Predicting stable gravel-bed river hydraulic geometry: A test of novel, advanced, hybrid data mining algorithms

Khabat Khosravi[1], Zohreh Sheikh Khozani*[2], James R.Cooper[3]

1- Department of Watershed Management Engineering, Ferdowsi University of Mashhad, Mashhad, Iran. Khabat.khosravi@gmail.com

2- Institute of Structural Mechanics, Bauhaus Universität Weimar, 99423 Weimar, Germany. (zohreh.khozani.sheikh@uni-weimar.de)

3- Department of Geography and Planning, University of Liverpool, Liverpool, UK. James.Cooper@liverpool.ac.uk

*Corresponding author: Zohreh Sheikh Khozani (zohreh.khozani.sheikh@uni-weimar.de)

Abstract

Accurate prediction of stable alluvial hydraulic geometry, in which erosion and sedimentation are in equilibrium, is one of the most difficult but critical topics in the field of river engineering. Data mining algorithms have been gaining more attention in this field due to their high performance and flexibility. However, an understanding of the potential for these algorithms to provide fast, cheap, and accurate predictions of hydraulic geometry is lacking. This study provides the first quantification of this potential. Using at-a-station field data, predictions of flow depth, water-surface width and longitudinal water surface slope are made using three standalone data mining techniques - , Instance-based Learning (IBK), KStar, Locally Weighted Learning (LWL) - along with four types of novel hybrid algorithms in which the standalone models are trained with Vote, Attribute Selected Classifier (ASC), Regression by Discretization (RBD), and Cross-validation Parameter Selection (CVPS) algorithms (Vote-IBK, Vote-Kstar, Vote-LWL, ASC-IBK, ASC-Kstar, ASC-LWL, RBD-IBK, RBD-Kstar, RBD-LWL, CVPS-IBK, CVPS-Kstar, CVPS-LWL). Through a comparison of their predictive performance and a sensitivity analysis of the driving variables, the results reveal: (1) Shield stress was the most effective parameter in the prediction of all geometry dimensions; (2) hybrid models had a higher prediction power than standalone data mining models, empirical equations and traditional machine learning algorithms; (3) Vote-Kstar model had the highest performance in predicting depth and width, and ASC-Kstar in

estimating slope, each providing very good prediction performance.. Through these algorithms, the hydraulic geometry of any river can potentially be predicted accurately and with ease using just a few, readily available flow and channel parameters. Thus, the results reveal that these models have great potential for use in stable channel design in data poor catchments, especially in developing nations where technical modelling skills and understanding of the hydraulic and sediment processes occurring in the river system may be lacking.

Keywords: gravel-bed rivers, hydraulic geometry, modelling, artificial intelligence, data mining, machine learning.

1. Introduction

Alluvial rivers form their own geometry in plan and cross-section, adjusting according to flow and sediment transport conditions. A river in a state of equilibrium over a specified period of time is said to be in regime or stable (Singh and Zhang 2008). This state of dynamic equilibrium occurs if the sediment transport rate is approximately equal to the upstream sediment supply, meaning that channel dimensions/geometry are maintained over this time period. Channel stability analysis involves analyzing how a channel adjusts its hydraulic geometry in response to changes in water and sediment discharge using river channel adjustment approaches (Gholami et al. 2017). This geometry is specified in terms of river flow width, depth, velocity and slope, and understanding how these hydraulic parameters vary with other variables, such as discharge, shear stress and median bed grain-size, is of paramount importance in stable channel design. The change in geometry is considered either over time at one cross-section (called at-a-station hydraulic geometry), focussing on temporal variations in the river geometry, or along the river length (called downstream hydraulic geometry). To design a stable geometry, accurate prediction of channel form in relation to the temporal and spatial variation in river hydraulics and sediment transport dynamics, is therefore required.

Thus far, various methods have been used to develop functional relationships for predicting stable hydraulic geometry dimensions. These approaches can be broadly classified into three methods, each using the same basic assumption of steady and uniform flow to achieve channel equilibrium. First, empirical equations of the regime have been obtained from the statistical rule/regression analysis of channel geometry data from different rivers (Thomas Blench 1952; Bray 1982; Hey and Thorne 1986a; Leopold and Wolman 1957; Wolman 1954). In these equations, flow discharge, bed shear stress and bed-grain diameters have been considered as the most effective parameters to predict the geometry of stable rivers (Deshpande and Kumar 2012; Parker et al. 2007). The major drawbacks of this approach is the lack of hydraulic, theoretical basis to the equations (Eaton and Church 2007; Hey and Thorne 1986b), and consequently low generalization and limited accuracy when applied to rivers in conditions that fall outside those used in the development of the equations (Bose 1936; Stevens and Nordin 1987). Another shortcoming of this method is that the equations are most often only developed only with flow discharge and bed-grain diameter as driving variables, while other important variables such as sediment transport rate or sediment concentration are neglected.

Secondly, theoretical and analytical models have been developed by river engineers and geomorphologists. For example, many studies have developed models based on regime theory (T Blench 1969; Hey and Thorne 1986a; Huang and Nanson 1998; Lane 1957), quantifying the critical control of bed and bank materials on river channel form either through using a 'silt factor' or by developing regime relations based on the character of these materials. However, no study has proposed a universally accepted rational theory, nor defined universal formulations for its parameters (Gleason 2015). Analytical models have been developed by solving the governing hydraulic equations, most often based on field observations (Henderson 1961). For example, Julien and Wargadalam (1995) created analytical equations for downstream hydraulic geometry as a function of flow discharge, sediment size, Shields number and streamline deviation angle. They argued these models are more accurate and reliable than empirical equations because they are based on the physics and theory of the process. Afzalimehr *et al*. (2010) tested

78 the performance of these analytical equations against empirical equations based on 85 at-a-station datasets
79 from Iranian rivers and found contrasting results. These contrasting results were reported because the
80 empirical equations were only tested with the datasets from which they were developed. This paper also
81 found that the grain size and the Shields parameter need not be taken into account when evaluating the
82 width and depth of an alluvial channel at a site.

83 Thirdly, numerical models have been developed based on the solution of flow friction equations, the law
84 of continuity, sediment transport capacity, and in some cases, the stability of the river banks (Chang 1980;
85 Millar 2005; White 1982). Although analytical equations provide a stronger logical framework for
86 examining possible changes in prevailing conditions (Ferguson 1986), the prediction performance of
87 numerical solutions can be similar to those of empirical models (Millar, 2005). Examples of numerical
88 equations for stable hydraulic geometry prediction are provided in commonly-used software, such as
89 HEC-RAS (Mehta et al. 2013; Shelley and Parr 2009). Although this type of model is developed based on
90 the physics of the process, they require lots of data to provide good model performance, and calibration is
91 difficult and time-consuming. Therefore, new ways to predict stable hydraulic geometry, that are
92 computationally simple, flexible, reliable and require small datasets, are required.

93 Since the 1980s, several Artificial Intelligence (AI) algorithms have been developed successfully to solve
94 hydraulic problems and are gaining more attention due to their high performance and flexibility. These
95 algorithms utilize data with different scale and are insensitive to missing data and the length of data. One
96 of the most commonly-used AI models in hydraulics is the Artificial Neural Network (ANN). This
97 algorithm has been used by many researchers to estimate hydraulic parameters, such as bed shear stress,
98 as well as inform the design of alluvial irrigation canals (Mohamed 2013; Sheikh Khozani et al. 2017;
99 Wan Mohtar et al. 2018), rainfall-runoff modelling (Antar et al. 2006), rainfall prediction (Mislan et al.
100 2015) and water quality assessment (Cuest Cordoba et al. 2014). ANN models can implicitly identify
101 complicated, nonlinear connections between independent and dependent parameters and can detect all
102 potential interactions across the predictor parameters. Given the nonlinear relationship between hydraulic

and sediment transport parameters, ANN models have thus been used in the prediction of channel geometry. For example, Khadangi et al. (2009) predicted three channel parameters (width, depth, and slope) using data collected from 371 rivers, and examined the prediction performance of two different ANNs structures. Their results showed good performance in the evaluation phase compared with measured values, performing better in estimating channel width than depth and slope. Mohamed (2013) applied an ANN model based on a back-propagation algorithm to estimate the wetted perimeter, hydraulic radius and water surface slope of 61 Egyptian irrigation canals. The prediction performance of Mohamed's (2013)model was compared against three empirical equations frequently used to predict hydraulic geometry. The ANN model had superior performance in all cases. Gholami *et al*. (2017) showed this was also the case for gravel-bed rivers. In another study, Tahershamsi *et al*. (2012) investigated the performance of multi-layer perceptron (MLP) and Radial Basis Function (RBF) models to forecast the width of alluvial channels. Both models had good prediction performance. However, despite these promising results, ANN models have slow coverage speed during the training procedure, and model performance can decrease significantly if the training dataset is not carefully chosen (i.e. when the testing dataset is out of range of the training dataset; Choubin et al., 2018).

