Flow Measurement and Instrumentation 65 (2019) 78–89

Contents lists available at ScienceDirect

Flow Measurement and Instrumentation

journal homepage: www.elsevier.com/locate/flowmeasinst

Artificial intelligence models for prediction of the aeration efficiency of the stepped weir

Ahmed A. Sattar[a], Mohamed Elhakeem[b], Mohammad Rezaie-Balf[c], Bahram Gharabaghi[d], Hossein Bonakdari[e,*]

[a] Department of Irrigation & Hydraulics, Faculty of Engineering, Cairo University, German University in Cairo, Giza, Egypt
[b] Civil Engineering Department, Abu Dhabi University, Abu Dhabi 59911, UAE
[c] Department of Civil Engineering, Graduate University of Advanced Technology-Kerman, Kerman, Iran
[d] School of Engineering, University of Guelph, Guelph, Ontario, Canada NIG 2W1
[e] School of Engineering, University of Guelph, Guelph, Ontario, Canada NIG 2W1

ARTICLE INFO

Keywords:
Stepped weir
Aeration efficiency
Model
Flow

ABSTRACT

Stepped weir is a commonly used hydraulic structure in water treatment plants to enhance the air-water transfer of oxygen or nitrogen and volatile organic components. The flow regimes on stepped weir are classified into nappe, transition and skimming flow. This study presents the novel application of artificial intelligence methods to evaluate the aeration efficiency over stepped weir for the three flow regimes. Two methods were adopted in this study, namely, the evolutionary polynomial regression (EPR) and the M5 model tree (M5 MT). A total of 151 laboratory experimental data sets were collected from the literature to train and test the artificial intelligence models. The Mallow's coefficient C_P was used to determine the effective variables affecting aeration efficiency. It was found that weir steps number, slope, the flow Reynolds number, and the ratio of the critical flow depth to the step height are the most important variables providing the lowest C_p. Both the EPR and M5 MT methods provided satisfactory predictions for the aeration efficiency. The two methods have high values of correlation coefficient $R > 0.93$ and low values for the root mean square error RMSE < 0.052 and relative mean absolute error RMAE < 0.065. However, the EPR method has an advantage over the M5 MT method that it provides one equation for each regime, while the M5 MT method provides a number of equations for each regime. This will make the equations of the EPR method more attractive to the practitioners compared to the equations of the M5 MT method. It was found that the equations obtained from artificial intelligence methods in this study perform better than the currently existing equations in the litrature obtained from regressive methods.

1. Introduction

The oxygen concentration in surface waters serves as an essential indicator of water quality for human use and to sustain aquatic life. It is also a significant factor in other applications, such as water treatment where it enhances the process of denitrification and volatile organic compounds can be removed. Several hydraulic structures can increase aeration of surface water, e.g., spillways, gated sills, weirs, stepped weir and ogee crests. However, stepped weir have been reported to be the most efficient hydraulic structure for aeration of surface water due to the intense turbulent mixing and large air bubble entrainment that they form, e.g., Toombes and Chanson [52], and Azimi et al. [1].

Abbreviations: ANFIS, neuro-fuzzy inference system; ANN, artificial neural networks; C_p, Mallows' coefficient; C_u, C_D, C_s, upstream, downstream and saturation dissolved oxygen concentrations; D, Diffusivity; E, transfer efficiency; E_{20}, transfer efficiency for 20 °C; EPR, evolutionary polynomial regression; EPR-MOGA, evolutionary polynomial regression- multi-objective genetic algorithm; F_r, Froude number; G, acceleration due to gravity; $f_{max}x_i$, maximum of the predicted output; $f_{min}x_i$, minimum of the predicted output; h_c, flow critical depth; H, total weir height; H, step height; K, explanatory variables; k and k', regression lines; l, weir step length; MT, model tree; N, weir number of steps; n, sample size; O, observed aeration efficiency; $\overline{O}$, mean of the observed values; P, predicted aeration efficiency; $\overline{P}$, mean of the predicted values; q, discharge per unit width; R, correlation coefficient; R_m, cross-validation coefficient; *RMAE*, relative mean absolute error; *RMSE*, root mean square error; RSS_p, residual sum of squares; R_e, Reynolds number; R_o^2, $R_o'^2$, squared correlation coefficients through the origin; SDR, standard deviation reduction; SVM, support vector machine; T, temperature; X1 × X2, input space by the M5 model tree; α, stepped weir slope; a_o, Bias; μ, Viscosity; ρ, density; σ, surface tension

* Corresponding author.
E-mail addresses: mohamed.elhakeem@adu.ac.ae (M. Elhakeem), bgharaba@uoguelph.ca (B. Gharabaghi), hbonakda@uoguelph.ca (H. Bonakdari).

https://doi.org/10.1016/j.flowmeasinst.2018.11.017
Received 22 October 2017; Received in revised form 29 October 2018; Accepted 18 November 2018
Available online 20 November 2018
0955-5986/ © 2018 Elsevier Ltd. All rights reserved.

The aeration efficiency of various hydraulic structures has been experimentally investigated by many researchers, including urban drainage structures (e.g., [57]; and [11]), drop manholes (e.g., [58,59]; and [56]), weirs (e.g., [25,54]). It was Gameson [20] who probably first reported potential of the aeration by stepped weir. Tebbutt et al. [50] presented experimental aeration data for stepped weir under the nappe and skimming flow regimes. Essery et al. [18] studied nappe flow in pooled stepped weir, while Novak [37] compared the work of Essery et al. [18] with a weir of pools and a single jet/pool configuration. Toombes and Chanson [51] studied the aeration on a stepped waterfall with flat chute slope and found that step length was more important than step height in increasing the oxygen concentration. They simulated nappe and skimming flow regimes on stepped weir and confirmed that the effective aeration occurs at nappe flow only. The experimental results have shown that the aeration efficiency on a stepped chute is a complex function of the step geometry (height, length) and inflow conditions (depth, velocity, initial level of air entrainment) and it is characterized by a high level of turbulence, which helps to distribute the solute and maintain maximum concentration gradient at the surface, while a high level of air entrainment increases the interface area available for mass transfer. Baylar et al. [5–7] studied the aeration efficiency and oxygen transfer on a large scale experimental setup over the three flow regimes, skimming, nappe and transition. They confirmed the effectiveness of aeration in case of nappe flow as compared to skimming flow.

Moulick et al. [35] have experimentally studied the optimum design criterion of a stepped weir to be used as a pre- or post- aeration unit in potable or wastewater treatment. They provided required optimum parameters for typical rectangular weir to maximize aeration efficiency. Khdhiri et al. [31] conducted experimental work to investigate the aeration performance of a stepped weir showing that aeration depends on operation condition and weir geometric parameters.

Based on the laboratory data, many researchers have developed regression models for the prediction of aeration efficiency as a function of flow and weir geometric parameters (e.g., [18,30,55]). As an alternative to traditional regression equations available in the literature, artificial intelligence has been utilized for modeling complex hydraulic problems. Many variants of the artificial intelligence data mining models have emerged, such as artificial neural networks (ANNs), adaptive neuro-fuzzy inference system (ANFIS), and support vector machine (SVM). These methods have been used to predict aeration efficiency under different flow conditions, e.g., [3,8,9,16,17,22,43]). The main drawback for the application of these methods is that the code itself is un-accessible for practitioners. In addition, they produce complex formulae that are less attractive to be used by the hydraulic engineers, and they don't directly take into account the physics of the analyzed phenomenon. Over the past few years, the evolutionary polynomial regression (EPR) and model tree (MT) techniques have been used to solve problems in hydraulic engineering. The MT model was used to predict the local scour depth at hydraulic structures under various waves and currents conditions (e.g., [10,60,61,62,63]; and [64]). On the other hand, the EPR technique was applied to evaluate the breakage rate in water distribution networks, sediment transport in pipes (e.g., [28]).

The primary objective of this study is to apply the evolutionary EPR and MT models to develop new regression equations for the prediction of the aeration efficiency of stepped weir. The proposed equations aims at covering some of the shortcomings in the currently existing equations, where some of them were developed from a limited number of laboratory data yielding to low correlation coefficient and relatively high error on their predictions. While others were too complicated with no clear relation between the input variables or developed for only one of the weir flow regimes. The proposed equations can serve as a basis for the design of weir according to required aeration efficiency and can substitute some of the previous complex models that practitioners avoid. The new equations are developed from laboratory data of a broad range of hydraulic conditions to cover the three flow regimes of the weir. For each flow regime, skimming, transition and nappe flow, an equation is developed for the aeration efficiency. A sensitivity analysis is performed to examine the predictions of equations and their predictions were further compared to the predictions of some of the existing equations in the literature.

2. Aeration and hydraulic regimes of stepped weir

Aeration or re-aeration is a mass transfer process between atmosphere and water. The mass transfer rate of a chemical across an interface in a quiescent fluid varies directly as the coefficient of molecular diffusion and the negative gradient of gas concentration. In stepped weir, aeration is caused from wave instabilities and turbulence fluctuations acting next to the air–water free-surface. Through this interface, there are continuous exchanges of both mass and momentum between water and atmosphere. Experimental has shown that the air–water flow behaves as a homogeneous mixture for void fraction C less than 90%. Typically, the water-air mixture is classified as bubbly flow for $C < 30\%$, spray for $C > 70\%$) and an intermediate for $30 < C < 70\%$ [14]. In turbulent water flows, air bubbles may be entrained when the turbulent kinetic energy is large enough to overcome both surface tension and gravity effects > This occurs when the instantaneous turbulent velocity normal to the flow direction is in the range of 0.1–0.3 m/s. The condition is nearly always achieved in stepped chute flows because of the high turbulence generated by the stepped invert (Fig. 1). Considering a given stepped weir, low flow rates give rise to a succession of free-falling nappes called nappe flow regime, and relatively little aeration is observed. With increasing flow rates, a transition flow regime occurs at intermediate flow rates. Dominant flow features included a chaotic appearance and strong splashing associated with irregular droplet ejections that may reach heights of up to 3–5 times the step height [13]. At larger flow rates, the waters skim over the pseudo-bottom formed by the step edges (skimming flow regime). Intense cavity recirculation is observed and the flow resistance is form drag predominantly [14]. In both transition and skimming flows, the upstream flow is non-aerated, but free-surface instabilities are observed. The location of the inception of free-surface aeration is clearly defined. Downstream the flow becomes rapidly aerated and free-surface aeration is very intense (Fig. 1).

