Determination of compound channel apparent shear stress: application of novel data mining models

Zohreh Sheikh Khozani, Khabat Khosravi, Binh Thai Pham, Bjørn Kløve, Wan Hanna Melini Wan Mohtar and Zaher Mundher Yaseen

ABSTRACT

Momentum exchange in the mixing region between the floodplain and the main channel is an essential hydraulic process, particularly for the estimation of discharge. The current study investigated various data mining models to estimate apparent shear stress in a symmetric compound channel with smooth and rough floodplains. The applied predictive models include random forest (RF), random tree (RT), reduced error pruning tree (REPT), M5P, and the distinguished hybrid bagging-M5P model. The models are constructed based on several correlated physical channel characteristic variables to predict the apparent shear stress. A sensitivity analysis is applied to select the best function tuning parameters for each model. Results showed that input with six variables exhibited the best prediction results for RF model while input with four variables produced the best performance for other models. Based on the optimised input variables for each model, the efficiency of five predictive models discussed here was evaluated. It was found that the M5P and hybrid bagging-M5P models with the coefficient of determination (R^2) equal to 0.905 and 0.92, respectively, in the testing stage are superior in estimating apparent shear stress in compound channels than other RF, RT and REPT models.

Key words | apparent shear stress, compound channel, data mining, prediction

Zohreh Sheikh Khozani
Wan Hanna Melini Wan Mohtar
Smart and Sustainable Township Research Center,
Faculty of Engineering & Built Environment,
Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor,
Malaysia

Khabat Khosravi
Department of Watershed Management
Engineering, Faculty of Natural Resources,
Sari Agricultural Science and Natural Resources
University,
Sari,
Iran

Binh Thai Pham
Institute of Research and Development,
Duy Tan University,
Da Nang 550000,
Vietnam

Bjørn Kløve
Water, Energy and Environmental Engineering
Research Unit, Faculty of Technology,
University of Oulu,
Finland

Zaher Mundher Yaseen (corresponding author)
Sustainable Developments in Civil Engineering
Research Group, Faculty of Civil Engineering,
Ton Duc Thang University,
Ho Chi Minh City,
Vietnam
E-mail: *yaseen@tdtu.edu.vn*

INTRODUCTION

The compound cross section is a typical cross section in natural rivers which consists of a main channel and one or two floodplains (Al-Khatib *et al.* 2013; Tang 2017). In flooded conditions, the floodplain has lower velocity than the main channel, leading to momentum transfers between the interfaces. In compound channels, the total flow resistance increases with transverse momentum transfer as first noted by Sellin (1964), and was then extensively studied until the millennium (Myers 1978; Fernandes *et al.* 2015). Improved methods for estimating momentum transfer at the main channel–floodplain interfaces have then been presented using apparent shear stress, compared to the traditional channel methods (Rajaratnam & Ahmadi 1981; Knight & Hamed 1984; Devi & Khatua 2016). Özbek *et al.* (2004) calculated the apparent shear stress and discharge in the symmetric compound channels with different floodplain widths, using directional division planes' methods. Based on the ratio of apparent shear stresses across the interfaces, the horizontal and diagonal interfaces' planes have better accuracy in estimating the discharge capacity than the vertical planes. Khatua *et al.* (2010) measured the shear stress distribution in channels with compound cross

doi: 10.2166/hydro.2019.037

section and extracted equations to compute the apparent shear force values for different considered interfaces. Moreta & Martin-Vide (2010) proposed a generalised formula for apparent shear stress prediction and validated it for an extensive range of experimental data. They studied the effect of geometry and roughness on a non-dimensional friction coefficient acting on the vertical main channel–floodplain interface (C_{fa}). They concluded that the variation aspect ratio (the flow depth of main channel divided by the width of main channel) influences the apparent friction coefficient, but the parameter of bank side slope is not very effective on C_{fa}.

All discussed formulae require realisation of the velocity gradient, the influence of roughness and channel geometry on the estimation of apparent shear stress. Since determining the velocity gradient without detailed measurements is difficult and C_{fa} has high uncertainty with different geometry and roughness based on the results of Moreta & Martin-Vide (2010), finding other methods that can estimate apparent shear stress without a need for these parameters is attractive.

One of the recent methods replacing or improvising the previous methods, which increased higher precision of results, is artificial intelligence (AI)-based numerical techniques. With progress in these methods, several studies have been successfully carried out to predict different phenomena in hydraulics and hydrology such as evaporation modelling, shear stress prediction and streamflow forecasting (Kisi 2008; Sheikh Khozani *et al.* 2016, 2017a; Yaseen *et al.* 2016). Higher efficiency in predicting the shear stress in circular channels was found using gene expression programming (GEP), extreme learning machines, the M5 model tree algorithm and a randomised neural network technique (Sheikh Khozani *et al.* 2015, 2017a, 2017b, 2017c, 2018). The neural network was able to simulate and capture the complexities of spatial momentum transfer at the main channel–floodplain interface for compound channels with and without vegetation (Huai *et al.* 2013). Prediction of apparent shear stress using the genetic algorithm–artificial neural network (GA-ANN) and genetic programming (GP), along with multiple linear regression (MLR), was studied by Bonakdari *et al.* (2018). They concluded that the GAA method was more powerful than GP and MLR methods and a matrix-based equation to forecast the apparent shear stress was presented. However, these AI methods, especially the ANN technique, as one of the most widely used models (Houichi *et al.* 2012; Hanspal *et al.* 2013; Choi *et al.* 2015; Yaseen *et al.* 2015; Fahimi *et al.* 2016; Tapoglou *et al.* 2019) in both spatial or time series data prediction, need long time series data for both testing and training dataset (Melesse *et al.* 2011). Furthermore, one of the known limitations is the poor prediction for the data which are not in the range of the learning values (Melesse *et al.* 2011; Khosravi *et al.* 2018b). Therefore, soft computing models are mostly integrated with other approaches to overcome the weakness of an individual model. The possibility of an efficient integrated model is immense, and proven techniques include the combination of ANN with fuzzy logic (FL) and an adaptive neuro-fuzzy inference system (ANFIS). Although this hybrid model has a higher prediction power than both the individual ANN and FL, and it has been prominently applied, the model still suffers setbacks in finding the optimal parameter for a neural fuzzy model and determination of the best weights in a membership analysis function (Khosravi *et al.* 2018a). Generally, those models which have a hidden layer in their structure have a lower prediction power than other models without a hidden layer, such as the decision tree algorithms (i.e., random forest, random tree and so on) (Kisi *et al.* 2012). In addition, researchers are always motivated to investigate more reliable and robust models and to explore other algorithms with better and solid results. As each available model has its own advantages and disadvantages, researchers are keen to find improved prediction models, which are more robust, and the advantages outweigh the disadvantages. Nowadays, data mining methods have been declared as dependable modelling approaches to solve regression problems in multiple engineering and science applications (Witten & Frank 2005; Baker & Yacef 2009; Witten *et al.* 2011). It is crucial to investigate the prediction capacity of newly developed data mining algorithms to improve the accuracy in prediction of hydraulic variables such as apparent shear stress.