Evolutionary models have gained a lot of attention in recent years (Ferreira 2001; Wang et al. 2016). In particular, Gene Expression Programming (GEP) is recognised as a strong and problem-independent technique for multivariate optimization (Ferreira 2002; Wu et al. 2013). Shaghaghi *et al*. (2018) applied three Non-linear Regression (NLR), GEP and, Generalized Structure of Group Method of Data Handling (GS-GMDH) models to estimate alluvial channel width, depth and slope. The Group Method of Data Handling (GMDH) model relates to the deterministic self-organizing method group, where the principle of a black box, connectionism and induction is used (Anastasakis and Mort 2001). Shaghaghi *et al*. (2018) investigated the impact of different input variable combinations and found that the most effective parameters in estimating width and depth were discharge and mean particle size, while for channel slope, the Shields parameter was the most effective. They compared the accuracy of their three models and

deduced that GEP and GS-GMDH had better predictive performance than the NLR model. However the weakness of the GMDH algorithm lies in its fixed configuration, using a deterministic approach to find the optimal partition of datasets and parameters (Robinson 1998). Sheikh Khozani *et al*. (2017) predicted shear stress distribution in circular channels by applying GEP and evaluating the performance of different input combinations. Their model showed better performance in estimating shear stress distribution than a Shannon entropy-based equation presented by Sterling and Knight (2002). Noori *et al*. (2016) compared ANN, Adaptive Neuro-fuzzy Inference system (ANFIS), and Support Vector Machine (SVM) models for predicting the longitudinal dispersion coefficient in rivers and reported that SVM had a higher performance followed by ANFIS and ANN. The ANFIS algorithm, however, suffers from a large number of model operators, each of which needs to be set accurately, especially the weights of membership function. Although SVM has a higher prediction power, the model can be time-consuming to train, since it is susceptible to hyper-parameter selection (Ahmad et al. 2018), and choosing the best kernel is problematic, reducing its wider application.

Consequently, a new form of AI, data mining, has been applied in the fields of hydrology and hydraulics to overcome the aforementioned weaknesses in traditional AI models. Some of these new algorithms, such as tree-based models [i.e. Random Tree (RT), Random Forest (RF), M5 Prime (M5P), Reduced Error Pruning Tree (REPT)], re-sampling algorithm (i.e. Bootstrap Aggregation, also called bagging), and other algorithms such as Random Subspace, and k nearest neighbor (IBK), were used to estimate apparent shear stress in a compound river cross section (Sheikh Khozani et al. 2019), suspended sediment transport (Khosravi et al. 2018), nitrate and strontium concentrations in groundwater (Bui et al. 2020). These data mining algorithms have higher predictive power than traditional AI models. For example, Hussain and Khan (2020) found RF had a 17.8 % and 33.6 % higher performance than ANN and SVM for predicting river streamflow. Further, Shamshirband *et al*. (2020) demonstrated the superiority of M5P over SVM for standardized streamflow index prediction. Also Khosravi *et al*. (2019) showed that data mining algorithms outperform standalone ANFIS algorithms in the prediction of reference evaporation, while

optimized ANFIS using metaheuristic algorithms performed slightly better than standalone data mining algorithms. Also some researchers have reported that hybridized algorithms improve the performance of standalone algorithms, not only for traditional AI algorithms, but also for data mining models in the prediction of water quality index and bedload transport rate (Bui et al. 2020a; Bui et al. 2020b; Khosravi et al. 2020). However, these new data mining algorithms have yet to be applied for the prediction of hydraulic geometry. Thus, a significant gap exists in understanding the potential of these data mining algorithms, and in the identification of the most flexible and accurate algorithm.

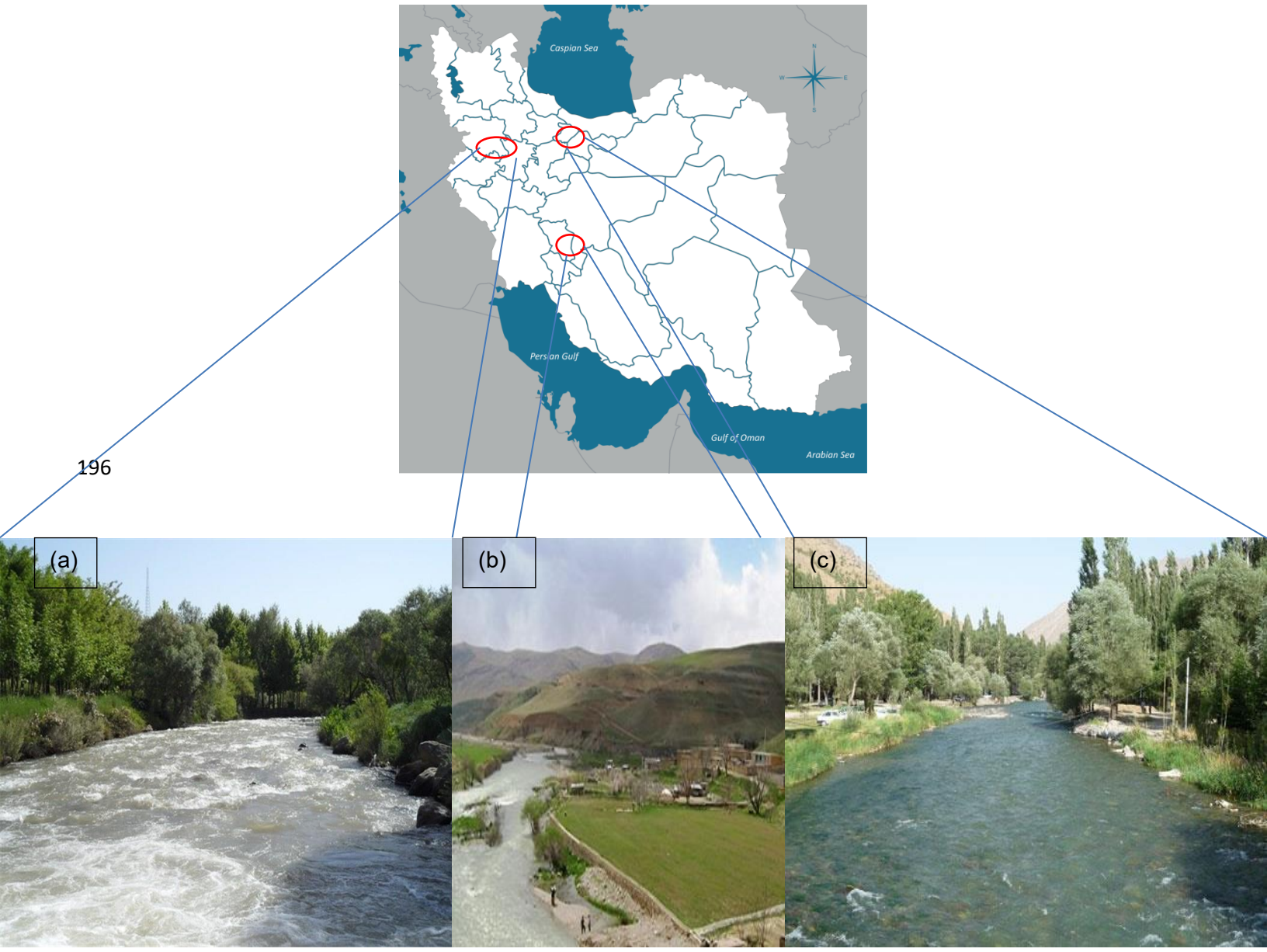
The present paper, therefore, aims to fill this gap in understanding by achieving the following objectives: (1) produce predictions of the three main hydraulic geometry parameters (mean flow depth, water-surface width and longitudinal water surface slope) using three standalone data mining techniques, namely Instance-based Learning (IBK), KStar, Locally Weighted Learning (LWL), along with four types of novel hybrid algorithms in which the standalone models are trained with Vote, Attribute Selected Classifier (ASC), Regression by Discretization (RBD), and Cross-validation Parameter Selection (CVPS) algorithms (Vote-IBK, Vote-Kstar, Vote-LWL, ASC-IBK, ASC-Kstar, ASC-LWL, RBD-IBK, RBD-Kstar, RBD-LWL, CVPS-IBK, CVPS-Kstar, CVPS-LWL; (2) compare the predictive power of these data-driven models; and (3) perform a sensitivity analysis of the driving variables used in each model. The performance of these algorithms is tested for the following reasons: (1) IBK can adapt to previously unseen data, storing a new instance or throwing an old instance away, making it potentially superior to other methods of machine learning. (2) The KStar algorithm uses entropic measure based on probability of transforming instance into another by randomly choosing between all possible transformations (Madhusudana et al. 2016). (3) LWL improves the overall performance of regression methods by adjusting the capacity of the models to the properties of the training data in each area of the input space (Reyes et al. 2018). (4) Vote algorithm can find the majority of a sequence of the elements by using linear time and constant space. Also, this algorithm is important for ultra-reliable system which are based on the multi-channel computation paradigm (Parhami 1994). (5) ASC model benefit from three main