Accordingly, there are three main flow regimes in a stepped weir, namely the skimming, the transition, and the nappe [12]. For high flow rates, the water flows down a stepped weir as a coherent stream "skimming" over the steps. The skimming flow consists of main flow with preferential direction imposed by the slope of the weir, secondary flows of large eddies formed between steps and biphasic flow due to the mixture of air and water. The external edges of the steps form a pseudo-bottom over which the flow skims. Beneath the pseudo-bottom, recirculating vortices develop (Fig. 1), and recirculation is maintained through the transmission of shear stress from the main stream [4].

Small-scale vorticity is also generated at the corner of the steps. The aerated flow region follows a region where the free-surface is smooth and glassy. Next to the boundary however, turbulence is generated and the boundary layer grows until the outer edge of the boundary layer reaches the surface. When the outer edge of the boundary layer reaches the free surface, the turbulence can initiate natural free surface aeration. The location of the start of air entrainment is called the point of inception. Downstream of the inception point of free-surface aeration, the flow becomes rapidly aerated and the free-surface appears white.

The transition regime is an intermediate flow condition between nappe flow which occurs at low flow rates and the skimming flow which occurs at high flow rates [38]. The transition flow is characterized by a chaotic flow behavior associated with rapid variations of the flow properties on each step and some instabilities for one discharge. Such unsteady fluctuations are associated with pressure fluctuations on the step faces and fluid-structure interactions. Further, strong

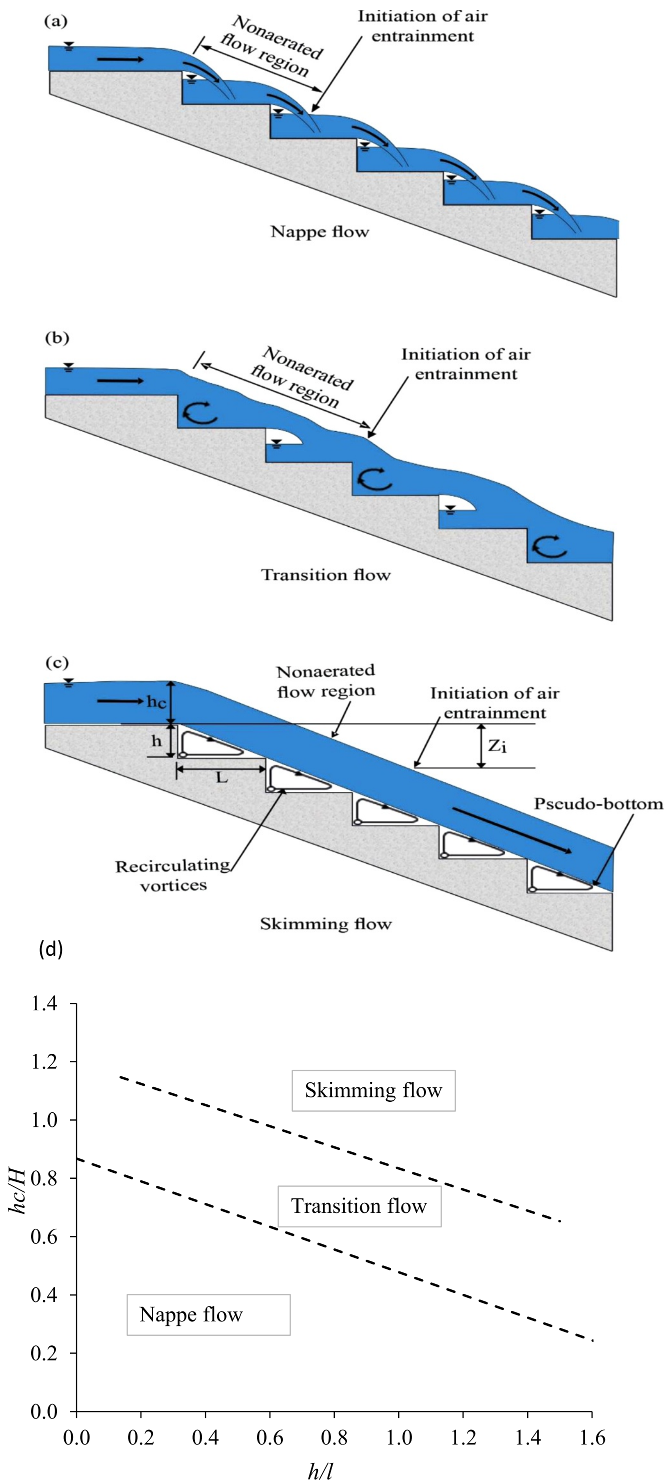

Fig. 1. Flow regimes above stepped chutes: (a) nappe flow; (b) transition flow; (c) skimming flow; (d) Range of the three flow regimes (after [13]).

stagnation occurs on the horizontal step faces associated with downstream splashing ([13]0. It is therefore strongly recommended to avoid the design of stepped spillways with transition flow regime unless a rigorous hydraulic and structural modeling study of the flow instabilities is conducted. Nonetheless, the transition flow is suitable for aeration weir used in water treatment because of the strong splashing and free-surface aeration.

The nappe flow occurs as a series of small consecutive falls impact onto the downstream of each step. At the upstream end of each step, the flow is characterized by a free-falling nappe, an air cavity and a pool of recirculating fluid. The nappe flow regime is classified according to the formation or absence of hydraulic jumps on the bed of the steps. In the nappe flow regime, the air is entrained at the jet, while de-aeration is often observed downstream. In the free-falling nappe, interfacial aeration takes place at both the upper and lower nappes. At the lower nappe, the developing shear layer is characterized by a high turbulence level of significant interfacial air entrainment [4].

Experiments have shown that the hydraulic flow regime above stepped weir depends on the step geometry and the discharge. Although each regime has its flow characteristics, the flow behavior is characterized by high turbulence increasing the potential for self-aeration.

Aeration is a physical process by which the oxygen transfers from the atmosphere to the surface water. The efficiency of this process can be enhanced by a stepped weir, which increase turbulent mixing of water. Experiments have shown that aeration efficiency is strongly affected by the type of the flow regime. Due to the different mechanisms of air entrainment in the nappe, transition and skimming flow regimes, the aeration efficiency of the three flow regimes differs significantly from each other. It was observed that the nappe flow condition leads to greater aeration efficiency compared to the other two flow regimes. This was attributed to that the nappe flow regime is associated with higher level of turbulence, larger residence time and substantial air bubble entrainment. According to Gulliver et al. [26], the aeration efficiency can be estimated as:

$$E = \frac{C_D - C_u}{C_s - C_u} \tag{1}$$

where C_u, C_D, C_s represent the upstream, downstream and saturation dissolved oxygen concentrations at the end of the stepped weir. The aeration efficiency varies between 0 for no oxygen transfer and 1 for the oxygen when it is fully transferred up to the saturation level of the water. No oxygen transfer implies zero efficiency. According to Gulliver et al. [26], the oxygen transfer depends on water temperature, viscosity (μ), density (ρ), surface tension (σ), and diffusivity (D). Gulliver et al. [26] expressed the transfer efficiency for any temperature or compounds as

$$\ln(1{-}E_i) = \ln(1{-}E)\left(\frac{D_i}{D}\right)^{1/2}\left(\frac{\mu}{\mu_i}\right)^{3/4}\left(\frac{\sigma}{\sigma_i}\right)^{3/5}\left(\frac{\rho_i}{\rho}\right)^{17/20} \tag{2}$$

The subscript i demonstrates the value of the parameter at the index temperature and for the index gas. It is essential to modify aeration efficiency to a reference temperature for evaluations of oxygen absorbency of the water at hydraulic structure. The aeration efficiency is normalized to a temperature of 20 °C to create a uniform basis for comparison of measurement results. The following equation was presented by Gulliver et al. [26] to describe the influence of temperature:

$$1{-}E_{20} = (1{-}E)^{1/f} \tag{3}$$

where E is transfer efficiency at actual water temperature, E_{20} is transfer efficiency for 20 °C and f is function of temperature T as $f = 1 + 2.1 \times 10^{-2}(T - 20) + 8.26 \times 10^{-5}(T - 20)^2$.

3. Regression equation for predicting aeration efficiency

Many laboratory studies have been carried out to measure the aeration efficiency in stepped weir. Experiments are usually conducted in prismatic rectangular channels with steps of various heights. Water flowing through the experiment is deoxygenated using sodium sulphite with the aid of cobalt chloride as a catalyst. Geometric dimensions of steeped weirs are know prior to running experiments, flow discharge is measured through an upstream weir, and the saturated dissolved oxygen concentration upstream and downstream the steps are measured using a portable oxygen meter 20 cm below the surface of water at any location. Using the above equations for each flow condition and geometric setup, the corresponding experimental aeration efficiency is calculated.

For various flow regimes, the aeration efficiency was estimated from simple regression equations fitted to laboratory data. One of the early equations for estimating aeration efficiency of stepped weir is due to Essery et al. [18] who performed experimental studies with various weir step heights under different flow conditions. They proposed the following regression equation:

$$E_{20} = 1 - \exp\left(-\frac{H}{\sqrt{gh}}\left(0.427 + 0.31\left(\frac{h_c}{h}\right)\right)\right) \tag{4}$$

where h_c is the flow critical depth, H is the total weir height, h is the step height, and g is the acceleration due to gravity. The weir water flow and the total height appeared as influential parameters in the above equation.