Despite the capability of data mining techniques to solve complicated problems (Khosravi *et al.* 2018a; Sharafati *et al.* 2019), application in the field of hydraulics is limited, and only a few studies have paid attention to forecasting the apparent shear stress in channels with compound cross section. The

accuracy of some innovative decision tree algorithms of random forest (RF), random tree model (RT), reduced error pruning tree (REPT) and M5P, and also a hybrid model of bagging-M5P structure is investigated in this study to assess the performance of these new algorithms for estimating the apparent shear stress in symmetric compound channels with smooth and rough floodplains. This study utilised the advantages of model trees, including faster in training, higher potential of convergence and assisting decision-making due to better transparency (Solomatine & Xue 2004). The RF model, in particular, lowers the risk of over-fitting and has less variance. As the determination of apparent shear stress is based on the geometrical characteristics of a channel, we attempt to obtain the best input variables (and the most effective parameters) with higher efficiency.

DATA COLLECTION AND PREPARATION

Theory and review

Estimation of the apparent shear stress requires knowledge of the velocity gradient (ΔU) and channel geometry and characteristics. The key empirical equations used to forecast the apparent shear stress and calculate discharge in compound channels (with both smooth and rough floodplains) are shown in Table 1. Most of the equations have similar geometrical parameters, particularly the velocity gradient, and widths and water depths for both main channel and floodplain, with noticeable variably dispersed numerical coefficients.

Data collection and preparation

Datasets are collected from several experimental studies on compound channels. For the smooth, symmetric, rectangular cross sections, the laboratory results of Knight & Demetriou (1983) were used. The researchers used a flume with dimensions of 15 m long, 0.61 m wide and 9.66×10^{-4} bed slope. The flume consisted of two floodplains with size of 0.076 m high and 0.229 m wide. Knight & Hamed (1984) used the same flume and investigated different roughness values in the floodplains and the main channel. Prinos & Townsend (1984) experimentally studied a compound channel with a trapezoidal cross section. The

Table 1 | Equations for calculating apparent shear stress

Researchers	Experimental condition	Presented equation
Ervine *et al.* (1982)	Symmetric and asymmetric, smooth and rough floodplains	$\tau_a = \rho \frac{7.1}{N_f} (\Delta U)^2$
Wormleaton *et al.* (1982)	Symmetric, smooth and rough floodplains	$\tau_a = 13.84(\Delta U)^{0.822} \left(\frac{H}{h}\right)^{-3.123} \left(\frac{B}{b}\right)^{-0.727}$
Knight & Demetriou (1983)	Symmetric, smooth and rough floodplains	$\tau_a = \frac{H}{2H(B/b) + 2h} - f\left(\frac{B}{b}, Hr\right)$
Baird & Ervine (1984)	Asymmetric, smooth floodplains	$\tau_a = \left[\frac{H}{H-h} - 1 + 1.5\left(\frac{h}{b}\right)^{1.25}\right]^{1.5} \left(\frac{h}{b}\right)^{0.5} \left[0.5 + 0.3 \ln\left(\frac{B_f}{h}\right)\right](\rho g(H-h)S_0)$
Prinos & Townsend (1984)	Symmetric, smooth and rough floodplains	$\tau_a = 0.874(\Delta U)^{0.92} \left(\frac{H-h}{H}\right)^{-1.129} \left(\frac{B}{b}\right)^{-0.514}$
Wormleaton & Merrett (1990)	Symmetric, smooth and rough floodplains (FCF)	$\tau_a = 3.325(\Delta U)^{1.451} (H-h)^{-0.354} B^{0.519}$
Christodoulou (1992)	Symmetric, smooth floodplains	$\tau_a = \frac{1}{2}\rho \times 0.01 \left(\frac{B}{b}\right) \times (\Delta U)^2$
Bousmar & Zech (1999)	Symmetric and asymmetric, smooth and rough floodplains (FCF)	$\tau_a = \rho \times 0.16 \times (\Delta U)^2$

N_f is the number of floodplains, B is the total width, b is the width of the main channel, H is the flow depth in a compound channel, h is the flow depth in the main channel and ΔU is the velocity difference between the main channel and the floodplain.

experiments were conducted in a main channel with dimensions of 1.02 m deep, 2:1 side slope and the bed channel slope was fixed at 3×10^{-4}. They measured the shear stress along the whole wetted perimeter, then calculated the apparent shear stress at different flow depths. In addition to the small-scale flume data, large-scale flood channel facility (FCF) experimental data from the work of Wormleaton & Merrett (1990) were used to estimate the apparent shear stress. This facility entailed a flume size of 56 m long and 10 m wide, with a trapezoidal main channel cross section. They examined several methods for calculating discharge and provided a modified form that had a realistic understanding of the interaction between the main channel and the flood interface. Figure 1 demonstrates the cross section of compound channels whose data were used. Considering the results of several researchers, the effective parameters in apparent shear stress values are channel height (H), total width (B), main channel widths (b), water depth in floodplain (h), floodplain roughness (n_f) and main channel roughness (n_c). Therefore, the apparent shear stress is a function of

$$\tau_a = F(H, h, B, b, n_f, n_c) \tag{1}$$

Using Buckingham's theorem, the dimensionless apparent shear stress ($\tau_a^* = \frac{\tau_a}{\rho g R_h S_0}$) is a function of six dimensionless parameters as:

$$\tau_a^* = f\left(\frac{B}{b}, \frac{n_f}{n_c}, \frac{(H-h)}{h}, \frac{h}{b}, \frac{H}{B}, \frac{BH}{bh}\right) \tag{2}$$

where R_h is the hydraulic radius and S_0 is the bed slope. A total of 100 data for each input were extracted from the four mentioned experimental studies. Out of available data, about 70% was used for the training procedure and the rest were reserved for the testing stage. Table 2 shows the statistics of the dimensionless parameters based on the experimental data.

MODELLING STRATEGY BACKGROUND

Random forest model

The RF model introduced by Breiman (2001), is one of the learning methods that generates many categories and provides the final outcome based on aggregated results. It

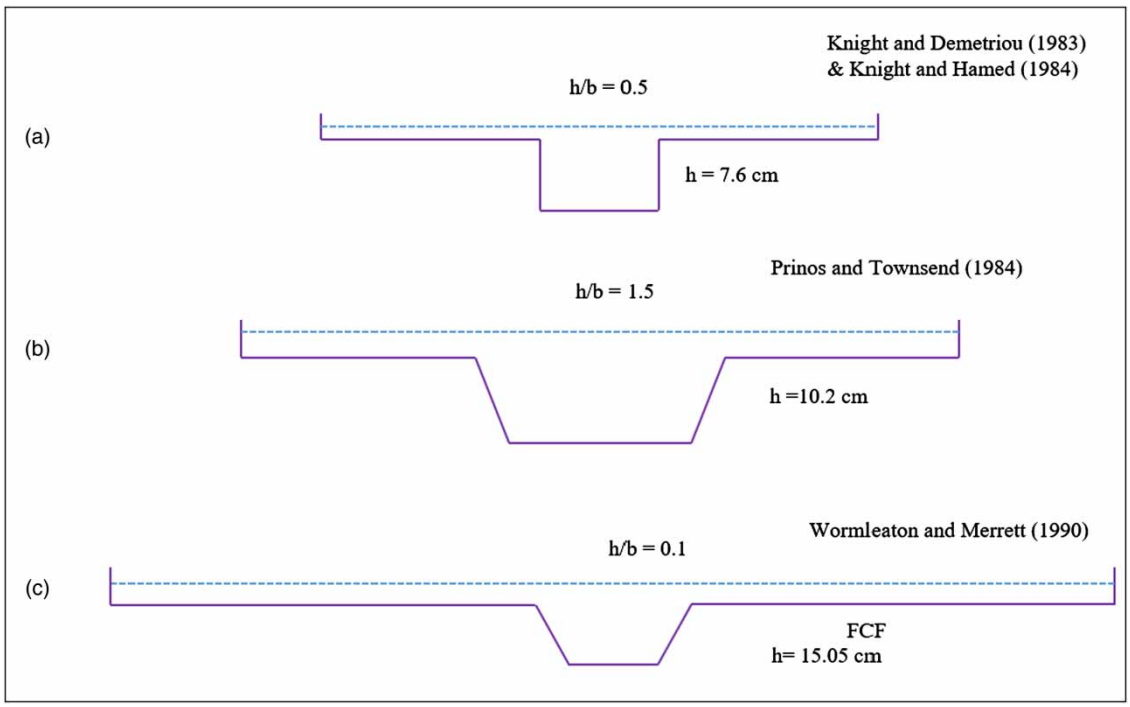

Figure 1 | Compound channel cross section used: (a) Knight & Demetriou (1983) also Knight & Hamed (1984), (b) Prinos & Townsend (1984), (c) Wormleaton & Merrett (1990).