components including base classifier, evaluator and search algorithm in its structure (Thornton et al. 2013). (6) In RBD method, the estimated value is the probable value of the mean class value for each discretized interval, according to the estimated probabilities for each interval (Frank and Bouckaert 2009). (7) CVPS is a technique of selecting parameters using cross-validation sampling. To the best of our knowledge, this study is the first to apply these hybridized algorithms in any branch of geoscience. The research offers new insight into which data mining algorithms offer the potential to provide relatively cheap and fast predictions of hydraulic geometry in situations when understanding of the physical processes at play may not be well understood.

2. Material and methods

2.1. Datasets

The paper uses a dataset compiled by Afzalimehr *et al*. (2010) for three stable gravel-bed rivers in Iran: Karaj river in Alborz Province, Behesht-Aabad river in Charmahal-and-Bakhtiari Province and Gamasiab River in Kermanshah province (Figure 1). This dataset includes measurements of flow discharge (Q), median sediment diameter (d_{50}), Shields number (τ^*) at 85 cross-sections (Table 1), used as inputs to predict hydraulic geometry. This geometry is defined by water-surface width (w), mean flow depth (h) and longitudinal water surface slope (S). Flow discharge in a cross section was estimated through three to five velocity profiles, with each profile containing 13-16 velocity measurements at different heights above the sediment bed, totalling 425 profiles



196

197

198 Fig. 1. Map and illustrative photographs of the three studied rivers: (a) Gamasiab river, (b) Behesht-Abad
199 river and (c) Karaj river.

200 and 6000 point velocities. At each cross section the top width (channel width at the water surface) was
201 measured along with the flow depth at 0.5 m intervals across the channel. The mean flow depth at each
202 cross section was calculated by dividing the cross-sectional area by this top width. The Wolman's walk
203 approach (Wolman 1954) was used to measure the bed sediment size distribution. The longitudinal water

surface slope was determined by dividing the difference in water surface elevations between two cross sections along the central axis of the reach. . Shields parameter was computed based on the following equation:

$$\tau^* = \frac{\tau}{(\rho_s - \rho)gd_{50}} \quad (1)$$

where are τ is the shear stress [-], ρ_s is sediment density [-], ρ is water density [kg m^{-3}] and g is gravitational acceleration [m s^{-2}]. Shear stress was calculated as follows:

$$\tau = (\rho.v^*)^2 \quad (2)$$

where v^* is the shear velocity [m s^{-1}] which was calculated as $v^* = (ghS)^{0.5}$. More information about the data collection methodology can be found in Afzalimehr *et al.* (2010).

Table 1. Descriptive statistics of the training and testing dataset

	Training dataset						Testing dataset					
	max	min	mean	Std	SK	K	max	min	mean	Std	SK	K
Q (m^3/s)	5.810	0.500	2.245	1.430	0.374	-0.709	5.300	0.550	2.299	1.338	0.181	-0.760
d_{50} (m)	0.130	0.004	0.032	0.031	1.316	1.080	0.094	0.004	0.031	0.029	0.819	-0.612
τ^* (-)	0.814	0.000	0.121	0.170	2.059	4.393	0.481	0.001	0.105	0.139	1.704	2.124
S (-)	0.028	0.0001	0.006	0.005	2.379	8.151	0.016	0.0001	0.005	0.003	1.686	3.774
h (m)	0.570	0.180	0.344	0.094	0.313	-0.877	0.570	0.230	0.337	0.085	0.937	0.738
w (m)	27.000	5.500	14.582	5.507	0.401	-0.749	23.000	7.000	14.072	4.748	0.217	-1.054

where max = maximum, min = minimum, Std = standard deviation, SK = skewness and K = kurtosis

2.2. Dataset preparation and sample size

The 85 datasets were split into two subgroups; 70% of the datasets were selected randomly to be used as training data for model development and the remaining 30% was applied as testing data for model validation. There is no agreement in the literature on this ratio. Some have used ratios of 80:20 (Zounemat-Kermani et al. 2019), and 75:25 (Hooshyaripor et al. 2014). Palani et al., (2008) and Barzegar et al. (2016) declared that the testing dataset should represent approximately 10 – 40% of the size of the

whole dataset. Also, Kisi et al. (2019) showed that by increasing the length of the training dataset from 50% to 75%, the modelling performance increased. With these considerations in mind a 70:30 ratio is the most commonly used (Bui et al. 2018; Chen et al. 2017; Taheri et al. 2019).

2.3. Model input, calibration and sensitivity analysis

Flow discharge, median sediment diameter, and Shields number are the three most important and widely used variables which affect stable river geometry (Deshpande and Kumar 2012; Gholami et al. 2017; Parker et al. 2007; Shaghaghi et al. 2018). These parameters were therefore used as an input in each model to predict the top width, flow depth and longitudinal slope at each river cross-section.

There are two main steps in using AI algorithms: (i) determination of the best input variable combination; and (ii) identifying the operator's optimum values. Each input variable has a differing impact on these hydraulic geometry parameters. Thus different input combinations were constructed and examined to find the most effective input combination (Table 2). These combinations were constructed by beginning with the variable with the highest Pearson correlation coefficient (*PCC)* (a measure of linear correlation between two sets of data) (τ^* for h and S, and d_{50} for w), and then exploring all other input combinations.. The effect of each input variables on the output was examined through a sensitivity analysis. To explore the most effective combination, the models were implemented using default models operators. Their effectiveness was assessed using Root Mean Square Error (*RMSE*); the lower the *RMSE*, the higher the effectiveness of the input combination.

Table 2. Different input combinations constructed to explore the most effective combination for model calibration.

No.	Input	Output		No.	Input	Output
1	τ^*	h, S		1	d_{50}	w
2	τ^*, Q	h, S		2	τ^*, Q	w

3	τ^*, d_{50}	*h, S*		3	τ^*, d_{50}	*w*
4	Q, d_{50}	*h, S*		4	Q, d_{50}	*w*
5	τ^*, Q, d_{50}	*h, S*		5	τ^*, Q, d_{50}	*w*

Along with data quality, length of data, and input variable choice, the calibration of model operator values has an important impact on prediction performance. There are no optimum operator values which work globally for model calibration. Hence, to enhance the prediction power of each algorithm, these values were set after the determination of the best input combination. At first, default values of each operator were considered, and then based on this result, lower and higher values were selected to find the optimum value. The most widely used approach of trial and error was performed in Waikato Environment for Knowledge Analysis (WEKA 3.9) software. The optimum operator values were achieved by minimizing the Root Mean Square Error (*RMSE*) during the testing phase.