Based on laboratory experiments, the effect of the weir inclination angle (α) on aeration efficiency has been considered in Baylar et al. [2] equation as:

$$E_{20} = 1 - \left[1 + 2.161(\sin\alpha)^{0.290}\, 0.549^{(h_c/h)}\right]^{-1} \tag{5}$$

Moulick et al. [35] conducted laboratory experiment to investigate the effect of number of weir steps on aeration efficiency. They concluded that the aeration efficiency of the stepped weir increases with the increase of the flow depth up to a certain depth level after that it remains constant. They included the weir number of steps (N) in their aeration efficiency equation as:

$$E_{20} = 0.803 \times [N^2]^{[0.032 - 0.409 \times 10^{-3}/(h_c/h)]} - 0.505 \times \left(\frac{h_c}{h}\right) \tag{6}$$

Khdhiri et al. [31] performed an experimental study to investigate aeration in stepped weir for only the nappe and transition flow regimes emphasizing the role of the Reynolds number, R_e. They developed a single equation for the three flow regimes using their laboratory data and data from other sources. They expressed their aeration efficiency equation as:

$$E_{20} = 0.211 R_e^{-0.033} N^{0.445} (\tan\alpha)^{0.083} \tag{7}$$

4. Predictive data mining approaches for model development

Two data-mining techniques, namely the evolutionary polynomial regression, EPR, and the M5 model tree were utilized in this study to develop a model for the prediction of the aeration efficiency of stepped weir. The specific flow diagram of the methodology adopted for this study is illustrated in Fig. 2.

4.1. Evolutionary polynomial regression

The EPR can be defined as a non-linear global stepwise regression that provides symbolic formulas of models. Differently, from the original stepwise regression of Draper and Smith (1998), EPR is non-linear because the relationships between variables may results in non-linear functions although they are linear on regression parameters. It is global since the search for the optimal model structure is based on the exploration of the entire space of models by leveraging a flexible coding of the candidate mathematical expressions. The expressions achievable by EPR are made of some additive terms multiplied by many polynomials as shown in the following general expression:

$$\hat{\mathbf{Y}} = a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right) \tag{8}$$

Where m is the maximum number of additive terms, $\mathbf{X}_i$ and $\hat{\mathbf{Y}}$ are model input and output variables, function f is chosen by the user and exponents of variables, i.e., **ES**(j, i)) are selected from a set **EX** of

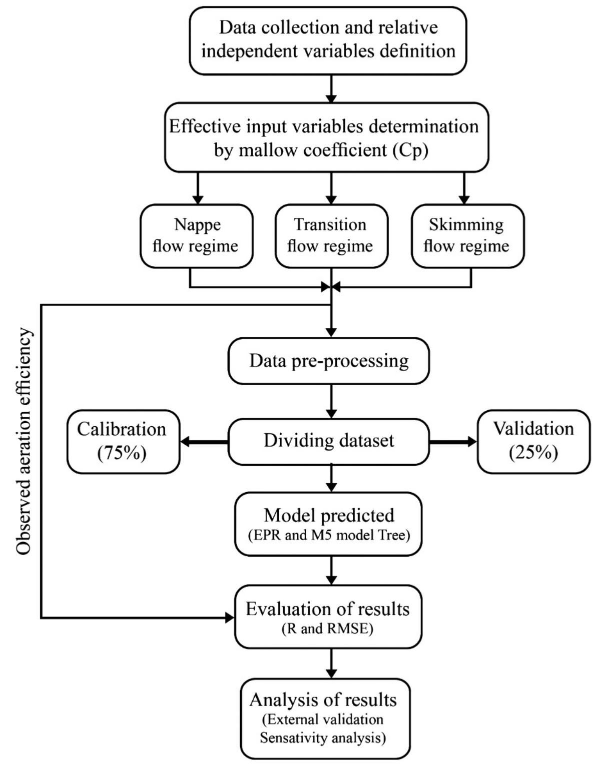

Fig. 2. The workmanship of proposed models for prediction of aeration efficiency.

candidates defined by the user, see [24] for details. The genetic algorithm is used to choose the exponents **ES**(j,i) from among the values in set **EX**. This means that an integer coding of possible alternative exponents **ES**(j,i) is adopted to achieve non-linear relationships. It is worth noting that, if the set of exponents contains zero and **ES**(j,i) = 0, the relevant input disappears from the final expression. Thus, although simple, structures like Eq. (8) are quite versatile and flexible to reproduce patterns in data.

A key point of the EPR model development strategy is that final expression are linear on coefficient a_j so that they are estimated using classical numerical regression (e.g., least squares). The parameter estimation is solved as a linear inverse problem to guarantee a unique relationship between each model structure and its parameters [24]. Regarding numerical regression strategy, EPR may produce a non-linear mapping of data (like that achievable by Artificial Neural Networks, [27]) although with few constants to estimate, and using linear regression for parameters estimation. These features, in turn, help in avoiding over-fitting to training data especially when the dataset is not large.

The most significant upgrade of the original EPR paradigm encompasses the multi-objective optimization strategy (i.e., EPR-MOGA) where the accuracy of data reproduction and parsimony of model structures are simultaneously maximized [23]. The parsimony refers to the number of variables and (or) additive terms of the mathematical expressions. Maximizing the parsimony of resulting formulas is aimed at facilitating the physical meaning of final expressions and, in turn, achieving a general description of the underlying phenomenon.

The search space in EPR-MOGA is defined by the user in terms of the base structure of mathematical expressions (e.g., as in Eq. (8) and type of function f), the maximum number of additive terms m, the cardinality of set **EX** of candidate exponents and number of candidate explanatory variables (i.e., k). The model search is performed by using the optimized Multi-Objective Genetic Algorithm (OPTIMOGA - Laucelli and Giustolisi, 2011) that is based on the Pareto dominance criterion [65] to accomplish multi-objective optimization.

EPR-MOGA explores the space of m-term formulas using two or three from the following objectives: (a) the maximization of model accuracy, (b) the minimization of the number of model coefficients (i.e., number of additive terms) and (c) the minimization of the number of actually used model inputs (i.e., whose exponent is not 0 in the resulting model structure). Note that the last two objectives represent as many measures of model parsimony. The EPR-MOGA finally obtains a set of optimal solutions (i.e., the Pareto front) that can be considered as trade-offs between structural complexity and accuracy [41].

4.2. M5 model tree

Among the data mining techniques, methods are used to solve the problem by dividing it into several sub-problems (sub-domains), and the result is a combination of these sub-problems. Classification trees classify data records by sorting them down the tree from the root node to some leaf nodes. That is, the initial tree firstly is built through a recursive splitting process to maximize the expected error reduction. The splitting criterion is based on the standard deviation of the subset in the training data that reaches a particular node. The splitting process terminates when: 1) the difference is relatively small among the output of the samples that reach a node, and 2) when the number of samples in a subset is less than a certain value. This is assisted by the speed of the Alpha-Beta pruning optimization technique that reduces the computational time by a huge factor allowing searching faster and going deeper in levels of model tree. The Alpha-Beta pruning is called so since it passes 2 extra parameters in the minimax function namely alpha and beta. It removes all nodes in the model tree that are not possibly affecting final decision. The linear regression models are then built for each subset of samples in the terminating (leaf) nodes. The difference between the better-known classification trees and the MT technique is that the later has a numeric value rather than a class label in connection with the leaves. The MT splits the entire input or parameter domain into sub-domains, and a linear multivariable regression model is applied for each of them [40]. In this way, MT models can be applied to solve continuous class problems and obtain a structural representation of the data sets using the piecewise linear models to approximate nonlinear relationships. Furthermore, this algorithm was known as one of the most effective approaches to present a meaningfully physical insight of the phenomenon. The tree-building procedure within four linear regression models and knowledge extraction from the structure for corresponding sub-domains were illustrated in Fig. 3a. Furthermore, a general tree structure of the MT approach was sketched in Fig. 3b. Based on the domain-splitting criterion, various approaches such as M5 model was frequently utilized to generalize the MT technique [53].

Through the MT approach, the basic tree is firstly generated using the splitting criterion of the standard deviation reduction (SDR) factor:

$$SDR = sd(E) - \sum_i \frac{|E_i|}{|E|} sd(E_i) \tag{9}$$

In which E, sd, and E_i are the set of examples (data records) that reach the node, the set that results from splitting the node according to the chosen attribute (parameter), and standard deviation, respectively. The M5 utilizes the sd parameter as an error measure of the class values that reach a node. Testing all parameters at a node, it calculates the expected reduction in error and then selects the parameter that maximizes SDR. This process stops when the standard deviation reduction becomes less than a certain percent of the standard deviation of the original dataset or when only a few data records remain [39,53]. Then, a linear regression model is developed for each sub-domain. Only the data in connection with the variables tested in that sub-domain are used in the regression.

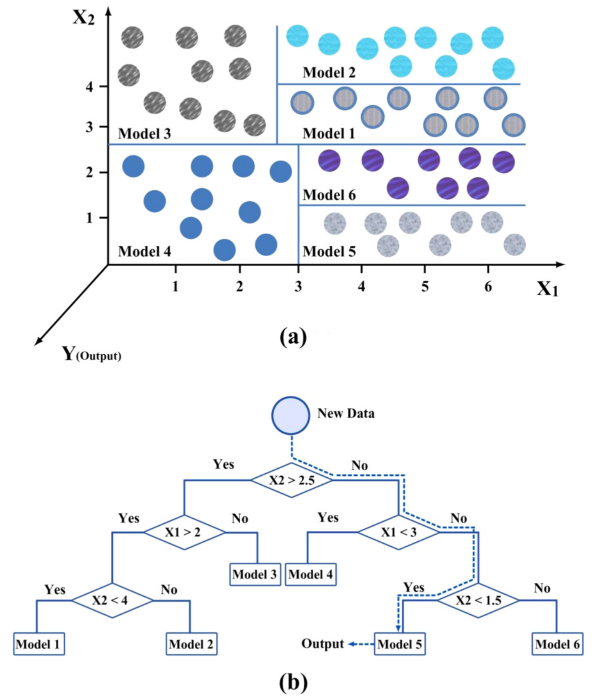

Fig. 3. Splitting the input space and prediction by the model tree for a new data record [36]: (a) splitting of the input space (X1 × X2) by the M5 model tree algorithm; (b) predicting a new data record by the model tree.