Table 2 | Range of geometric data used for the compound channel

Datasets	Variables	Minimum	Maximum	Mean	Skewness	Kurtosis
Training dataset	B/b	2.00	6.67	3.97	0.08	0.73
	$(H-h)/H$	0.02	0.51	0.21	0.53	−1.07
	H/h	1.02	2.02	1.32	1.02	0.02
	BH/bh	2.24	8.10	5.16	0.10	−0.59
	h/b	0.10	2.00	1.05	0.26	−1.84
	H/B	0.02	0.58	0.31	0.05	1.12
	n_f/n_c	1.00	3.03	1.42	1.28	−0.01
Testing dataset	B/b	2.00	5.26	3.96	−0.54	−0.01
	$(H-h)/H$	0.04	0.50	0.24	0.73	−0.17
	H/h	1.05	2.00	1.47	1.33	0.89
	BH/bh	2.49	8.00	5.31	−0.26	0.26
	h/b	0.10	2.00	0.97	0.68	−1.52
	H/B	0.00	0.56	0.28	0.28	−1.08
	n_f/n_c	1.00	2.83	1.37	1.43	1.39

uses a different bootstrap sample from the data to build a tree for regression (Liaw & Wiener 2002). In addition, it changes how regression trees are constructed. In a random forest, the best among a subset of predictors randomly selected in a node is used to divide that node (Liaw & Wiener 2002). This is also an advantage of the random forest which helps it to perform well compared with other models like support vector machine (SVM) and neural networks. The RF model is robust against over-fitting (Breiman 2001). In the RF model, three parameters require optimisation: (1) the number of different predictors, (2) the number of regression trees, (3) the minimal size of the terminal nodes (Mutanga *et al.* 2012). Random forest regression performs as shown below (Liaw & Wiener 2002):

1. n_{tree} bootstrap samples are drawn from the original data.
2. An unpruned regression tree is grown for each bootstrap sample with the following modification: sample randomly m_{try} of the predictors at each node and select the best split among the variables.
3. Aggregate the predictions of the n_{tree} trees to predict new data using averaged values for regression.

Random tree model

Random tree is a quick and flexible tree model which has been applied for path planning of robotics and many real-world problems (Haitao *et al.* 2007). It builds the decision trees on a random subset of columns based on a stochastic process. Random tree works similarly to other traditional decision trees like C45 or J48 except that only a random subset is available for each split (Dehling *et al.* 2008). Suppose R is defined as a plane rooted tree called a family tree with n nodes, E is defined as a class of a plane rooted tree, thus, each $R \in E$ the size $|R|$ by the number of nodes R comprises a weight indicated as the following equation (Drmota & Gittenberger 1997):

$$\omega(R) = \prod_{k\geq 0} \alpha_k^{n_k(R)} \tag{3}$$

where $n_k(R)$ is the number of the nodes $\upsilon \in R$ without-degree k, $(\alpha_k; k \geq 0)$ are the non-negative numbers. Let us set $a_n = \sum_{R:|R|=n} \omega(R)$ then the corresponding function $a(r) = \sum_{n\geq 0} a_n r^n$ must satisfy the functional function as below:

$$a(r) = r\varphi(a(r)) \tag{4}$$

where

$$\varphi(x) = \sum\nolimits_{k\geq 0} \varphi_k x^k \tag{5}$$

Finally, equip the sets $C_n = \{R \in E : |R| = n\}$ with the probability distribution caused by the weight function $\omega(R)$.

REP tree model

A REP tree model which creates a regression tree based on the reduction of variance or increased information

(Mohamed *et al.* 2012) is a fast decision tree method. It has been applied in many studies for solving many real-world problems such as bankruptcy prediction (Cielen *et al.* 2004), academic performance prediction (Vandamme *et al.* 2007) and heart disease prediction (Pandey *et al.* 2013). The training step of the REP tree model is carried out in two main stages as: (1) the regression tree is utilised to generate multiple trees in various iterations and (2) the best tree is selected from the generated trees. The advantage of the REP tree model compared with other decision trees is that it uses the reduced error pruning (REP) technique to prevent the over-fitting problems and handle the missing values (Elomaa & Kaariainen 2001).

Suppose $[U, V]$ is explained as the regression tree of which U and V demonstrate the leaf of the tree and the output variable, respectively. The REP tree model constructs a tree with maximum information gain with stopping criteria of the sum of squared errors indicated as follows:

$$E = \sum_{Eeleaves(RT)} p_c G_c \tag{6}$$

where G_c is known to be the within variable and p_c is the class prediction.

M5P model

M5P model is a machine learning decision tree method which was first presented by Quinlan (1992). It combines a linear regression function and a conventional decision tree to construct the regression tree for prediction. In the M5P tree, the conventional decision tree is used at the node whereas the linear regression function is utilised as the leaves. Generation of the M5P tree is carried out through two main steps, including (1) a decision tree, using the division criteria based on the standard deviation of class values, and (2) the linear regression function for replacing the sub-trees, and the tree is grown by using the pruning process (Pal 2006). In addition, the standard deviation reduction (SD) is used to reduce errors of the M5P algorithm expressed as follows (Sanikhani *et al.* 2018):

$$SD = S(D) - \sum \frac{|D_i|}{|D|} sd(D_i) \tag{7}$$

where D is inferred as a set of samples which reach the mode of the tree, D_i is defined as the subsets of samples which have the i-th output of the potential set, $S(D)$ is defined as the standard deviation of D.

Hybrid model of bagging-M5P

Bagging-M5P is a hybrid machine learning model built by combining the bagging ensemble and M5P predictor. The bagging ensemble was proposed by Breiman (1996) and has been applied effectively in many studies to deal with many real-world problems such as ecological prediction (Prasad *et al.* 2006), Lyme disease risk prediction (Rizzoli *et al.* 2002) and early prediction of heart disease (Chaurasia & Pal 2013).

The robustness of the bagging is that it can enhance the performance of the weak predictors, as it can raise the recognition rate of unstable predictors and weaken the defects of component predictors (Breiman 1996). There are three main steps to train the bagging ensemble: (1) choosing independently and randomly the data from the original training dataset, (2) designating the M5P learning algorithm to train the various sub-datasets for gaining the sequence of predictive function, (3) voting for the results and choosing the final outcome with the most votes (Breiman 1996).