2.4. Model descriptions

2.4.1. Iinstance-based Llearning (IBK)

Instance-based Learning, also known as K-Nearest Neighbor classification, is a lazy learning algorithm, well known for its ability to recognise data patterns. The algorithm applies a relatively simple method to store training data and identify new undefined data by measuring the distance between similar recorded samples. The IBK utilises an election system to determine the class of new samples; the number of votes defines the *k* value. The distance is defined after the *k* value is determined. The application of the IBK algorithm involves three steps. (1) reading the *k* value, distance type and test data, (2) finding the *k* nearest neighbor to the test data, and (3) setting the maximum label class to the test data. The WEKA Machine Learning Software (Witten et al. 2016) was utilized for running the IBK algorithm.

2.4.2. Kstar

The Kstar algorithm, first introduced by Cleary and Trigg (1995), is another type of lazy algorithm, which uses an entropy-based distance function to transform one sample probability to another by selecting arbitrarily all feasible transformations. The classification with Kstar is performed by summing the new instance probabilities to all the members of a group. This classification must be achieved for the other groups in order to eventually choose the one with the highest probability (Cleary and Trigg 1995). For missing values, Cleary and Trigg (1995) assumed that the likelihood of transformation to these values is the average of the likelihood of transformation to each of the defined attributes in the whole dataset. The algorithm is defined as follows. Consider I as a set of instances and T as a set of transformations on I (Clearly *et al*., 1995). Each instance ($t \in T$) maps to another instance as $t: I \rightarrow I$. T has a special member ω to map samples to themselves ($\omega(a) = a$). Let P be the set of all prefix codes from T^*, which is terminated by ω. The T^* members define a transformation on I:

$$\bar{t}(a) = t_n\big(t_{n-1}(\dots t_1(a) \dots)\big) \quad where\ \bar{t} = t_1, \dots t_n \tag{3}$$

The probability function of T^* is defined as p:

$$0 \leq \frac{p(\bar{t}u)}{p(\bar{t})} \leq 1 \tag{4}$$

$$\sum_u p(\bar{t}u) = p(\bar{t}) \tag{5}$$

$$\sum_{\bar{t} \in P} p(\bar{t}) = 1 \tag{6}$$

Furthermore, the probability of the entire path from such an instance to b is defined as P^*:

$$P^*(b|a) = \sum_{\bar{t} \in P: \bar{t}(a) = b} p(\bar{t}) = 1 \tag{7}$$

2.4.3. Locally Weighted Learner (LWL)

Locally weighted Learner is another lazy learning algorithm, The algorithm has an optimal convergence speed and its minimum performance is higher than all possible linear regressions (Stone 1982). The LWL

method is able to manage a wide range of data distribution types and can prevent boundary and cluster effects (Hastie and Loader 1993). The LWL depends on the distance function, which is used to recover the nearest neighbours of a given query example (Atkeson *et al.*, 1997). The method also depends on a smoothing parameter and weighting function. The weighting function calculates the weight of the sample neighbor query. This function should have a maximum value at a distance of zero, and as the distance increases, the performance slowly decreases., A bandwidth parameter (k) acts as the smoothing parameter, determining the size or the range in which generalisation is accomplished. This parameter is defined as follows.

Let a non-linear system be defined as (Arif et al. 2001):

$$y(k) = z\big(x(k), u(k)\big) \tag{8}$$

$$u_d(k) = z^{-1}\big(x_d(k), y_d(k)\big) \tag{9}$$

in which a non-linear function is defined as $z(.)$, the states as $x_d(k)$, and the output parameter as $y_d(k)$.

2.4.4. Vote

The meta algorithm Vote was used to train the IBK, Kstar, and LWL models and produced three hybrid models, Vote-IBK, Vote-Kstar, and Vote-LWL. This algorithm combined each basic-level classifier using a vote approach. The simplest voting approach is majority voting, in which the basic-level classifier casts one vote for its predictions. The instance is categorised into the class which obtains the most votes. For the situation where class probability distributions are estimated by the basic-level classifiers, the plurality voting method is modified (Dietterich 1997), defined as follows Assume $P_s(x)$ is the estimated class probability distribution by the basic-level classifier S on sample x. The probability distribution components restored by the basic-level classifiers are summed to reach the probability distribution class of meta-level voting classifier as:

$$P_{s(ML)}(x) = \frac{1}{|s|}\sum P_s\,(x) \quad (10)$$

2.4.5. Attribute Selection Committees (ASC)

The Attribute Selected Classifier algorithm was applied to train the IBK, Kstar, and LWL models and produced three hybrid models, ASC-IBK, ASC-Kstar, and ASC-LWL. The Attribute Selected Classifier is an ensemble technique, generally considered as a black-box form of classifier. The structure of ensemble classifiers is such that much information can be obtained by using bi-product data (Gislason et al. 2006). making it possible to determine an attribute based on the training set before learning the predefined classification.

The advantages of applying the attribute subsets in ensemble learning are, according toThornton et al. (2013): (1) reduction in the dimension of the data, which decreases the effect of the "curse of dimensionality"; (2) decrease in the connection between classifiers through training them on several characteristics; and (3) improvement in the classifiers output of the ensemble.

2.4.6. Regression by Discretization (RBD)

The Regression by Discretization algorithm was used to train the standalones models and produce the following hybrid models: RBD-IBK, RBD-Kstar, and RBD-LWL. This algorithm is a meta classifier technique, based on conditional density prediction via the class probabilities. The output parameter is discretized in non-overlapping periods which are called "bins". These bins can be produced of equal frequency and equal width. If a bin is defined as k_y which consist of the output value y, the whole number of output values in the training stage is n, the number of output values in bin m is n_m and $p(k_y|X)$ is the estimated probability of specified class X forecasted from the class probability predictor. The weight, for a specified output value y_i in case X, was computed as:

$$w(y_i|X) = m\frac{p(k_{yi}|X)}{m_{k_{yi}}} \quad (11)$$

The weight $w(y_i|X)$ can be seen as an approximation of the likelihood of a future target predicted value correlated with X being close to y_i, based on the class probability prediction model derived from discrete training data.

2.4.7. Cross-Validation Parameter Selection (CVPS)

The Cross-Validation Parameter Selection algorithm was used to train the standalone models IBK, Kstar, and LWL and produce the following three hybrid models: CVPS-IBK, CVPS-Kstar, and CVPS-LWL. Cross-validation is one of the most widely used statistical methods for assessing predictor model performance by using an *a priori* modelling procedure (Stone 1974). The method is based on data splitting; a portion of the data is used to fit each competing method and the remaining data is used to calculate the predictive model's performance, and the model with the best overall efficiency is chosen. Using continuous cycles, the training and validation sets are cross-overed so that each data point has a chance of being verified against all other data points. The CVPS algorithm is one of the meta-classifier techniques which was extended in WEKA environment by Garg and Khurana (2014) and is used to improve the prediction power of standalone algorithms through hybridization.

2.5. Model validation

Five frequently used metrics for assessing model performance were applied: coefficient of determination (R^2), Root Mean Square Error (*RMSE*), Mean Absolute Error (*MAE*), Nash-Sutcliffe Efficiency (*NSE*) and percent bias (*PBIAS*). These metrics were calculated as follows (Dawson et al. 2007; Legates and McCabe Jr 1999; Moriasi et al. 2007):

$$R^2 = \left(\frac{\sum_{i=1}^{n}(X_o - \overline{X}_o)(X_e - \overline{X}_e)}{\sqrt{\sum_{i=1}^{n}(X_o - \overline{X}_o)^2 \sum_{i=1}^{n}(X_e - \overline{X}_e)^2}}\right)^2 \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_e - X_o)^2} \quad (13)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_e - X_o\right| \quad (14)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(X_e - X_o)^2}{\sum_{i=1}^{n}(X_o - \overline{X}_o)^2} \quad (15)$$

$$PBIAS = \left(\frac{\sum_{i=1}^{n}(X_o - X_e)}{\sum_{i=1}^{n}X_e}\right) * 100 \quad (16)$$

where, X_o and X_e are observed and predicted values, $\overline{X}_0$ and $\overline{X}_e$ are mean observed and predicted values, respectively, and n is the number of data points. The performance classification of the model evaluation metrics is shown in Table 3. The *PBIAS* metric reports over- (*PBIAS* < 0) or under-prediction (*PBIAS* > 0).