4.3. Variable selection

In practical problems, many independent variables exist, and the choice of the most influencing variables is necessary for the development of descriptive models. The predictive ability of the developed models depends largely on the chosen variables subset [21]. The problem of subset selection arises, for example, when a researcher seeks a predictive model that cross-validates well, or when there is redundancy amongst the variables leading to multicollinearity [19]. Collinearity is the case when one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. When more than two predictor variables are associated, this leads to a substantial decrease in statistical significance of model and this is termed multicollinearity. Unlike the collinearity, the multicollinearity may not be possible to predict before observing its effect on model since some of the predictor variables may have low degree of association. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only increases model complexity and affects calculations regarding individual predictors. Consequently, a model that is simple yet contains useful information is always desirable.

There are several approaches to predictor influencing subset, including forward selection, backward elimination, Akaike Information Criterion, Bayesian Information Criteria, and Mallow's coefficient, C_p [15]. The Mallows' coefficient is used herein for the choice of the minimum number of variables. The Mallow's coefficient displays a tradeoff between the regression equation precision and its complexity by seeking for an equation that has the best fit to the data but fewer variables. It compromises between the goodness of fit of an equation and the number of variables incorporated in the equation, by imposing a penalty for increasing the number of variables.

The Mallow's coefficient has an inversely oriented score, where lower values indicate the preferred regression equation that has the fewest number of variables but still provides an adequate fit to the data.

If p predictors are to be selected from a set of k available predictors for a specific problem ($k > p$), the Mallows' coefficient C_p for this particular subset can be calculated as [34]:

$$C_p = \frac{RSS_p}{MSE_k} + 2p - n \tag{10}$$

where RSSp is the residual sum of squares for the model with p variables, MSE_k is the mean squared error for regression of the model with the complete set of k variables, and n is the sample size. As such, C_P indexes the mean squared error of prediction for 'subset' models relative to the full model with a penalty for the inclusion of unimportant variables [32].

After selecting the important variables, a sensitivity analysis to assign the importance of each input variable on the aeration efficiency. The analysis was conducted by eliminating one input variable each time and evaluating the effect of that variable on the model prediction of E_{20}. The sensitivity value in percentage (%) of the dependent variable to each independent variable is calculated using the following relationships:

$$N_i = f_{max}(x_i) - f_{min}(x_i) \tag{11}$$

$$S_i = \frac{N_i}{\sum_{j=1}^{n} N_j} \times 100 \tag{12}$$

where $f_{max}(x_i)$ is the maximum of the predicted output and $f_{min}(x_i)$ is the minimum of the predicted output over the i-th input domain, where other variables are equal to their mean values.

4.4. Models performance evaluation and validation

The performance and validity of the developed models were evaluated through statistical measures of the models mean errors and graphical representation of their predictions against the laboratory data. The correlation coefficient (R) and two mean error criteria are used to examine the models accuracy, namely the root mean square error (*RMSE*) and the relative mean absolute error (*RMAE*) expressed as:

$$R = \frac{\sum_{i=1}^{n}(P_i - \bar{P}).(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2 \sum_{i=1}^{n}(O_i - \bar{O})^2}} \tag{13}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \tag{14}$$

$$RAE = \frac{\sum_{j=1}^{n}|O_{(ij)} - P_j|}{\sum_{j=1}^{n}|P_j - \bar{P}|} \tag{15}$$

where P is the predicted aeration efficiency, and O is the observed aeration efficiency value, $\bar{O}$ is the mean of the observed values, and $\bar{P}$ represents the mean of the predicted values, n is the total number of dataset samples. The R measures how well considered independent variables account for the measured dependent variable. The RMSE and RMAE are used to measure model prediction accuracy. The *RMSE* is commonly used in the error analysis to quantify the absolute errors in the model predictions however; it is biased towards the observations with larger values. The *RMAE* eliminates the bias towards the observations with high values by giving equal weight to all values over the observed range. Therefore, it is important to consider the *RMAE* in the model predictions as well, when the observations span several orders of magnitude as in the case herein. The *RMSE* and *RMAE* would be zero for perfect agreement between the observed and predicted values and increase as the difference increases. The low value of RMSE and the high value of R are expected in case the models well represent the laboratory data.

Reliable prediction models require that at least one of the regression lines of predicted versus observed values (k or k') are close to unity [43,45]. Herein, k and k' defined as

$$k = \sum_{i=1}^{n}(P_i \times O_i)/O_i^2 \tag{16}$$

$$k' = \sum_{i=1}^{n}(P_i \times O_i)/P_i^2 \tag{17}$$

Additionally, the coefficient of determination for the regression lines through the origin m' and n' should be less than 0.1, where m and n' defined as [47–49]:

$$m' = (R^2 - R_O^2)/R^2, \quad n' = (R^2 - R_O'^2)/R^2 \tag{18}$$

Moreover, the cross-validation coefficient R_m should satisfy ([46, and 49]):

$$R_m = R^2 \times (1 - \sqrt{|R^2 - R_O^2|}) > 0.5 \tag{19}$$

where R^2 is the squared correlation coefficient as given in Eq. (11), and R_o^2 and R'^2_o are the squared correlation coefficients through the origin between the predicted and observed values and between the observed and predicted values, respectively, defined as: $R_O^2 = 1 - \sum_{i=1}^{n} O_i^2(1-k)^2/\sum_{i=1}^{n}(O_i - \bar{O})^2$ and $R_O'^2 = 1 - \sum_{i=1}^{n} P_i^2(1-k')^2/\sum_{i=1}^{n}(P_i - \bar{P})^2$.

5. Results and discussions

5.1. Aeration data set

To develop the models, a total of 151 experimental measurements of aeration efficiency over stepped weir were collected from Baylar and Emiroglu [2], Baylar et al. [7], and Moulick et al. [35]. The chosen data provide the aeration efficiency over the three possible flow regimes over stepped weir. The data were divided into two subsets: approximately 75% of the datasets were used for the training phase (calibration) to train network of proposed models over a number of cases while the remainder data (25% of the data) were kept unseen to MT/EPR during the models development process for testing purposes (validation). Data partitioning was randomly made for each flow regime such that the statistical measures of both datasets are similar. Four-cross fold validation was utilized herein.

Table 1 illustrates the collected aeration data and their statistical measures. The collected data include the critical depth (hc), the flow Reynolds number (R_e), stepped weir slope (α), stepped weir number of steps (N), and the total weir height (H). The stepped weir had a step height that varies from 5 cm to 15 cm, total height that varies from 1 to 3 m and a slope that varies from 14° to 50°. The Froude number varied from 0.1 to 1 in the nappe flow regime and reached a value of 9.5 in the skimming flow regime. The calculated critical flow depth for all three flow regimes was similar and ranged from 0.05 m to 0.15 m.

The variables affecting aeration efficiency are given in the following functional relationship:

Table 1
Descriptive Statistics for prediction of aeration efficiency.

Parameters	N	R_e	h_c(m)	H(m)	α(deg)	$E20$
Nappe flow regime						
Maximum	50	181387	0.1	3	50	0.902
Minimum	6	1963	0.0046	1.2	14.48	0.43
mean	13.309	60779	0.041	2.166	26.3463	0.712
Standard Deviation	7.0961	48011	0.0243	0.7845	10.9487	0.141
Transition flow regime						
Maximum	50	285891	0.1414	2.5	50	0.79
Minimum	8	46180	0.0304	1.2	14.48	0.39
mean	19	169363	0.0872	1.8129	30.233	0.607
Standard Deviation	10.99	72403	0.0333	0.626	13.8103	0.1411
Skimming flow regime						
Maximum	50	263308	0.14	2.5	50	0.82
Minimum	12	74601	0.0483	1.2	14.48	0.16
mean	31.094	194895	0.10755	1.966	31.3762	0.5464
Standard Deviation	14.267	61265	0.0301	0.618	12.647	0.1801

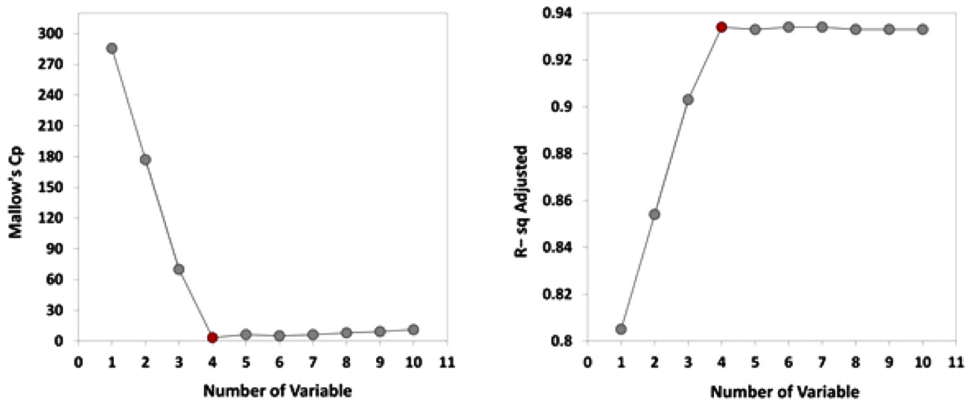

Fig. 4. Number of variable against the criteria where the red point corresponds the number of variable to achieve the criteria, i.e. minimum for Cp and maximum for R^2 adjusted.

Table 2
Values of Mallows' Cp results from Best Subsets Regression analysis.

Model No.	R^2	Cp	S	Variables									
				q(m^2/s)	N	h(m)	hc(m)	Re	h/l	hc/h	hc/H	α(deg)	Fr
1	0.806	285.6	0.076306								X		
2	0.856	176.9	0.066044									X	X
3	0.905	69.9	0.053893		X					X		X	
4	**0.936**	**3.3**	**0.044363**		**X**			**X**			**X**	**X**	
5	0.935	6.3	0.044825		X	X		X		X	X		
6	0.936	5.1	0.044491		X	X	X	X		X	X		
7	0.937	6.2	0.0445		X	X	X		X	X	X		X
8	0.937	7.9	0.044607		X	X	X	X	X	X	X		X
9	0.937	9.2	0.044661		X	X	X	X	X	X	X	X	X
10	0.937	11	0.044782	X	X	X	X	X	X	X	X	X	X

Table 3
Sensitivity analysis S_i (%) for independent variables.