Models' performance evaluation

The models' performances are evaluated using six statistical parameters, i.e., coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), Nash–Sutcliffe efficiency (NSE), percentage of BIAS (PBIAS) and the RMSE of the standard deviation of observation ratio (RSR) (Tao *et al.* 2018). These statistical parameters are defined as:

$$R^2 = \frac{\left(n\sum_{i=1}^{n} x_{ip}x_{io} - \sum_{i=1}^{n} x_{ip}\sum_{i=1}^{n} x_{io}\right)^2}{\left(n\sum_{i=1}^{n} x_{ip}^2 - \left(\sum_{i=1}^{n} x_{ip}\right)^2\right)\left(n\sum_{i=1}^{n} x_{io}^2 - \left(\sum_{i=1}^{n} x_{io}\right)^2\right)} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_{ip} - x_{io})^2}{n}} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_{ip} - x_{io}| \tag{10}$$

$$NSE = \frac{\sum_{i=1}^{n} (x_{ip} - x_{im})^2}{\sum_{i=1}^{n} (x_{ip} - \bar{x}_{ip})^2} \quad (11)$$

$$PBIAS = 100 \times \left[\frac{\sum_{i=1}^{n} (x_{io} - x_{ip})}{\sum_{i=1}^{n} x_{ip}} \right] \quad (12)$$

$$RSR = \sqrt{\frac{\sum_{i=1}^{n} (x_{ip} - x_{io})^2}{\sum_{i=1}^{n} (x_{ip} - \bar{x}_{ip})^2}} \quad (13)$$

where x_{io} and x_{ip} are observed and predicted values of apparent shear stress, also $\bar{x}_{io}$ and $\bar{x}_{ip}$ are the observed and predicted mean value of apparent shear stress, respectively.

The box plot diagram is a useful tool for understanding the distribution and scattering of data, and is widely used in many cases. The graphical presentation offers instantaneous and prompt information, which is more useful than simply listed data in a table. This approach is a good option in terms of examining the maximum, minimum and other useful statistical analysis information, particularly in comparing the same values in different sets. This chart prevents the misjudgement of data by reference to central parameters such as the average and the median, by allowing both of them and making comparison possible. Therefore, attention is drawn to the problem of data dispersion and variability. Besides the statistical index-based assessment, the box plot analysis is also used to evaluate the performance of the five discussed models.

Methodology flowchart

Methodology of this study can be carried out in five main steps, namely, (1) preparing the data, (2) pre-processing data, (3) generating the datasets, (4) processing the models and (5) validating the models as shown in Figure 2. A detailed description of these steps is given below.

(1) Preparing the data: The data were collected from various credible experimental studies as previously described. As mentioned before, the data include input parameters, namely, channel height (H), total width (B), main channel widths (b), water depth in floodplain (h), floodplain roughness (n_f), main channel roughness (n_c) and an output variable named the apparent shear stress.

(2) Pre-processing data or selection of input variables: At first, different dimensionless parameters were considered based on the input and output parameters and

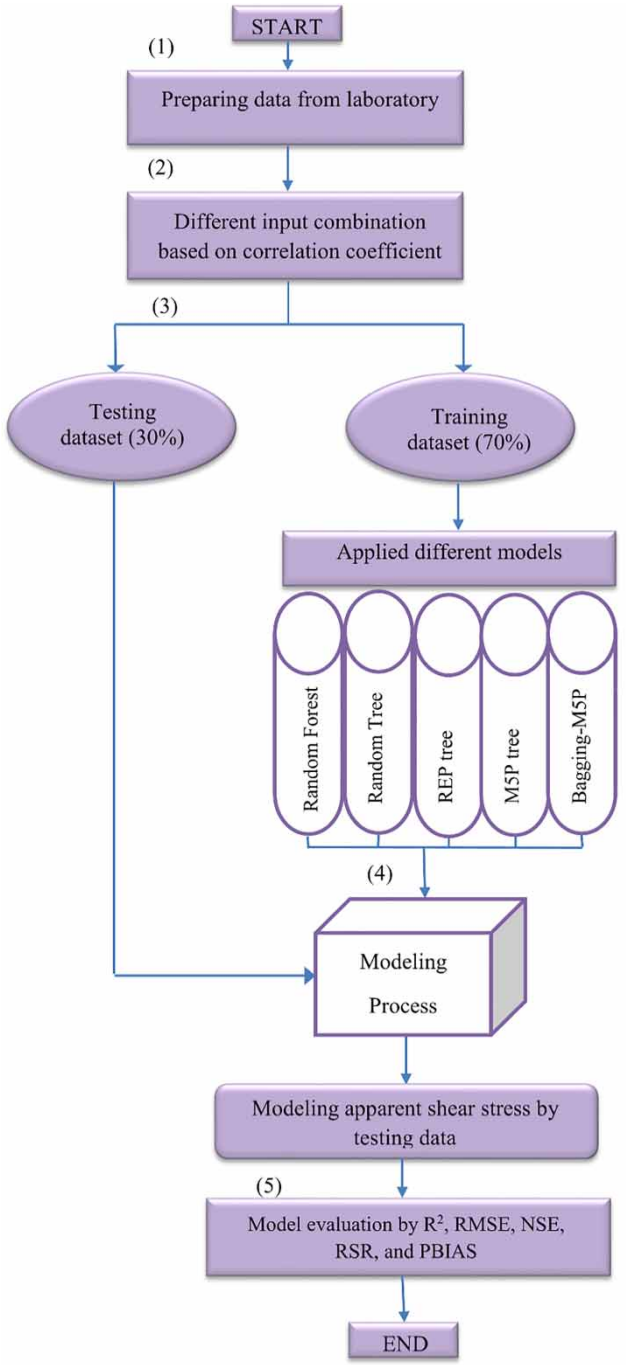

Figure 2 | The applied modeling procedure flowchart.

next, different input combinations were carried out using correlation coefficient to find the best one and have a higher prediction power.

(3) Generating the datasets: In this step, all input data were divided into two parts. Of these parts, about 70% was used for the training procedure and the rest (30%) was reserved for the testing stage.

(4) Processing the models: In this step, the training dataset was used to construct the models and train them. RF model was constructed with a bag size percentage, batch size, maximum depth of tree, number of decimal places, number of execution slots, number of features, number of iterations and seeds with optimal values of 1, 100, 100, 0, 2, 1, 6, 130 and 1, respectively. Bag size percentage is the size of each bag as a percentage of the training set size, preferred number of instances to process for batch prediction, the number of decimal places to be used for the number output in the model, the number of execution slots (threads) used to construct the ensemble and the number of iterations required:
   1. For RT model, the K value of 1, batch size of 100, maximum depth of tree of 0, minimum number of 1, minimum variance probability of 0.001, number of decimal places of 2, number of folds of 0 and seeds of 3 are used.
   2. The REP tree model was built with the batch size of 100, initial count of 0, maximum depth of −1, minimum number of 1, minimum variance probability of 0.001, number of decimal places of 2, number of folds of 8 and seeds of 3.
   3. The optimal values for M5P model operators, including of batch size, minimum number of instances and number of decimal places were 100, 4 and 2, respectively.
   4. The optimal values of these operators for bagging-M5P model were 20, 100, 2, 0, 10 and 1, for bag size percentage, batch size, number of decimal places, number of execution slots, and number of iterations and seeds, respectively. Finally the test data were predicted by the constructed models.

(5) Validating the models: In this step, the models were validated based on the testing datasets. Various statistical methods, namely, R^2, RMSE, NSE, RSR and PBIAS were used for validation of the developed models.