Table 3. Performance classification of the model evaluation metrics

Objective function	Value range	Performance classification	References
R^2	$0.7 < R^2 < 1$ $0.6 < R^2 < 0.7$	Very good Good Satisfactory	Moriasi et al. (2007); Ayele et

	$0.5 < R^2 < 0.6$	Unsatisfactory	al. (2017)
	$R^2 < 0.5$		
RMSE		The lower the *RMSE*, the better the model performance	Dawson et al. (2006)
MAE		The lower the *MAE*, the better the model performance	Dawson et al. (2006)
NSE	$0.75 < NSE \leq 1.00$	Very good	Moriasi et al. (2007); Boskidis et al. (2012)
	$0.65 < NSE \leq 0.75$	Good	
	$0.50 < NSE \leq 0.65$	Satisfactory	
	$0.4 < NSE \leq 0.50$	Acceptable	
	$NSE \leq 0.4$	Unsatisfactory	
PBIAS	$PBIAS < \pm 10$	Very good	Legates et al. (1999)
	$10 \leq \lvert PBIAS \rvert < 15$	Good	
	$15 \leq \lvert PBIAS \rvert < 25$	Satisfactory	
	$PBIAS \geq \pm 25$	Unsatisfactory	

For a visual assessment of the applied models, boxplots of observed and predicted values were compared (Figure A, Supplementary material). These were used to shows how well a model predicts extreme, median and quartile values.

3. Results

The *PCC* values in Figure 2 show the level of correlation between input variables and hydraulic geometry parameters. First, Shields's stress had the highest correlation with longitudinal slope (*PCC* = 0.85) followed by flow depth (*PCC*=0.29) and width (*PCC*=0.01). Second, median sediment diameter e had the highest correlation with width (*PCC* = -0.39), followed by slope (*PCC* = -0.32) and depth (*PCC* = 0.08). Finally, discharge had the highest correlation coefficient with longitudinal slope (*PCC*=0.53) followed by width (*PCC*=0.2) and depth (*PCC*=0.07).

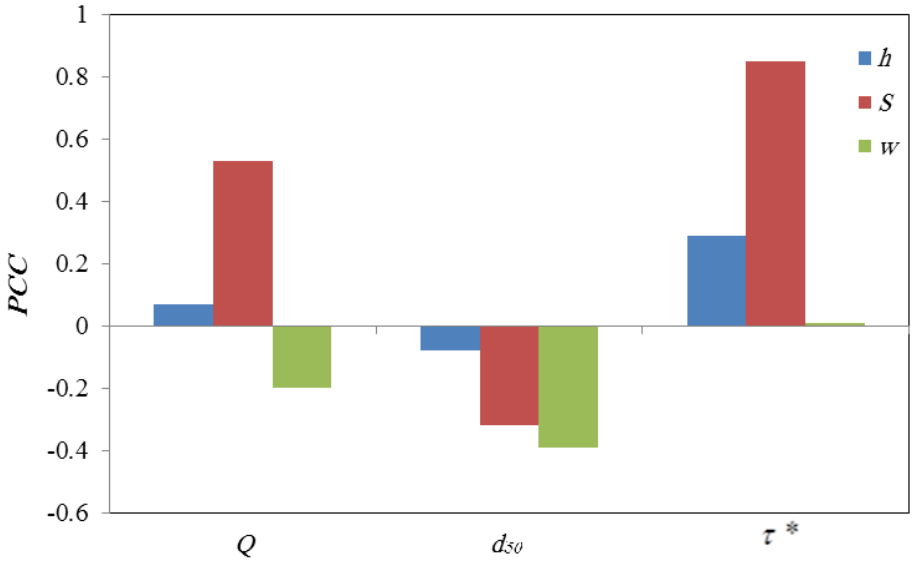


Fig 2. Pearson correlation coefficient (*PCC*) between input variables and hydraulic geometry parameters

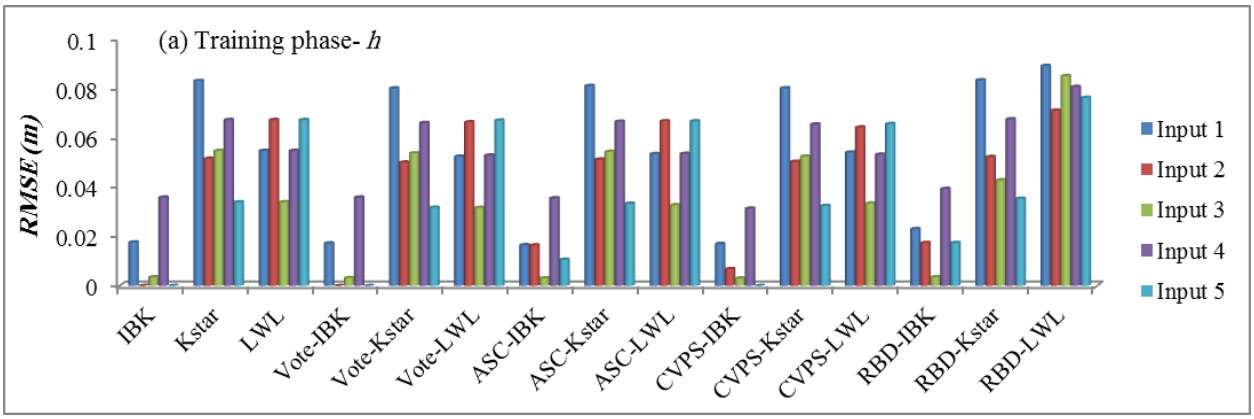
3.1. Determination of the best input variable combination

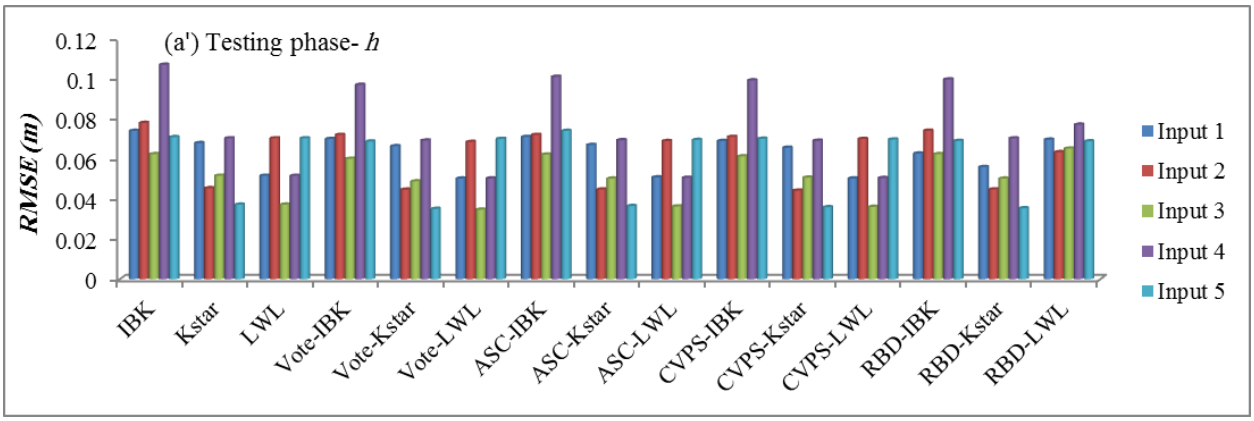
Figure 3 shows that, due to the different structures of each model, the optimal input variable combinations differ between the models. Input combination No. 3 (d_{50} and τ^*) and No. 5 (Q, d_{50} and τ^*) were most influential in both the training and testing phase for flow depth prediction; No. 2 input combination, was only the most effective for the RBD-LWL algorithm. This result reveals that overall Q is not a particularly effective variable, as neither No. 2 nor No.4 input combinations could predict flow depth accurately. This finding is in accordance with the *PCC* values displayed in Figure 2.