Model	Flow condition	N	R_e	$\frac{h_c}{H}$	α
MT	Nappe flow regime	22.85	**26.79**	26.35	23.99
	Transition flow regime	23.26	**27.85**	24.26	24.60
	Skimming flow regime	24.11	**30.16**	21.93	20.78
EPR	Nappe flow regime	21.94	**27.35**	26.40	24.30
	Transition flow regime	22.86	**28.27**	25.62	23.25
	Skimming flow regime	23.46	**27.39**	26.97	22.16

$$E20 = f\left(q,\, H,\, h_{c,} \frac{h_c}{H},\, R_e,\, N,\, F_r,\alpha, \frac{\mathrm{h}}{l}\right) \tag{20}$$

Where E20 is the aeration efficiency corrected to a temperature of 20 degrees, q is the discharge per unit width, and N is the number of weir steps.

The best subset is selected from Eq. (20) using the Mallows' coefficient C_P and adjusted squared correlation coefficient R^2 where illustrated in Fig. 4. Fig. 4 shows that 4 variables were sufficient to minimize the Mallows' coefficient C_p and maximize the adjusted R^2. Table 2 provides the values of R^2 and C_p obtained by each subset of variables considering the aeration efficiency as the model output. From Table 2, it can be seen that N, R_e, α, and h_c/H are the most important variables providing the lowest C_p. Therefore, this subset of variables has been used to develop the aeration efficiency models. While other subsets may have slightly higher R^2 values, more weight was given to the C_p value in selecting the representative subset. The $C_{p,}$ S and adjusted R^2 values for the selected subset were 3.3, 0.044 and 0.93.6, respectively. The sensitivity analysis has further shown that $\frac{h_c}{H}$ and the Re have the highest influence on the predicted values of E20 in nappe, transition and

Table 4
The tree structure of the MT technique for nappe, transition, and skimming flow regimes.

<table>
<tr><th>Rules of MT approach</th><th>LM</th><th>LM No.</th></tr>
<tr><td>Nappe flow regime
h_c/H < = 0.022</td><td>$E20 = 1.108(N)^{0.0271}(R_e)^{-0.0388}(h_c/H)^{0.0144}(\alpha)^{0.028}$</td><td>(1)</td></tr>
<tr><td>| α < = 28: LM1</td><td>$E20 = 1.107(N)^{0.002}(R_e)^{-0.0284}(h_c/H)^{0.0095}(\alpha)^{0.004}$</td><td>(2)</td></tr>
<tr><td>| α > 28
| | a < = 39.72
| | | N < = 11.32: LM2</td><td>$E20 = 1.12(N)^{0.005}(R_e)^{-0.0296}(h_c/H)^{0.0095}(\alpha)^{0.004}$</td><td>(3)</td></tr>
<tr><td>| | | N > 11.32: LM3</td><td></td><td></td></tr>
<tr><td>| | α > 39.72: LM4</td><td>$E20 = 1.163(N)^{0.004}(R_e)^{-0.032}(h_c/H)^{0.0095}(\alpha)^{-0.004}$</td><td>(4)</td></tr>
<tr><td>h_c/H > 0.022
| α < = 26: LM5</td><td>$E20 = 0.75(N)^{0.133}(R_e)^{-0.1026}(\alpha)^{0.1908}$</td><td>(5)</td></tr>
<tr><td>| α > 26: LM6</td><td>$E20 = 0.474(N)^{0.134}(R_e)^{-0.0709}(\alpha)^{0.2457}$</td><td>(6)</td></tr>
<tr><td>Transition flow regime</td><td>$E20 = 0.05(N)^{0.34}(R_e)^{0.096}(\alpha)^{0.1027}$</td><td>(1)</td></tr>
<tr><td>α < = 26</td><td>$E20 = 0.08(N)^{0.33}(R_e)^{0.055}(\alpha)^{0.1027}$</td><td>(2)</td></tr>
<tr><td>| Re < = 171395.7: LM1</td><td></td><td></td></tr>
<tr><td>| Re > 171395.7: LM2</td><td>$E20 = 0.056(N)^{0.265}(R_e)^{0.11}(\alpha)^{0.116}$</td><td>(3)</td></tr>
<tr><td>α > 26: LM3</td><td></td><td></td></tr>
<tr><td>Skimming flow regime</td><td>$E20 = 0.097(N)^{-0.13}(R_e)^{0.06}(h_c/H)^{-0.38}(\alpha)^{0.13}$</td><td>(1)</td></tr>
<tr><td>hc/H < = 0.06</td><td></td><td></td></tr>
<tr><td>| hc/H < = 0.044: LM1</td><td>$E20 = 0.038(N)^{-0.162}(R_e)^{0.13}(h_c/H)^{-0.38}(\alpha)^{0.17}$</td><td>(2)</td></tr>
<tr><td>| hc/H > 0.044</td><td></td><td></td></tr>
<tr><td>| | N < = 20: LM2</td><td>$E20 = 0.0376(N)^{-0.162}(R_e)^{0.123}(h_c/H)^{-0.38}(\alpha)^{0.197}$</td><td>(3)</td></tr>
<tr><td>| | N > 20: LM3</td><td></td><td></td></tr>
<tr><td>hc/H > 0.06: LM4</td><td>$E20 = 4.11(N)^{-0.46}(R_e)^{-0.237}(h_c/H)^{-0.59}(\alpha)^{0.097}$</td><td>(4)</td></tr>
</table>

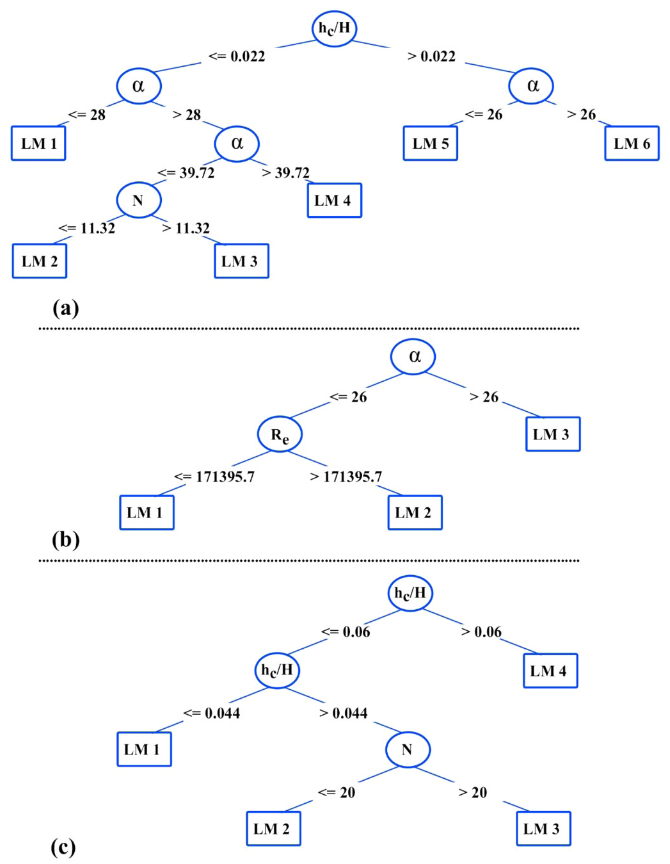

Fig. 5. Proposed MT structure for prediction of the aeration efficiency under different flow regimes: (a) nappe flow; (b) transition flow; (c) skimming flow.

skimming flow regimes, respectively. Table 3 summarizes sensitivity analysis results for each flow regime.

5.2. Aeration efficiency prediction from the EPR method

During the development process in the EPR approach, some limitations can be defined to control the output models regarding the type of functions used, the number of generations, the range of exponents, the number of terms, etc. In this process, there is a potential to achieve different models for a particular problem, which allows the user to obtain supplementary information for different conditions [42]. The EPR method was applied for nappe, transition, and skimming flow regimes. The input variables for model development were the number of weir steps, the flow Reynolds number, the slope of the stepped weir, and the ratio of the critical flow depth to the step height. The model output was the aeration efficiency over stepped weir.

In the configuration of the EPR, the exponent with values in the range between -2 and 2, spanned with a 0.5 long step, was used to improve the fitting performance of the EPR procedure. The maximum number of polynomial terms is set to $m = 3$, without assuming a bias a_O, and admitting only positive coefficients, i.e., $a_j > 0$. The optimization strategy of the objective functions was: (i) to the maximize the model accuracy, (ii) to minimize the number of model coefficients (i.e., whose exponent is not 0 in the resulting model structure). Different runs were performed using various variables and forms to determine the most robust and practical relationships for nappe, transition, and skimming flow regimes. Ultimately, among several models returned by EPR-MOGA-XL, the following models have been selected for the three regimes, as a trade-off between accuracy (on the training set) and parsimony. For the nappe flow regime, the aeration efficiency prediction model can be written as:

$$E20 = 0.000262 R_e^{0.5} N^{-0.5}\left(\frac{h_c}{H}\right)^{-1}\alpha^{-0.5}\exp\left(-2\frac{h_c}{H}\right) + 0.0003 R_e^{0.5}\left(\frac{h_c}{H}\right)^{-0.5}\exp\left(-2\frac{h_c}{H}\right) + 0.025 N\alpha^{-0.5}\exp\left(-2\frac{h_c}{H}\right) \tag{21}$$

And for the transition flow regime, the aeration efficiency prediction model can be written as:

$$E20 = 0.000052\, R_e^{-0.5} N^{-0.5}\left(\frac{h_c}{H}\right)\alpha^{-0.5}\exp\left(-2\frac{h_c}{H}\right) + 71191.536\, R_e^{-1}\left(\frac{h_c}{H}\right)^{0.5}\exp\left(-2\frac{h_c}{H}\right) + 0.00028\, R_e^{0.5}\left(\frac{h_c}{H}\right)^{-0.5}\exp\left(-2\frac{h_c}{H}\right) \tag{22}$$

And for the skimming flow regime, the aeration efficiency prediction model can be written as:

$$E20 = 0.0063\, R_e^{0.5} N^{-1.5}\left(\frac{h_c}{H}\right)^{-2}\alpha^{-1}\exp\left(-2\frac{h_c}{H}\right) + 0.00000043\, R_e^{0.5}\alpha^{2}N^{-0.5}\exp\left(-2\frac{h_c}{H}\right) + 0.0000056\, R_e^{0.5}\alpha^{0.5}\left(\frac{h_c}{H}\right)^{-0.5} N^{0.5}\exp\left(-2\frac{h_c}{H}\right) \tag{23}$$

All calculations have been implemented using the package (EPR-MOGA-XL), working in the MS-Excel environment [33].