RESULTS AND DISCUSSION

Selection of input variables

In order to identify the most effective input variables based on the channel geometry and characteristics, different combinations of these inputs were examined. Based on the primary correlation analysis reported in Table 3, the ratio of (H/B) attained the highest R^2 value followed by (h/b). According to the correlation trend, the input combinations were constructed to represent the characteristics of the predictive models (see Table 4). A total of nine input combinations were constructed using those input variables. It is worth mentioning here that the ratio of H/B was incorporated in all constructed input combinations owing to its essential variability to predict the apparent shear stress of the compound channel.

Table 3 | The analysis results of the Pearson correlation coefficient values towards the apparent shear stress

Variables	Pearson correlation coefficient
B/b	−0.13
$(H-h)/H$	0.074
H/h	0.008
BH/bh	−0.12
h/b	−0.66
H/B	−0.72
n_f/n_c	−0.035

Table 4 | The constructed input combinations for the applied data mining predictive models

Input no.	Input variables
1	H/B
2	H/B, h/b
3	H/B, h/b, B/b
4	H/B, h/b, B/b, BH/bh
5	H/B, h/b, B/b, BH/bh, $(H-h)/H$
6	H/B, h/b, B/b, BH/bh, $(H-h)/H$, n_f/n_c
7	H/B, h/b, B/b, BH/bh, $(H-h)/H$, n_f/n_c, H/h
8	B/b, $(H-h)/H$, n_f/n_c, h/b
9	B/b, H/B, n_f/n_c, h/b

Table 5 | The sensitivity examination of the various input combinations to predict the apparent shear stress over all the applied predictive models

Models	Criteria	Input 1	Input 2	Input 3	Input 4	Input 5	Input 6	Input 7	Input 8	Input 9
RF	CC	0.63	0.65	0.77	0.72	0.76	0.78	0.77	0.91	0.77
	RMSE	1.5	1.4	1.1	1.2	1.1	1	1	0.60	1
RT	CC	0.61	0.61	0.78	0.55	0.67	0.79	0.2	0.72	0.67
	RMSE	1.6	1.6	1.1	1.8	1.5	0.99	2.2	1.3	1.1
REPT	CC	0.67	0.67	0.67	0.68	0.65	0.65	0.65	0.76	0.67
	RMSE	1.5	1.5	1.5	1.5	1.6	1.6	1.6	1.3	1.5
M5P	CC	0.64	0.64	0.67	0.73	0.73	0.73	0.73	0.9	0.67
	RMSE	1.4	1.4	1.4	1.4	1.4	1.3	1.3	0.71	1.4
Bagging M5P	CC	0.7	0.71	0.72	0.73	0.72	0.68	0.68	0.86	0.7
	RMSE	1.4	1.3	1.3	1.2	1.2	1.2	1.3	0.87	1.2

Based on the constructed input combination in Table 4, a sensitivity analysis was performed to review the appropriate input combination for each developed model. This is accomplished through the calculation of correlation coefficient (CC) and root mean square error (RMSE) for each applied predictive model. The four applied models, i.e., RF, REPT, M5P and bagging-M5P were found to have a consistent behaviour with the combination of input 8: (B/b, $(H-h)/H$, n_f/n_c, h/b) as an applicable attribute to predict the apparent shear stress. On the other hand, the RT model was performed accurately using the sixth input combination through incorporating the following parameters (i.e., H/B, h/b, B/b, BH/bh, $(H-h)/H$, n_f/n_c) (see Table 5).

The attained results in Table 5 indicate two facts: (i) the initiated input combinations based on the Pearson correlation coefficient are reliably suitable for predicting the τ_a^* due to the consistency of the input variables. However, (ii) RT acted differently using another order of input variables' information and this is somehow very normal as those stochastic models visualised the regression problem from one case to another in different manners.

Evaluation of models in predicting apparent shear stress

Following several prediction researches conducted in the literature and within the hydraulic engineering perspective, several performance metrics preferably should be conducted for the predictability assessment (Chadalawada & Babovic 2019). Based on the quantitative examination, different prediction skills metrics computed the predictability of the proposed data mining models including R^2, RMSE, MAE, NSE, PBIAS and RSR. Multiple indicators for prediction accuracy assessment give a more comprehensive vision of the capacity of the developed models. In general, it was observed that the applied models showed good predictability performance, as shown in Table 6. However, all analysed statistical parameters showed that the bagging-M5P model performed with superior prediction accuracy in comparison with the other models. The coefficient of determination (R^2) exhibited a very good index between the simulated and the actual experiment τ_a^*. The results were presented based on their performance, starting with the highest, the integrated bagging-M5P ($R^2 = 0.92$), M5P ($R^2 = 0.90$), RF ($R^2 = 0.80$), REPT ($R^2 = 0.73$), RT ($R^2 = 0.72$). In harmony with the displayed scatter plot (see Figure 3), bagging-M5P demonstrated a reliable predictive model with minimum diversion from the best fit line for all the range of τ_a^* data (0–5). As a matter of fact, R^2 value is a sensitive indicator towards the outlier observations (Legates & Mccabe 1999; Yaseen *et al.* 2016), and other indicators are required to be evaluated. With respect to

Table 6 | The performance metrics results of the applied data mining predictive models over the test modelling phase

Models	R^2	RMSE	MAE	NSE	PBIAS	RSR
RF	0.805	0.698	0.457	0.802	2.242	0.441
RT	0.727	0.876	0.565	0.689	14.404	0.556
REPT	0.732	0.908	0.673	0.666	−14.160	0.577
M5P	0.905	0.487	0.334	0.903	2.073	0.309
Bagging-M5P	0.920	0.460	0.320	0.914	−5.517	0.292