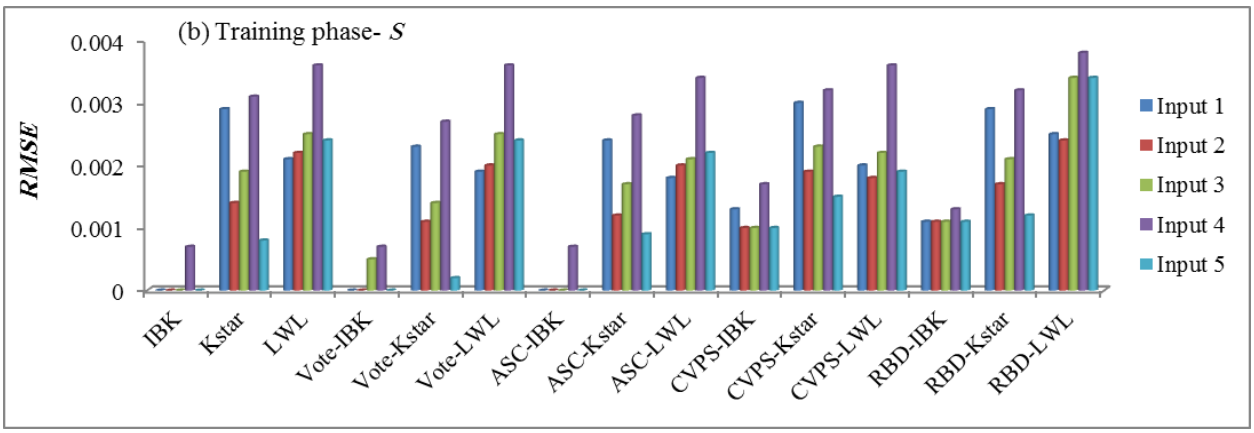
The best input combinations for predicting longitudinal slope were No. 2 and 3. Combination No. 4 (Q, d_{50}) could not predict slope accurately, revealing that Shields stress was the most effective parameter.

Contrasting results were found for predicting width. No. 2 and 3 were the optimum input combination for just a few of the models, while No.1 and No.5 input combinations were the optimum combination in most cases. Input No.1, which only contains d_{50}, predicted flow depth accurately in all models, reflecting its high *PCC* value (Figure 2). In all models, the *RMSE* is larger for the testing than the training phase as

commonly found in AI methods because the training data are assessed on the same data that have been learnt before, while the test dataset has data that is unknown to the algorithm and gives rise to more errors or misclassification. Overall, the results show the single, most effective parameter is not able to predict hydraulic geometry dimensions with the highest degree of accuracy.







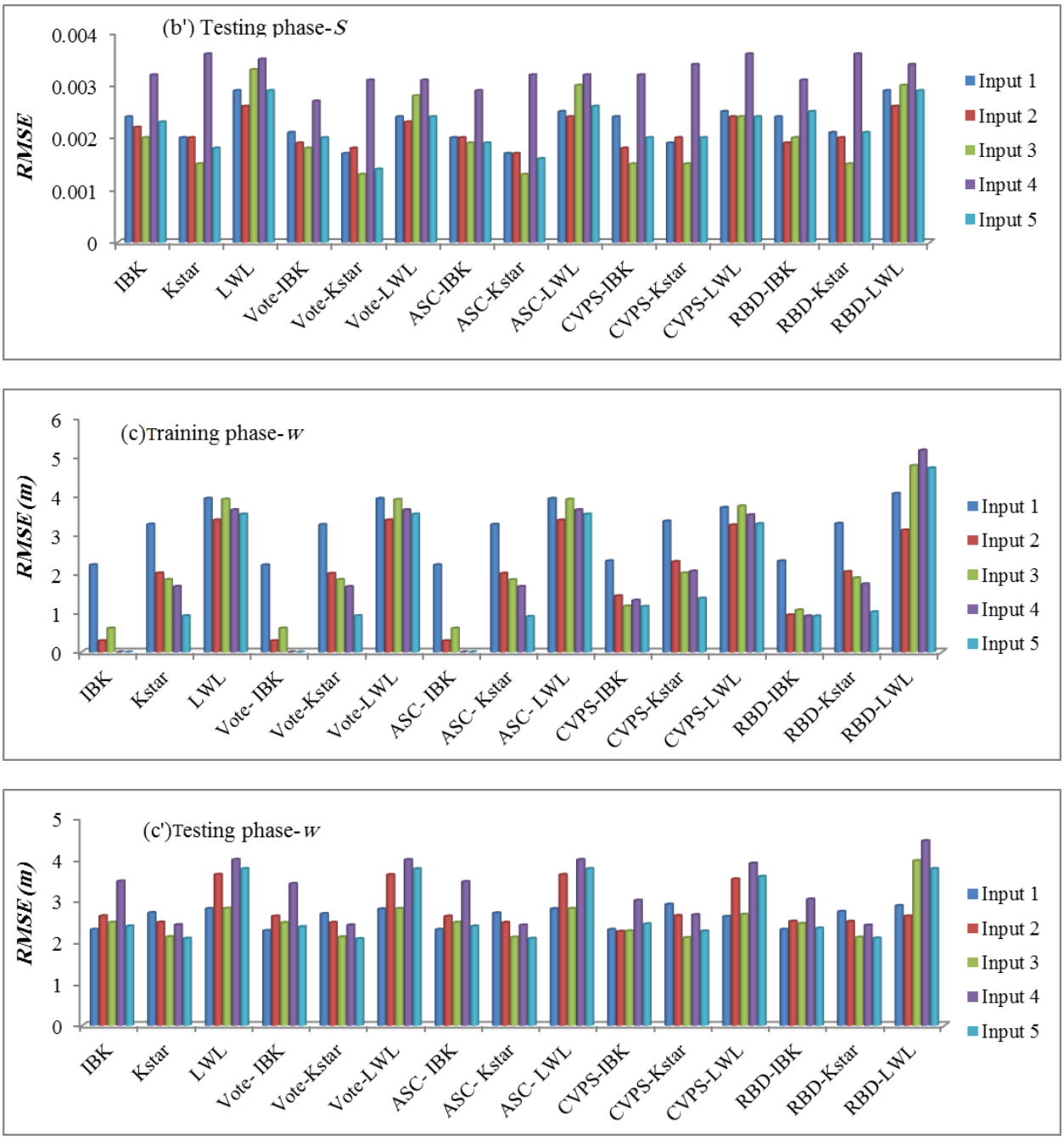
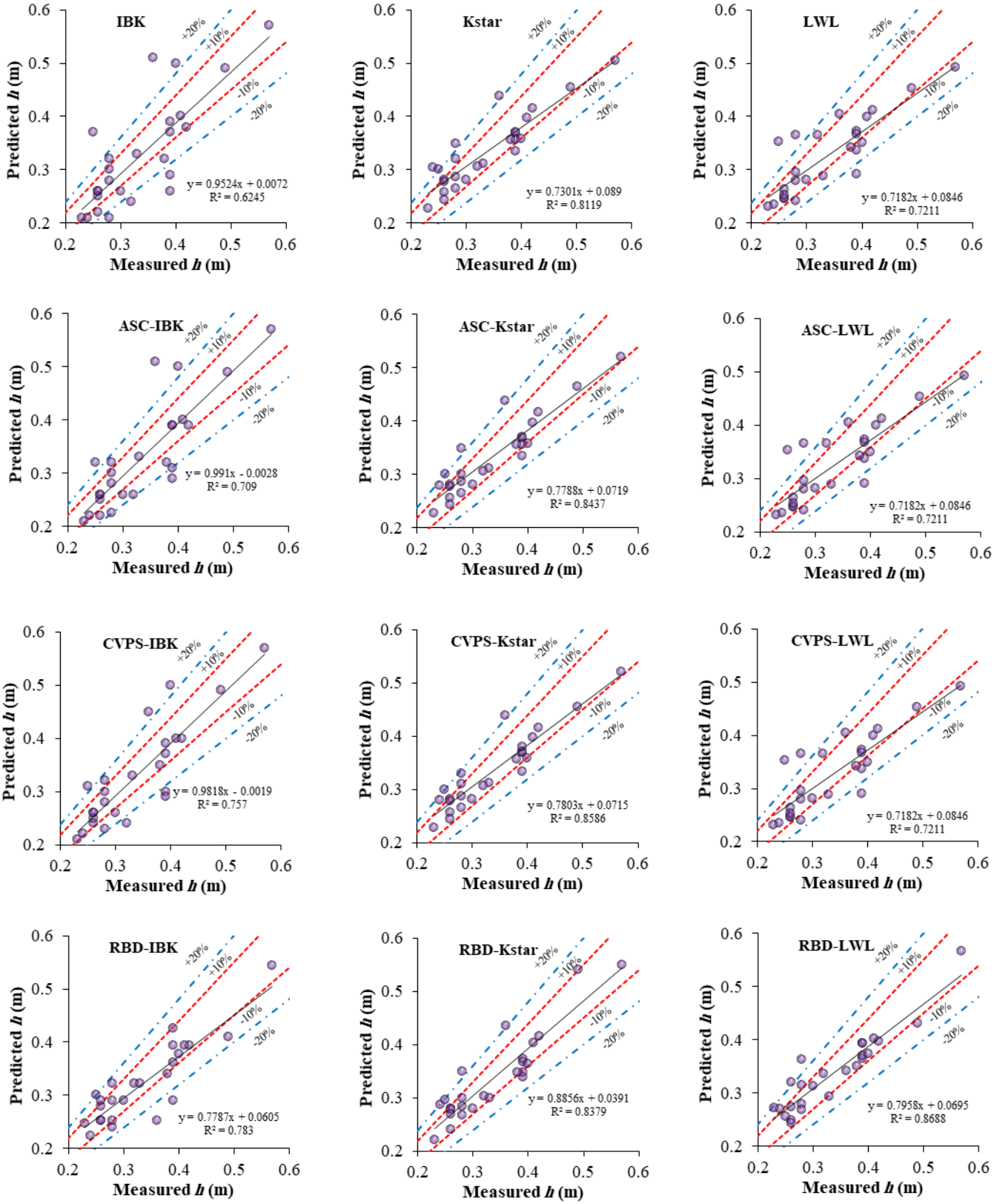