5.3. Aeration efficiency prediction from the M5 model tree

The M5 model tree procedure for prediction of aeration efficiency was built using machine learning software, Weka V3.6. The same predictors as calculated from Mallow's coefficient are used in M5 model tree development.

The M5 model tree provides linear equations linking independent and dependent variables; to overcome this constraint, we used the logarithmic forms of the variables to produce a nonlinear equation for our model. Therefore, a nonlinear prediction was achieved in the form of output $E20 = \gamma X_1^{\varphi_1} X_2^{\varphi_2} \ldots X_N^{\varphi_N}$.

In order to reduce the final size of the decision tree, and to improve the predictive accuracy by reduction of overfitting, the alpha-beta pruning algorithm is employed in model development. However, the generation of small discontinuity between adjacent linear models was unavoidable [29]. Therefore, a pruning factor of 4.0 and smoothing option was utilized. After classifying, M5 MT technique provides 6, 3, and 4 nonlinear relationships for the nappe, transition, and skimming flow regimes, respectively. The developed equations based on conditional sentences for each flow regime are presented in Table 4.

In Fig. 5, the classification of data set by using an M5 algorithm for nappe, transition, and skimming flow regimes are shown. From the diagram, α is the most important variable as a principal splitting attribute for prediction of E20 and the corresponding value is obtained 26.00 for transition and skimming flow regimes. In addition, N with a value of 11 was considered as splitting variable for production of 2nd and 3rd nonlinear models for nappe flow condition.

5.4. Comparison of the EPR and the M5 models

The perditions of the ERP and M5 MT methods were evaluated graphically followed by statistical analysis of their models errors. Fig. 6 shows scatter plot comparisons of the two methods for three flow

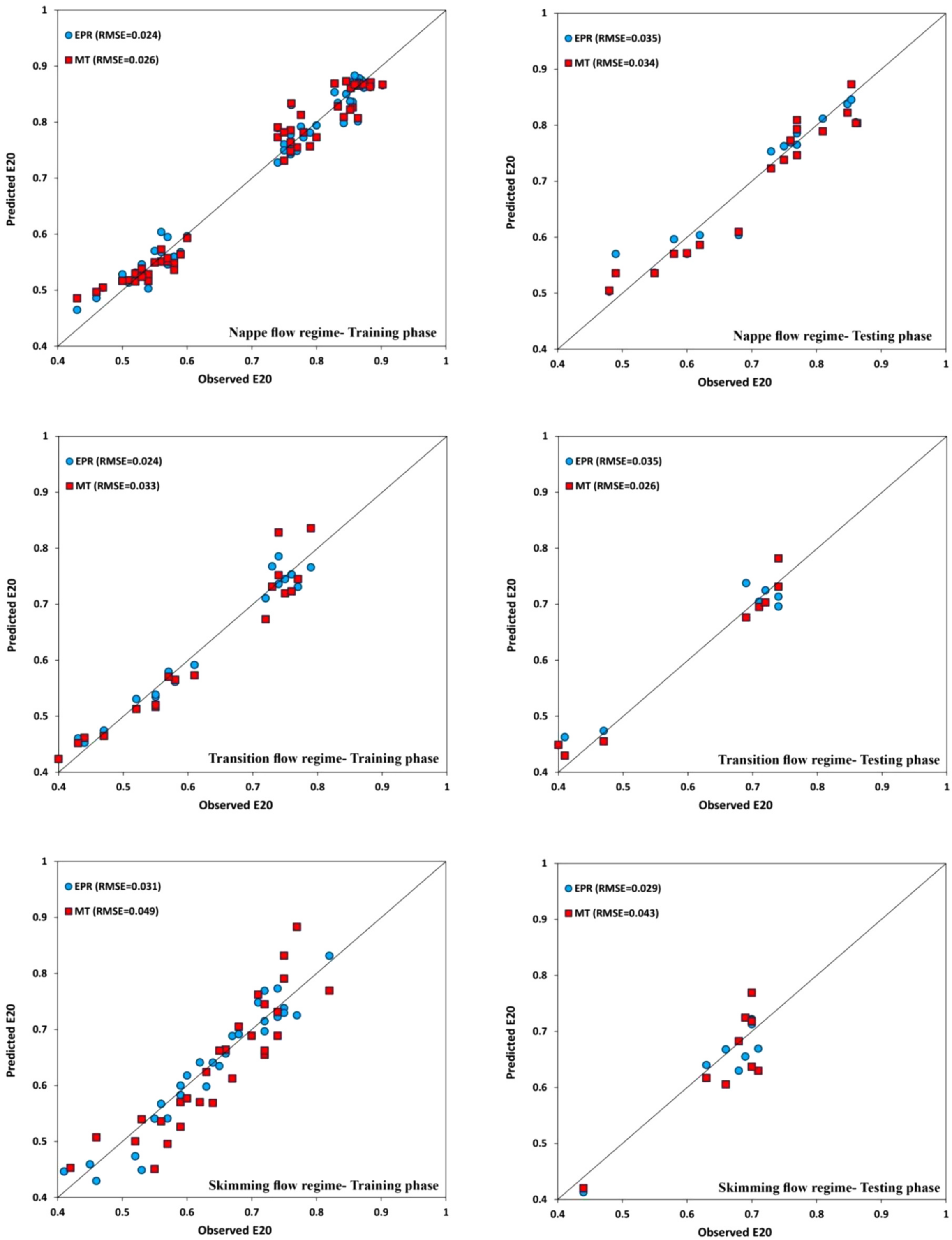

Fig. 6. Scatter plots of observed and predicted E20 in training and testing phases of the proposed models.

Table 5
External validation statistical measures for developed MT/EPR models.

Models	R (R > 0.8)	k (0.85 < k < 1.15)	k' (0.85 < k' < 1.15)	m (m < 0.1)	n (n < 0.1)
EPR	0.95	0.92	0.88	− 0.11	− 0.10
MT	0.97	1.11	0.72	− 0.07	− 0.03

Table 6
Results of Performances for proposed models for different flow regimes.

Methods	R^2	RMSE	RAE
Nappe flow regime			
EPR-Training	0.94	0.038	0.0399
EPR-Testing	0.88	0.052	0.0481
MT-Training	0.98	0.040	0.0649
MT-Testing	0.93	0.039	0.0135
Essery et al. [18]	0.25	0.184	0.5894
Baylar et al. [2]	0.56	0.148	0.5667
Moulick et al. [35]	0.26	0.141	0.1191
Khdhiri et al. [31]	0.15	0.277	0.4804
Transition flow regime			
EPR-Training	0.96	0.029	0.0211
EPR-Testing	0.98	0.029	0.0084
MT-Training	0.93	0.035	0.0447
MT-Testing	0.97	0.032	0.0138
Essery et al. [18]	0.93	0.055	0.0287
Baylar et al. [2]	0.92	0.122	0.296
Khdhiri et al. [31]	0.85	0.181	0.2463
Skimming flow regime			
EPR-Training	0.98	0.032	0.0162
EPR-Testing	0.94	0.041	0.0115
MT-Training	0.96	0.041	0.0216
MT-Testing	0.95	0.039	0.0111
Essery et al. [18]	0.17	0.384	0.3866
Baylar et al. [2]	0.56	0.162	0.9596
Khdhiri et al. [31]	0.52	0.155	0.3222

regimes in the training and testing phases. Experimental data for each flow regime has been obtained as identified by the relevant researcher and according to Fig. 1d. In general, these graphs highlight the good alignment of the data points along the graph bisector, affirming the good agreement between the models predictions and the observed values in both the training and testing phases. In addition, graphical representation of the models prediction requires that the gradients of the regression line (graph bisector) satisfies some or all of the following validation criteria: k and k' are close to 1, m and n < 0.1, and Rm $>$ 0.5 to be considered valid for prediction (Pan et al., 1996). This was true for the models developed by the two methods as they satisfy all of the related validation criteria, which indicate their strong prediction ability. Table 5 summarizes the statistical validation measures for the proposed models.

The performance of the models prediction was further evaluated using the statistical measures of R^2, RMSE and RMAE as summarized in Table 6. It can be seen from Table 6 that both the EPR and the M5 MT methods provide high values of $R^2 > 0.865$ with negligible difference between the two methods, which confirm their reliability. The EPR method in general provides slightly higher values for R^2 in the training phase while the M5 MT provides slightly higher values for R^2 in the testing phase. In addition, both methods provide low values for the RMSE < 0.052 and RMAE < 0.065. The EPR method in general provides slightly lower error values in the training phase while the M5 MT provides slightly lower values of errors in the testing phase. Thus, the two methods provide satisfactory predictions for the aeration efficiency. However, it should pointed out here that the EPR method has an advantage over the M5 MT method that it provides one equation for each regime, while the M5 MT method provides a number of equations for each regime. This will make the equations of the EPR method more attractive to the practitioners compared to the equations of the M5 MT method.

5.5. Comparison with available models

The predictions of the E20 aeration equations proposed in this study were compared tothe predictions of the equations proposed by Essery et al. [18], Baylar et al. [2], Moulick et al. [35], and Khdhiri et al. [31] for the three flow regimes in a stepped weir. Table 7 shows the ranges of experimental measurements, which were used in development of these classical equations for the prediction of aeration efficiency. It is observed from this table that the range of experimental data used for model development for Baylar et al. [2] and Khdhir et al. [31] were the widest range spanning three flow regimes similar to data used in this study. On the other hand, data range for Essery et al. [18] and Moulick et al. [35] was narrower since they only studied Nappe flow regime. Thus it is expected that the equations developed by both Essery et al. [18] and Moulick et al. [35] would yield lowest scores and highest error.