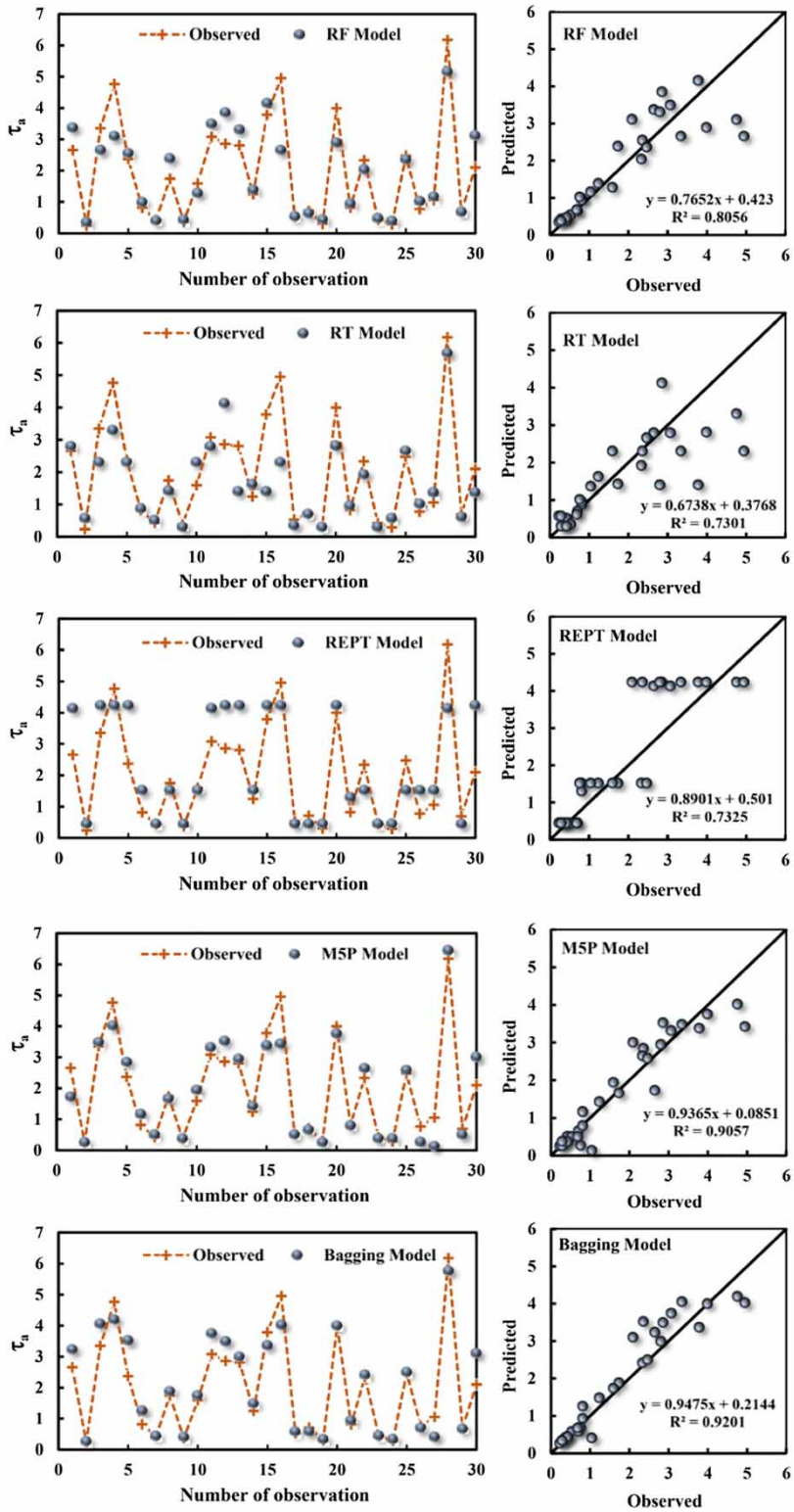

Figure 3 | Predicted and observed apparent shear stress time series and scatter plot presentation using all the applied data mining predictive models for the best input combination and over the test modeling phase.

the absolute error metrics, the integrated bagging-M5P model displayed the minimum values of RMSE = 0.46 and MAE = 0.32 in comparison with the other applied data mining predictive models. Another excellent performance metric was computed that measures the overestimation (i.e., negative PBIAS value) and underestimation (i.e., positive PBIAS value) (Moriasi *et al.* 2007). Based on the reported statistical results, the bagging-M5P and REPT showed overestimation values, with slightly higher magnitude of the bagging-M5P model; whereas the other models indicated underestimation behaviour. Other metrics (e.g., NSE and RSR) behaved similarly in indicating the superiority of the integrated bagging-M5P model over the other data mining models.

The time series and scatter plot presentation between the observed experimental and predicted shear (for the validation phase) are displayed in Figure 3. The context of the results presented in Figure 3 somewhat follows the reported statistical performance in Table 6. The bagging-M5P model performed the best prediction capacity in terms of agreement between the observed and predicted values. The scatter plot is useful in visually reflecting the variation of prediction values around the ideal line. It can be said that all models (except for REPT) have relatively good accuracy in predicting low apparent shear stress, particularly for $\tau_a^* < 2$. Most of the predicted values fall on the best fit line 45°. However, it is evident that the high apparent shear stresses were not very well predicted using RF, RT, REPT and M5P models. REPT model visibly did not perform well where all higher τ_a^* (that is >2) have a consistent value of 4. Having said this, even at lower range of $1 < \tau_a^* < 3$, the prediction values were fixed at ≈1.5 and 0.4 for $\tau_a^* < 1$. Although the RF, RT and M5P models may capture the dynamic fluctuating behaviour, the predicted values are either overestimated or underestimated in a rather large envelope along the line of agreement (with RT model observably having the widest stretch). Out of the five models discussed here, only the bagging-M5P model managed to mimic those high values with better accuracy.

Another important graphical visualisation generated is the box plot. Figure 4 shows the capability of the applied predictive models in the form of box plots. Based on the displayed statistical presentation with respect to the medians, quartiles and data ranges, the bagging-M5P model attained the best performance to the observed data pattern with slight variation as seen in box plots (Figure 4). This is followed by the M5P and RF models based on their capability to have a close prediction with the observed data.

Based on the attained predictability performance of the proposed data mining, clearly evidenced is the potential of data mining for simulating the apparent shear stress of compound channels. Data mining models revealed the opportunity to be integrated for channel design, management and sustainability. Finally, it is worth highlighting the possibility for future research application. Incorporating other experimental methods from the literature related to the apparent shear stress with the other channel characteristics might enhance the predictability of the proposed

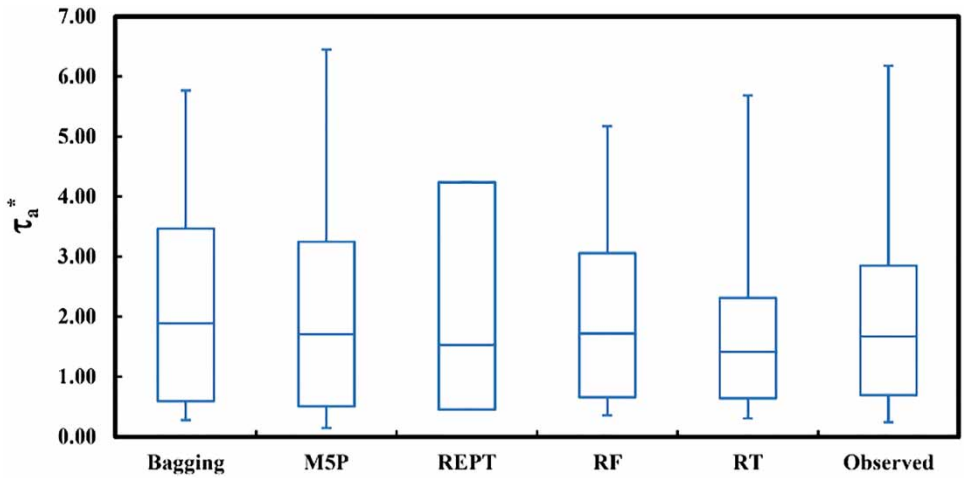

Figure 4 | Box plot visualisation of the applied data mining predictive models and the observed data of the experimental apparent shear stress.

data mining models. In addition, integrating nature-inspired optimisation algorithms (Yang 2013, 2014; Ajay Adithyan *et al.* 2018) is highly recommended as a prior stage for the prediction process where the most anticipated variables toward the shear value can be abstracted and fed as reliable input attributes for the prediction model.

CONCLUSION

Apparent shear stress is a vital component in prior channel design and thus it is highly emphasised that it should be quantified with accurate and reliable magnitude, in particular for flood design. In this study, five different versions of new data mining (i.e., RF, RT, REPT, M5P and bagging-M5P) were used. The models were constructed based on several channel properties' variables built in nine input combinations. The four models, RF, REPT, M5P and bagging-M5P behaved similarly to the best input combination 8 using B/b, $(H-h)/H$, n_f/n_c and h/b. On the other hand, the RT model attained its best results using input combination 6 based on H/B, h/b, B/b, BH/bh, $(H-h)/H$ and n_f/n_c. After selecting the best input combination, the predictions of apparent shear stress using all mentioned models were compared. It was found that the models' results improved significantly when a hybrid model was used, such that the bagging-M5P with lower statistical parameter values (R^2 of 0.92) demonstrated better performance than those of the RF, RT, REPT and M5P models. As traditional relations for calculating apparent shear stress require complete awareness of the velocity gradient and the compound channel's geometry, the results obtained from these methods are helpful for predicting the apparent shear stress in symmetric compound channels because the proposed method is based on knowledge of channel geometry and roughness.