Fig 3. The change in model *RMSE* for different input variable combinations: (a) training phase, *h*; (a') testing phase, *h*; (b) training phase, *S*; (b') testing phase, *S*; (c) training phase, *w*; (c') testing phase, *w*.

3.2. Model performance

After determination of the most effective input variables and optimised model operators, three standalone data mining models, along with 12 types of novel hybrid models were developed to predict the hydraulic geometry. The models were built by a training dataset. A comparison of the observed and predicted values from the testing dataset (Figure 4) shows that of the three standalone models, IBK had the lowest prediction power for flow depth (R^2 = 0.624), and Kstar had the highest (R^2 = 0.812). All hybrid algorithms performed better than the standalone models, with, the hybrid Vote-Kstar algorithm performing the best of all models (R^2 = 0.889).

Kstar was also the best performing standalone model for predicting slope (R^2 = 0.792; Figure 5) and width (R^2 = 0.792; Figure 6), and LWL was the lowest (R^2 = 0.792; R^2 = 0.754, respectively). Hybridization of the standalone algorithms increased the model performance for slope and width by a greater degree than for flow depth. The RBD-IBK algorithm outperformed all other algorithms in the prediction of slope (R^2 = 0.913), followed very closely by RBD-LWL (R^2 = 0.910) and ASC-Kstar (R^2 = 0.909). Whereas for width, this order was CVPS-Kstar (R^2 = 0.914), Vote-Kstar (R^2 = 0.911) and RBD-IBK (R^2 = 0.908). According to the classification of performance based on the R^2 metric (Ayele et al. 2017; Legates and McCabe Jr 1999; Moriasi et al. 2007), all models had a 'very good' performance, except the IBK model for depth and slope, and the LWL model for slope, which had 'good' performance.


IBK
y = 0.9524x + 0.0072
R² = 0.6245
Kstar
y = 0.7301x + 0.089
R² = 0.8119
LWL
y = 0.7182x + 0.0846
R² = 0.7211
ASC-IBK
y = 0.991x - 0.0028
R² = 0.709
ASC-Kstar
y = 0.7788x + 0.0719
R² = 0.8437
ASC-LWL
y = 0.7182x + 0.0846
R² = 0.7211
CVPS-IBK
y = 0.9818x - 0.0019
R² = 0.757
CVPS-Kstar
y = 0.7803x + 0.0715
R² = 0.8586
CVPS-LWL
y = 0.7182x + 0.0846
R² = 0.7211
RBD-IBK
y = 0.7787x + 0.0605
R² = 0.783
RBD-Kstar
y = 0.8856x + 0.0391
R² = 0.8379
RBD-LWL
y = 0.7958x + 0.0695
R² = 0.8688
+20%
+10%
-10%
-20%
Predicted h (m)
Measured h (m)
0.2
0.3
0.4
0.5
0.6

414
415

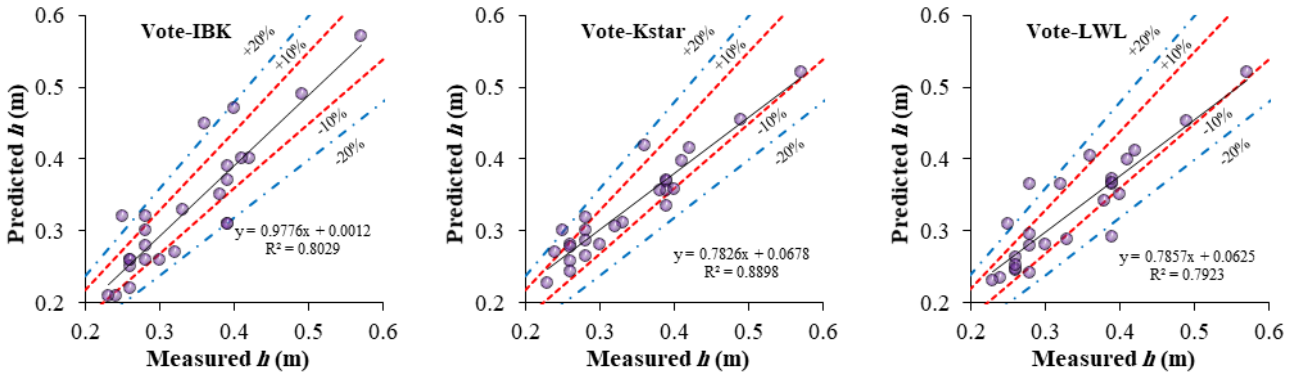
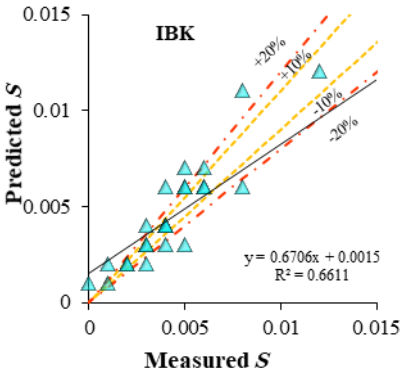
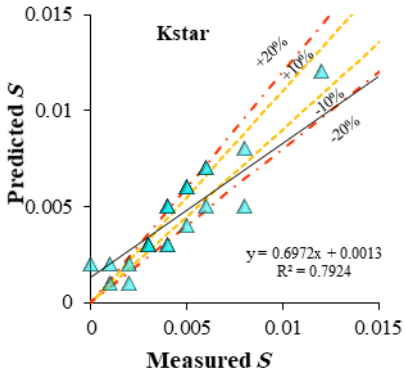
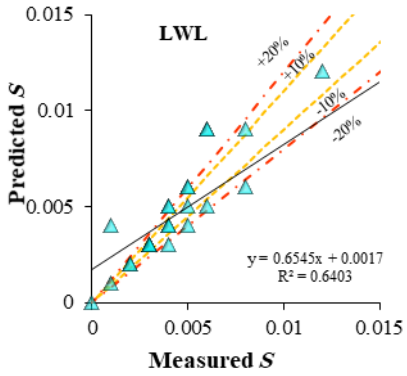
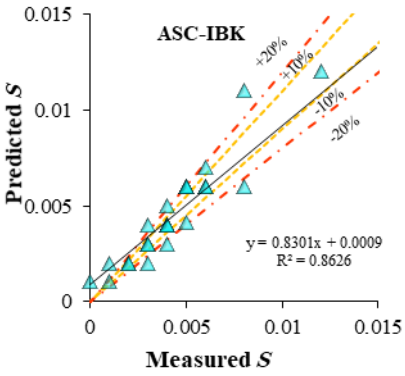