As shown in Table 4, for the three flow regimes the equations developed by the EPR and M5 MT methods have higher R^2 values, and lower RMSE and RMAE values when compared with the other examined equations for the prediction of E20. Thus, the proposed equations show improvement in the prediction of aeration efficiency compared to the currently existing equations. Nonetheless, the examined equations provide satisfactory predictions as well in the transition flow regime with $R^2 >$ 0.86 and low mean errors. The equation of Baylar et al. [2] performs better than the other examined equations in the nappe and skimming flow regimes with R^2 of 0.56 since it was originally developed for predicting the aeration efficiency in the three flow regimes. On the other hand, the equation of Khdhiri et al. [31] showed the lowest R^2 value of 0.39 in the nappe flow regime despite being developed from experimental data for Nappe flow. This might be due to the absence of important variables describing the hydraulic characteristics of flow in equation form. However it has a value for R^2 of 0.52 in the skimming flow regime. While the equation of Essery et al. [18] gave the lowest R^2 value of 0.16 in skimming regime compared to the other examined equations. This might be attributed to the low range of flow used in their experiments.

Fig. 7 shows scatter plot comparisons of the examined equations for three flow regimes. In general, these graphs highlight the high scatter of the data points along the graph bisector, affirming the poor agreement between the models predictions and the observed values. The graphs show that for nappe flow regime all equations in general underestimate the predictions of E20. For the skimming flow regime, it can be noted that the equation by Baylar et al. [2] underestimtes the predictions of E20, while the equation by Essery et al. [18] overestimates the predictions of E20. In the skimming flow regime, the equation by Baylar et al. [2] underestimtes the predictions of E20, while the equations by Essery et al. [18] and Khdhiri et al. [31] overestimate the predictions of E20. Nonetheless, the equation by Baylar et al. [2] show satisfactory predictions for low values of E20 < 0.6 in the three flow regimes compared to the other examined equations despite being developed for values of E20 up to 0.9 as shown in Table 7.

By comparing Fig. 7 with 6, it can be seen how good is the agreement between the observed data and the values obtained by the new developed models compared to the values obtained from the currently

Table 7
Range of experimental measurements used to develop aeration prediction models.

Researcher	Equation	Flow regime	q (m^2/s x10^{-3})	N	F_r	h_c(m)	H(m)	l (m)	α(deg)	*E20*
Essery et al. [18]	(4)	Skimming, Nappe flow	1.5–21.75	–	–	0.0047–0.465	0.025–0.5	–	0–20	0.05–0.85
Baylar et al. [2]	(5)	Skimming, Transition, Nappe flow	16–166	8–50	0.1–9.5	0.03–0.14	0.05–0.15	0.042–0.58	14–50	0.16–0.82
Moulick et al. [35]	(6)	Nappe flow	1–17	6–14	0.001–0.95	0.01–0.03	0.2–0.5	0.61	19–39	0.76–0.90
Khdhiri et al. (2014)	(7)	Transition, Nappe flow	0.3–2.5	3–10	0.1–0.60	0.007–0.03	0.25–0.5	0.1–0.14	14–50	0.1–0.90
EPR and MT models	This paper	Skimming, Transition, Nappe flow	1–166	6–50	0.002–9.5	0.0047–0.14	0.05–3	–	14–50	0.16–0.90

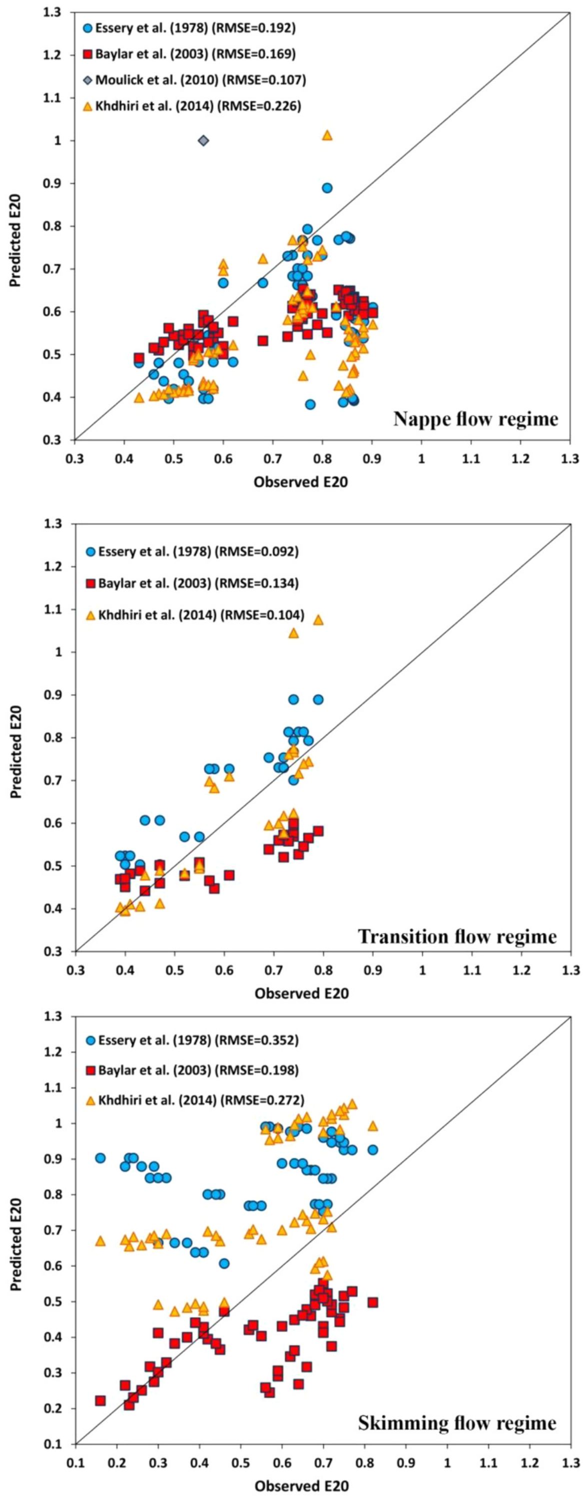

Fig. 7. Scatter plots of observed and predicted E20 for nappe, transition and skimming flow condition of traditional equations.

existing equations. This is due to the utilization of the power of the proposed artificial intelligent prediction models in addition to the usage of more comprehensive experimental data set for aeration efficiency spanning the three flow régimes. Thus, it can be concluded that the equations obtained from artificial intelligence methods in this study perform better than the currently existing equations obtained from regressive methods.

6. Conclusion

This study presents one of the first successful attempts in using the adaptive learning network (EPR and M5 MT) to develop models for the prediction of aeration efficiency in stepped weir under the three flow regimes, namely, the nappe, the transition and the skimming. Available laboratory data in literature were compiled to train and test the developed models. The Mallow's coefficient C_P was used to determine the effective variables affecting aeration efficiency. It was found that weir steps number, slope, the flow Reynolds number, and the ratio of the critical flow depth to the step height are the most important variables providing the lowest C_p. Both the EPR and M5 MT methods provide satisfactory predictions for the aeration efficiency. The two methods have high values of $R^2 > 0.86$ and low values for the RMSE < 0.052 and RMAE < 0.065. However, the EPR method has an advantage over the M5 MT method that it provides one equation for each regime, while the M5 MT method provides a number of equations for each regime. It was found that the equations obtained from artificial intelligence methods in this study perform better than the currently existing equations in the litrature obtained from regressive methods. Nonetheless, we acknowledge the relatively small size and range of the experimental datasets and recommend future studies to include more comprehensive field datasets to allow the models to better capture the effect of scale on the aeration efficiency.

References

[1] H. Azimi, H. Bonakdari, I. Ebtehaj, B. Gharabaghi, F. Khoshbin, Evolutionary design of generalized group method of data handling-type neural network for estimating the hydraulic jump roller length, Acta Mech. 229 (3) (2018) 1197–1214.
[2] A. Baylar, M.E. Emiroglu, Study of aeration efficiency at stepped channels, Proc. Inst. Civil. Eng. - Water Marit. Eng. 156 (3) (2003) 257–263.
[3] A. Baylar, M. Batan, Usage of artificial intelligence methods in free flowing gated closed conduits for estimation of oxygen transfer efficiency, Adv. Eng. Softw. 41 (5) (2010) 729–736.
[4] A. Baylar, T. Bagatur, Aeration efficiency with nappe flow over stepped cascade, Proc. Inst. Civil. Eng. - Water Manag. 160 (1) (2007) 43–50.
[5] A. Baylar, T. Bagatur, M.E. Emiroglu, Aeration efficiency with nappe flow over stepped cascades, Proc. Inst. Civil. Eng. - Water Manag. 160 (1) (2007) 43–50.
[6] A. Baylar, T. Bagatur, M.E. Emiroglu, Prediction of oxygen content of nappe, transition and skimming flow regimes in stepped-channel chutes, J. Environ. Eng. Sci. 6 (2) (2007) 201–208.
[7] A. Baylar, M.E. Emiroglu, T. Bagatur, An experimental investigation of aeration performance in stepped spillways, Water Environ. J. 20 (1) (2006) 35–42.
[8] A. Baylar, M. Unsal, F. Ozkan, GEP modeling of oxygen transfer efficiency prediction in aeration cascades, KSCE J. Civil. Eng. 15 (5) (2011) 799–804.
[9] H. Bonakdari, A.H. Zaji, B. Gharabaghi, I. Ebtehaj, M. Moazamnia, More accurate prediction of the complex velocity field in sewers based on uncertainty analysis using extreme learning machine technique, ISH J. Hydraul. Eng. (2018) 1–12, https://doi.org/10.1080/09715010.2018.1498753.
[10] B. Bhattacharya, D.P. Solomatine, Neural networks and M5 model trees in modelling water level–discharge relationship, Neurocomputing 63 (2005) 381–396.
[11] G.A. Camino, D.Z. Zhu, N. Rajaratnam, Hydraulics of stacked drop manholes, J. Irrig. Drain. Eng. 137 (8) (2011) 537–552.
[12] H. Chanson, Comparison of energy dissipation between nappe and skimming flow regimes on stepped chutes, J. Hydraul. Res. 32 (2) (1994) 213–218.
[13] H. Chanson, The Hydraulics of Stepped Chutes and Spillways, Swets & Zeitlinger, Lisse, The Netherlands, 2001.
[14] H. Chanson, L. Toombes, Air–water flows down stepped chutes: turbulence and flow structure observations, Int. J. Multiph. Flow. 28 (2002) (2002) 1737–1761.
[15] J. Cohen, P. Cohen, S.G. West, L.S. Aiken, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition, Lawrence Erlbaum, Mahwah, New Jersey, USA, 2003.
[16] I. Ebtehaj, H. Bonakdari, B. Gharabaghi, Development of more accurate discharge coefficient prediction equations for rectangular side weirs using adaptive neuro-fuzzy inference system and generalized group method of data handling, Measurement 116 (2018) 473–482.
[17] I. Ebtehaj, H. Bonakdari, A.H. Zaji, H. Azimi, A. Sharifi, Gene expression programming to predict the discharge coefficient in rectangular side weirs, Appl. Soft Comput. 35 (2015) 618–628.
[18] I.T.S. Essery, T.H.Y. Tebbutt, S.K. Rajaratnam Design of Spillways for Reaeration of Polluted Waters. CIRIA Report No. 72, Jan. London, UK, p. 36.
[19] J. Fox, Regression diagnostics: An introduction 79 Sage Publications, Newbury Park, California, USA, 1991.
[20] A.L.H. Gameson, Weirs and aeration of rivers, J. Inst. Water Eng. 11 (1957) 477–490.
[21] E.I. George, The variable selection problem, J. Am. Stat. Assoc. 95 (2000)