ACKNOWLEDGEMENT

The first author acknowledges Universiti Kebangsaan Malaysia (Grant No. MI-2018-011) for financial support. The authors are also very grateful to the respected reviewers' constructive comments that enhanced the visualisation of the reported research.

REFERENCES

Ajay Adithyan, T., Sharma, V., Gururaj, B. & Thirumalai, C. 2018 Nature inspired algorithm. In: *Proceedings of the International Conference on Trends in Electronics and Informatics, ICEI 2017*, Tirunelveli, India, pp. 1131–1134. doi:10.1109/ICOEI.2017.8300889.

Al-Khatib, I. A., Abu-Hassan, H. M. & Abaza, K. A. 2013 Development of empirical regression-based models for predicting mean velocities in asymmetric compound channels. *Flow Measurement and Instrumentation* **33**, 77–87. doi:10.1016/j.flowmeasinst.2013.04.013.

Baird, J. I. & Ervine, D. A. 1984 Resistance to flow in channels with overbank flood-plain flow. In: *Proceedings of the 1st International Conference on Hydraulic Design in Water Resources Engineering: Channels and Channel Control Structures*, Southampton, UK, pp. 561–574.

Baker, R. S. J. D. & Yacef, K. 2009 The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining* **1**, 3–17. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8.

Bonakdari, H., Sheikh Khozani, Z., Zaji, A. H. & Asadpour, N. 2018 Evaluating the apparent shear stress in prismatic compound channels using the genetic algorithm based on multi-layer perceptron: a comparative study. *Applied Mathematics and Computation* **338**, 400–411. doi:10.1016/J.AMC.2018.06.016.

Bousmar, D. & Zech, Y. 1999 Momentum transfer for practical flow computation in compound channels. *Journal of Hydraulic Engineering* **125**, 696–706.

Breiman, L. 1996 Bagging predictors. *Machine Learning* **24**, 123–140. doi:10.1007/BF00058655.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32. doi:10.1023/A:1010933404324.

Chadalawada, J. & Babovic, V. 2019 Review and comparison of performance indices for automatic model induction. *Journal of Hydroinformatics* **21**. doi:10.2166/hydro.2017.078.

Chaurasia, V. & Pal, S. 2013 Early prediction of heart diseases using data mining techniques. *Carribean Journal of Science and Technology* **1**, 208–217.

Choi, S.-U., Choi, B. & Choi, S. 2015 Improving predictions made by ANN model using data quality assessment: an application to local scour around bridge piers. *Journal of Hydroinformatics* **17**, 977–989. doi:10.2166/hydro.2015.097.

Christodoulou, G. C. 1992 Apparent shear stress in smooth compound channels. *Water Resources Management* **6**, 235–247.

Cielen, A., Peeters, L. & Vanhoof, K. 2004 Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research* **154**, 526–532.

Dehling, H. G., Fleurke, S. R. & Külske, C. 2008 Parking on a random tree. *Journal of Statistical Physics* **133**, 151–157.

Devi, K. & Khatua, K. K. 2016 Prediction of depth averaged velocity and boundary shear distribution of a compound

channel based on the mixing layer theory. *Flow Measurement and Instrumentation* **50**, 147–157. doi:10.1016/j.flowmeasinst.2016.06.020.

Drmota, M. & Gittenberger, B. 1997 On the profile of random trees. *Random Structures and Algorithms* **10**, 421–451.

Elomaa, T. & Kaariainen, M. 2001 An analysis of reduced error pruning. *Journal of Artificial Intelligence Research* **15**, 163–187.

Ervine, D. A., Baird, J. I., Noutsopoulos, G. C., Hadjipanos, P. A., Bulman, R. B. & Holland, P. G. 1982 Rating curves for rivers with overbank flow. *Proceedings of the Institution of Civil Engineers* **73**, 849–855.

Fahimi, F., Yaseen, Z. M. & El-shafie, A. 2016 Application of soft computing based hybrid models in hydrological variables modeling: a comprehensive review. *Theoretical and Applied Climatology* **128**, 875–903. doi:10.1007/s00704-016-1735-8.

Fernandes, J. N., Leal, J. B. & Cardoso, A. H. 2015 Assessment of stage-discharge predictors for compound open-channels. *Flow Measurement and Instrumentation* **45**, 62–67. doi:10.1016/j.flowmeasinst.2015.04.010.

Haitao, G., Qingbao, Z. & Shoujiang, X. 2007 Rapid-exploring random tree algorithm for path planning of robot based on grid method. *Journal of Nanjing Normal University (Engineering and Technology Edition)* **2**, 58–61.

Hanspal, N. S., Allison, B. A., Deka, L. & Das, D. B. 2013 Artificial neural network (ANN) modeling of dynamic effects on two-phase flow in homogenous porous media. *Journal of Hydroinformatics* **15**, 540–554. doi:10.2166/hydro.2012.119.

Houichi, L., Dechemi, N., Heddam, S. & Achour, B. 2012 An evaluation of ANN methods for estimating the lengths of hydraulic jumps in U-shaped channel. *Journal of Hydroinformatics* **15**, 147–154. doi:10.2166/hydro.2012.138.

Huai, W., Chen, G. & Zeng, Y. 2013 Predicting apparent shear stress in prismatic compound open channels using artificial neural networks. *Journal of Hydroinformatics* **15**, 138–146.

Khatua, K. K., Patra, K. C. & Jha, R. 2010 Apparent shear stress in a compound channel. *ISH Journal of Hydraulic Engineering* **16**, 1–14.

Khosravi, K., Mao, L., Kisi, O., Yaseen, Z. M. & Shahid, S. 2018a Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile. *Journal of Hydrology* **567**, 165–179. doi:10.1016/j.jhydrol.2018.10.015.

Khosravi, K., Panahi, M. & Tien Bui, D. 2018b Spatial prediction of groundwater spring potential mapping based on an adaptive neuro-fuzzy inference system and metaheuristic optimization. *Hydrology and Earth System Sciences* **22**, 4771–4792. doi:10.5194/hess-22-4771-2018.

Kisi, O. 2008 The potential of different ANN techniques in evapotranspiration modelling. *Hydrological Processes* **22**, 2449–2460. doi:10.1002/hyp.6837.

Kisi, O., Ozkan, C. & Akay, B. 2012 Modelling discharge-sediment relationship using neural networks with artificial bee colony algorithm. *Journal of Hydrology* **428–429**, 94–103. doi:10.1016/j.jhydrol.2012.01.026.

Knight, D. W. & Demetriou, J. D. 1983 Flood plain and main channel flow interaction. *Journal of Hydraulic Engineering* **109**, 1073–1092. doi:10.1061/(ASCE)0733-9429(1983)109:8(1073).

Knight, D. W. & Hamed, M. E. 1984 Boundary shear in symmetrical compound channels. *Journal of Hydraulic Engineering* **110**, 1412–1430. doi:10.1061/(ASCE)0733-9429(1984)110:10(1412).

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**, 233–241.

Liaw, A. & Wiener, M. 2002 Classification and regression by random Forest. *R News* **2**, 18–22.

Melesse, A. M., Ahmad, S., McClain, M. E., Wang, X. & Lim, Y. H. 2011 Suspended sediment load prediction of river systems: an artificial neural network approach. *Agricultural Water Management* **95**, 855–866. doi:10.1016/j.agwat.2010.12.012.

