (7), 639–646.

Uyumaz, A. 1982 Yan savaklardaki akImlO teorik ve deneysel incelenmesi (Theoretical and experimental investigations of flow over side weirs). PhD thesis, Istanbul Tech. Univ., Istanbul, Turkey.

Uyumaz, A. & Muslu, Y. 1985 Flow over side weirs in circular channels. *J. Hydraul. Eng.* **111** (1), 144–160.

Uyumaz, A. & Smith, R. H. 1991 Design procedure for flow over side weirs. *J. Irrig. Drain. Eng.* **117** (1), 79–90.

Vatankhah, A. R. 2012 Analytical solution for water surface profile along a side weir in a triangular channel. *Flow Meas. Instrum.* **23** (1), 76–79.

Vatankhah, A. R. 2013 Water surface profiles along a rectangular side weir in a U-shaped channel (Analytical Findings). *J. Hydrol. Eng.* **18** (5), 595–602.

Whigham, P. A. & Crapper, P. F. 2001 Modelling rainfall-runoff using genetic programming. *Math. Comput. Model.* **33** (6–7), 707–721.

Willmott, C. J. & Matsuura, K. 2005 Advantages of the mean absolute error (MAE) over the s (RMSE) in assessing average model performance. *Clim Res.* **30**, 79–82.

Yüksel, E. 2004 Effect of specific energy variation on lateral overflows. *Flow Meas. Instrum.* **15** (5–6), 259–269.

First received 13 October 2013; accepted in revised form 23 April 2014. Available online 22 May 2014