1304–1308.
[22] A. Gholami, H. Bonakdari, I. Ebtehaj, M. Mohammadian, B. Gharabaghi, S.R. Khodashenas, Uncertainty analysis of intelligent model of hybrid genetic algorithm and particle swarm optimization with ANFIS to predict threshold bank profile shape based on digital laser approach sensing, Measurement 121 (2018) 294–303.
[23] O. Giustolisi, D.A. Savic, Advances in data-driven analyses and modelling using EPR-MOGA, J. Hydroinform. 11 (3–4) (2009) 225–236.
[24] O. Giustolisi, D.A. Savic, Symbolic data-driven technique based on evolutionary polynomial regression, J. Hydroinform. 8 (3) (2006) 207–222.
[25] P. Guenther, S. Felder, H. Chanson, Flow aeration, cavity processes and energy dissipation on flat and pooled stepped spillways for embankments, Environ. Fluid Mech. 13 (5) (2013) 503–525.
[26] J.S. Gulliver, J.R. Thene, A.J. Rindels, Indexing gas transfer in self-aerated flows, J. Environ. Eng. 106 (3) (1990) 503–523.
[27] S. Haykin, Neural Networks—A Comprehensive Foundations, Prentice-Hall International, New Jersey, USA, 1999.
[28] F. Karimian, H. Elsawah, T. Zayed, O. Moselhi, A. Al Hawari, Forecasting breakage rate in water distribution networks using evolutionary polynomial regression, in: Proceedings of the 5th International/11th Construction Specialty Conference, Vancouver, Canada, 2015.
[29] A. Kaveh, T. Bakhshpoori, S.M. Hamze-Ziabari, Derivation of new equations for prediction of principal ground-motion parameters using M5′ algorithm, J. Earthq. Eng. 20 (6) (2016) 910–930.
[30] R. Khatibi, F. Salmasi, M.A. Ghorbani, H. Asadi, Modelling energy dissipation over stepped-gabion weirs by artificial intelligence, Water Ressour. Manag. 28 (7) (2014) 1807–1821.
[31] H. Khdhiri, O. Potier, J.P. Leclerc, Aeration efficiency over stepped cascades: better predictions from flow regimes, Water Res. 55 (2014) 194–202.
[32] M. Kobayashi, S. Sakata, Mallows' Cp criterion and unbiasedness of model selection, J. Econ. 45 (1990) 385–395.
[33] D. Laucelli, L. Berardi, A. Doglioni, O. Giustolisi, EPR-MOGA-XL: an excel based paradigm to enhance transfer of research achievements on data-driven modeling, in: R. Hinkelmann, M.H. Nasermoaddeli, S.Y. Li-ong, D. Savic, P. Fröhle (Eds), Proceedings of the 10th International Conference on Hydroinformatics HIC 2012, 14-18 July, Hamburg, Germany, 2012.
[34] C.L. Mallows, Some comments on Cp, Technometrics 15 (4) (1973) 661–675.
[35] S. Moulick, N.V. Tambada, B.K. Singh, B.C. Mal, Aeration characteristics of a rectangular stepped cascade system, Water Sci. Technol. 61 (2) (2010) 415–420.
[36] M. Najafzadeh, M. Rezaie Balf, E. Rashedi, Prediction of maximum scour depth around piers with debris accumulation using EPR, MT, and GEP models, J. Hydroinformatics (2016) jh2016212.
[37] P. Novak, Improvement of water quality in rivers by aeration at hydraulic structures, in: M. Hino (Ed.), Water quality and its control, 5 IAHR Hydraulic Structures Design Manual, Rotterdam, Netherland, 1994.
[38] I. Ohtsu, Y. Yasuda, Characteristics of flow conditions on stepped channels. in: Proceedings of 27th IAHR Biennial Congress, San Francisco, USA, Theme D, 583–588, 1997.
[39] J.R. Quinlan, Learning with continuous classes, in: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence AI'92. World Scientific, Singapore, 343-348, 1992.
[40] A. Rahimikhoob, Comparison between M5 Model Tree and Neural Networks for Estimating Reference Evapotranspiration in an Arid Environment, Water Resour. Manag. 28 (2014) 657–669.
[41] P. Reed, J.B. Kollat, V.K. Devireddy, Using interactive archives in evolutionary multiobjective optimization: a case study for long-term groundwater monitoring design, Environ. Model. Softw. 22 (5) (2007) 683–692.
[42] M. Rezania, A. Faramarzi, A.A. Javadi, An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacement, Eng. Appl. Artif. Intell. 24 (1) (2011) 142–153.
[43] A. Safarzadeh, A.H. Zaji, H. Bonakdari, Comparative assessment of the hybrid genetic algorithm–artificial neural network and genetic programming methods for the prediction of longitudinal velocity field around a single straight groyne, Appl. Soft Comput. 60 (2017) 213–228.
[44] A.M. Sattar, Gene expression models for the prediction of longitudinal dispersion coefficients in transitional and turbulent pipe flow, J. Pipeline Syst. Eng. Pract. 5 (1) (2014) 04013011.
[45] A.M. Sattar, Gene expression models for prediction of dam breach parameters, J. Hydroinformatics 16 (3) (2014) 550–571.
[46] A.M. Sattar, B. Gharabaghi, E.A. McBean, Prediction of timing of watermain failure using gene expression models, Water Resour. Manag. 30 (5) (2016) 1635–1651.
[47] A.M. Sattar, Ö.F. Ertuğrul, B. Gharabaghi, E.A. McBean, J. Cao, Extreme learning machine model for water network management, Neural Comput. Appl. (2017) 1–13.
[48] A.M.A. Sattar, B. Gharabaghi, F. Sabouri, A.M. Thompson, Urban stormwater thermal gene expression models for protection of sensitive receiving streams, Hydrol. Process. 31 (13) (2017) 2330–2348.
[49] A.M. Sattar, K. Plesiński, A. Radecki-Pawlik, B. Gharabaghi, Scour depth model for grade control structures, J. Hydroinform. 20 (1) (2018) 117–133.
[50] T.H.Y. Tebbutt, I.T.S. Essery, S.K. Rasaratnam, Reaeration performance of stepped cascades, J. Inst. Water Eng. Sci. 31 (4) (1977) 285–297.
[51] L. Toombes, H. Chanson, Air–water mass transfer on a stepped waterway, J. Environ. Eng. 131 (10) (2005) 1377–1386.
[52] L. Toombes, H. Chanson, Air water flow and gas transfer at aeration cascades a comparative study of smooth and stepped chutes. in: International Workshop on Hydraulics of Stepped Spillways, in: H.E. Minor, W.H. Hager, (eds.), Balkema, Rotterdom, The Netherlands, 77–84, 2000.
[53] Y. Wang, I.H. Witten, Induction of model trees for predicting continuous lasses. in: Proceedings of the poster papers of the European conference on machine learning. Prague, Czech Republic, 1997.
[54] P. Wormleaton, C. Tsang, Aeration performance of rectangular planform labyrinth weirs, J. Environ. Eng. 126 (5) (2000) 456–465.
[55] G. Zhang, H. Chanson, Hydraulics of the developing flow region of stepped spillways. II: Pressure and velocity fields, J. Hydraul. Eng. 142 (7) (2016) 04016016.
[56] G. Zhang, H. Chanson, Self-aeration in the rapidly-and gradually-varying flow regions of steep smooth and stepped spillways, Environ. Fluid Mech. 17 (1) (2017) 27–46.
[57] S.C. Jain, Air transport in vortex-flow drop shaft, J. Hydraul. Eng. 114 (12) (1988) 1485–1497.
[58] F. Granata, G. de Marinis, R. Gargano, Air-water flows in circular drop manholes. urban Water, Journal 12 (6) (2014) 477–487.
[59] F. Granata, Dropshaft cascades in urban drainage systems, Water Sci. Technol. 73 (9) (2016) 2052–2059.
[60] D.P. Solomatine, Y. Xue, M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China, J. Hydrol. Eng. 9 (6) (2004) 491–501.
[61] K.K. Singh, M. Pal, V.P. Singh, Estimation of mean annual flood in Indian catchments using backpropagation neural network and M5 model tree, Water Resour. Manag. 24 (10) (2010) 2007–2019.
[62] M.K. Goyal, Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression, Water Resour. Manag. 28 (7) (2014) 1991–2003.
[63] F. Granata, S. Papirio, G. Esposito, R. Gargano, G. de Marinis, Machine Learning algorithms for the forecasting of wastewater quality indicators, Water 9 (2) (2017) 105.
[64] F. Granata, G. de Marinis, Machine learning methods for wastewater hydraulics, Flow. Meas. Instrum. 57 (2017) 1–9.
[65] D.A. Van Veldhuizen, G.B. Lamont, Multiobjective Evolutionary Algorithms: Analysing the State-of-the-Art, Evol. Comput. 8 (2) (2000) 125–147.