Mohamed, W. N. H. W., Salleh, M. N. M. & Omar, A. H. 2012 A comparative study of reduced error pruning method in decision tree algorithms. In: *IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 2012*, Penang, Malaysia, pp. 392–397.

Moreta, P. J. M. & Martin-Vide, J. P. 2010 Apparent friction coefficient in straight compound channels. *Journal of Hydraulic Research* **48**, 169–177.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Binger, R. L., Harmel, R. D. & Veith, T. L. 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* **50**, 885–900. doi:10.13031/2013.23153.

Mutanga, O., Adam, E. & Cho, M. A. 2012 High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* **18**, 399–406.

Myers, W. R. C. 1978 Momentum transfer in a compound channel. *Journal of Hydraulic Research* **16**, 139–150.

Özbek, T., Koçyiğit, M. B., Koçyiğit, Ö. & Cebe, K. 2004 Comparison of methods for predicting discharge in straight compound channels using the apparent shear stress concept. *Turkish Journal of Engineering and Environmental Sciences* **28**, 101–109.

Pal, M. 2006 M5 model tree for land cover classification. *International Journal of Remote Sensing* **27**, 825–831.

Pandey, A. K., Pandey, P., Jaiswal, K. L. & Sen, A. K. 2013 A heart disease prediction model using decision tree. *IOSR Journal of Computer Engineering (IOSR-JCE)* **12**, 83–86.

Prasad, A. M., Iverson, L. R. & Liaw, A. 2006 Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**, 181–199. doi:10.1007/s10021-005-0054-1.

Prinos, P. & Townsend, R. D. 1984 Comparison of methods for predicting discharge in compound open channels. *Advances in Water Resources* **7**, 180–187. doi:10.1016/0309-1708(84)90016-2.

Quinlan, J. R. 1992 Learning with continuous classes. In: *5th Australian Joint Conference on Artificial Intelligence*, Singapore, pp. 343–348.

Rajaratnam, N. & Ahmadi, R. 1981 Hydraulics of channels with flood-plains. *Journal of Hydraulic Research* **19**, 43–60.

Rizzoli, A., Merler, S., Furlanello, C. & Genchi, C. 2002 Geographical information systems and bootstrap aggregation (bagging) of tree-based classifiers for Lyme disease risk prediction in Trentino, Italian Alps. *Journal of Medical Entomology* **39**, 485–492.

Sanikhani, H., Deo, R. C., Yaseen, Z. M., Eray, O. & Kisi, O. 2018 Non-tuned data intelligent model for soil temperature estimation: a new approach. *Geoderma* **330**, 52–64. doi:10.1016/j.geoderma.2018.05.030.

Sellin, R. H. J. 1964 A laboratory investigation into the interaction between the flow in the channel of a river and that over its flood plain. *La Houille Blanche* **7**, 793–802.

Sharafati, A., Khosravi, K., Khosravinia, P., Ahmed, K., Salman, S. A., Mundher, Z. & Shamsuddin, Y. 2019 The potential of novel data mining models for global solar radiation prediction. *International Journal of Environmental Science and Technology*. doi:10.1007/s13762-019-02344-0.

Sheikh Khozani, Z., Bonakdari, H. & Zaji, A. H. 2015 Application of a soft computing technique in predicting the percentage of shear force carried by walls in a rectangular channel with non-homogenous roughness. *Water Science and Technology* **73**, wst2015470. doi:10.2166/wst.2015.470.

Sheikh Khozani, Z., Bonakdari, H. & Zaji, A. H. 2016 Application of a genetic algorithm in predicting the percentage of shear force carried by walls in smooth rectangular channels. *Measurement* **87**, 87–98. doi:10.1016/j.measurement.2016.03.018.

Sheikh Khozani, Z., Bonakdari, H. & Ebtehaj, I. 2017a An analysis of shear stress distribution in circular channels with sediment deposition based on Gene Expression Programming. *International Journal of Sediment Research* **32**, 575–584. doi:10.1016/J.IJSRC.2017.04.004.

Sheikh Khozani, Z., Bonakdari, H. & Zaji, A. H. 2017b Estimating the shear stress distribution in circular channels based on the randomized neural network technique. *Applied Soft Computing* **58**, 441–448.

Sheikh Khozani, Z., Bonakdari, H. & Zaji, A. H. 2017c Efficient shear stress distribution detection in circular channels using Extreme Learning Machines and the M5 model tree algorithm. *Urban Water Journal* **14**, 999–1006. doi:10.1080/1573062X.2017.1325495.

Sheikh Khozani, Z., Bonakdari, H. & Ebtehaj, I. 2018 An expert system for predicting shear stress distribution in circular open channels using gene expression programming. *Water Science and Engineering* **11**, 167–176. doi:10.1016/j.wse.2018.07.001.

Solomatine, D. P. & Xue, Y. 2004 M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering* **9**, 491–501. doi:10.1061/(ASCE)1084-0699(2004)9:6(491).

Tang, X. 2017 An improved method for predicting discharge of homogeneous compound channels based on energy concept. *Flow Measurement and Instrumentation* **57**, 57–63. doi:10.1016/j.flowmeasinst.2017.08.005.

Tao, H., Diop, L., Bodian, A., Djaman, K., Ndiaye, P. M. & Yaseen, Z. M. 2018 Reference evapotranspiration prediction using hybridized fuzzy model with firefly algorithm: regional case study in Burkina Faso. *Agricultural Water Management* **208**, 140–151.

Tapoglou, E., Varouchakis, E. A., Trichakis, I. C. & Karatzas, G. P. 2019 Hydraulic head uncertainty estimations of a complex artificial intelligence model using multiple methodologies. *Journal of Hydroinformatics*. https://doi.org/10.2166/hydro.2019.137

Vandamme, J., Meskens, N. & Superby, J. 2007 Predicting academic performance by data mining methods. *Education Economics* **15**, 405–419.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, Amsterdam, The Netherlands.

Witten, I. H., Frank, E. & Hall, M. A. 2011 *Data Mining: Practical Machine Learning Tools and Techniques, Annals of Physics*. Morgan Kaufmann, Los Altos, CA, USA. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823> 3.3.CO;2-C.

Wormleaton, P. & Merrett, D. 1990 An improved method of calculation for steady uniform flow in prismatic main channel/flood plain sections. *Journal of Hydraulic Research* **28**, 157–174.

Wormleaton, P. R., Allen, J. & Hadjipanos, P. 1982 Discharge assessment in compound channel flow. *Journal of the Hydraulics Division* **108**, 975–994.

Yang, X.-S. 2013 Metaheuristic optimization: Nature-inspired algorithms and applications. In: *Artificial Intelligence, Evolutionary Computing and Metaheuristics*. Springer, Berlin, Heidelberg, pp. 405–420.

Yang, X. S. 2014 *Nature-Inspired Optimization Algorithms*. Elsevier, London, UK. doi:10.1016/C2013-0-01368-0.

Yaseen, Z. M., El-Shafie, A., Afan, H. A., Hameed, M., Mohtar, W. H. M. W. & Hussain, A. 2015 RBFNN versus FFNN for daily river flow forecasting at Johor River, Malaysia. *Neural Computing and Applications* **25**, 1533–1542. doi:10.1007/s00521-015-1952-6.

Yaseen, Z. M., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J. & El-shafie, A. 2016 Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. *Journal of Hydrology* **542**, 603–614. doi:10.1016/j.jhydrol.2016.09.035.

First received 9 February 2019; accepted in revised form 27 May 2019. Available online 18 June 2019