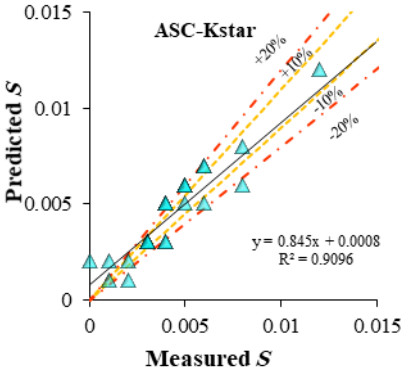
Fig 4. Scatter plot of measured versus predicted flow depth *h*.

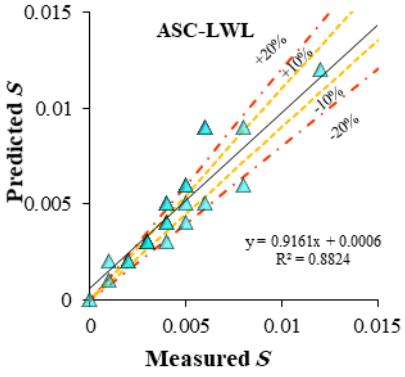

IBK
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.6706x + 0.0015
R² = 0.6611
0
0.005
0.01
0.015

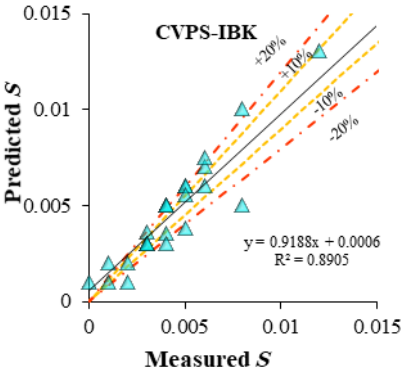

Kstar
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.6972x + 0.0013
R² = 0.7924
0
0.005
0.01
0.015

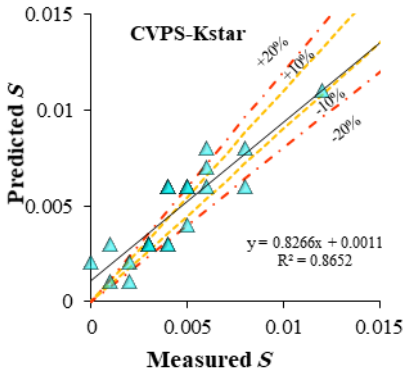

LWL
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.6545x + 0.0017
R² = 0.6403
0
0.005
0.01
0.015

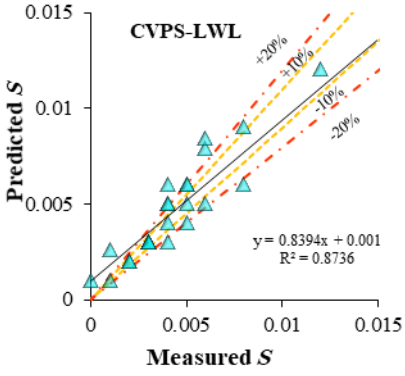

ASC-IBK
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.8301x + 0.0009
R² = 0.8626
0
0.005
0.01
0.015


ASC-Kstar
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.845x + 0.0008
R² = 0.9096
0
0.005
0.01
0.015

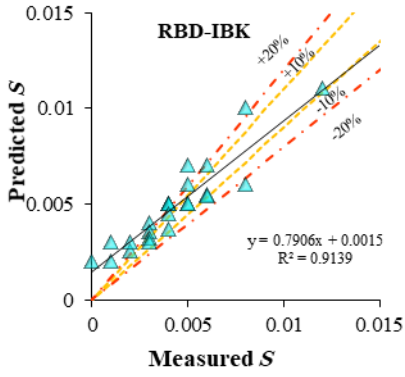

ASC-LWL
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.9161x + 0.0006
R² = 0.8824
0
0.005
0.01
0.015

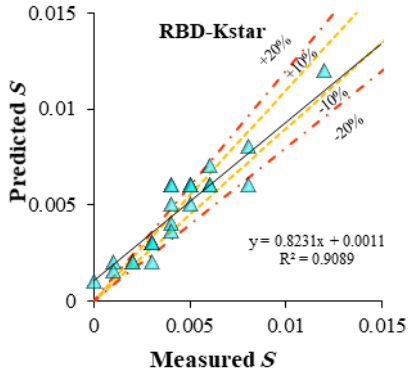

CVPS-IBK
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.9188x + 0.0006
R² = 0.8905
0
0.005
0.01
0.015

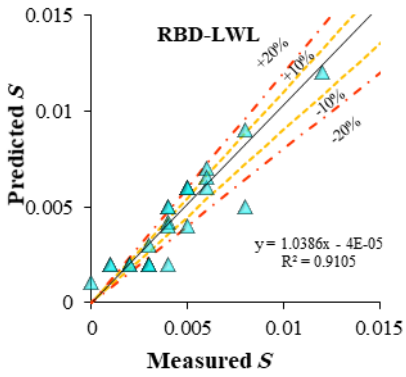

CVPS-Kstar
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.8266x + 0.0011
R² = 0.8652
0
0.005
0.01
0.015

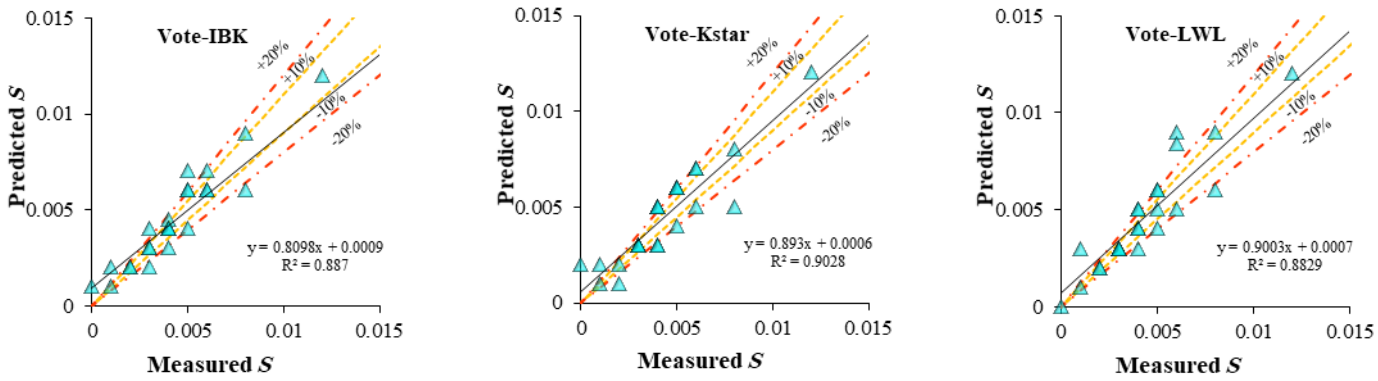

CVPS-LWL
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.8394x + 0.001
R² = 0.8736
0
0.005
0.01
0.015

437


RBD-IBK
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.7906x + 0.0015
R² = 0.9139
0
0.005
0.01
0.015


RBD-Kstar
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 0.8231x + 0.0011
R² = 0.9089
0
0.005
0.01
0.015


RBD-LWL
Predicted S
Measured S
+20%
+10%
-10%
-20%
y = 1.0386x - 4E-05
R² = 0.9105
0
0.005
0.01
0.015



438

439 Fig 5. Scatter plot of measured versus predicted longitudinal slope *S*.

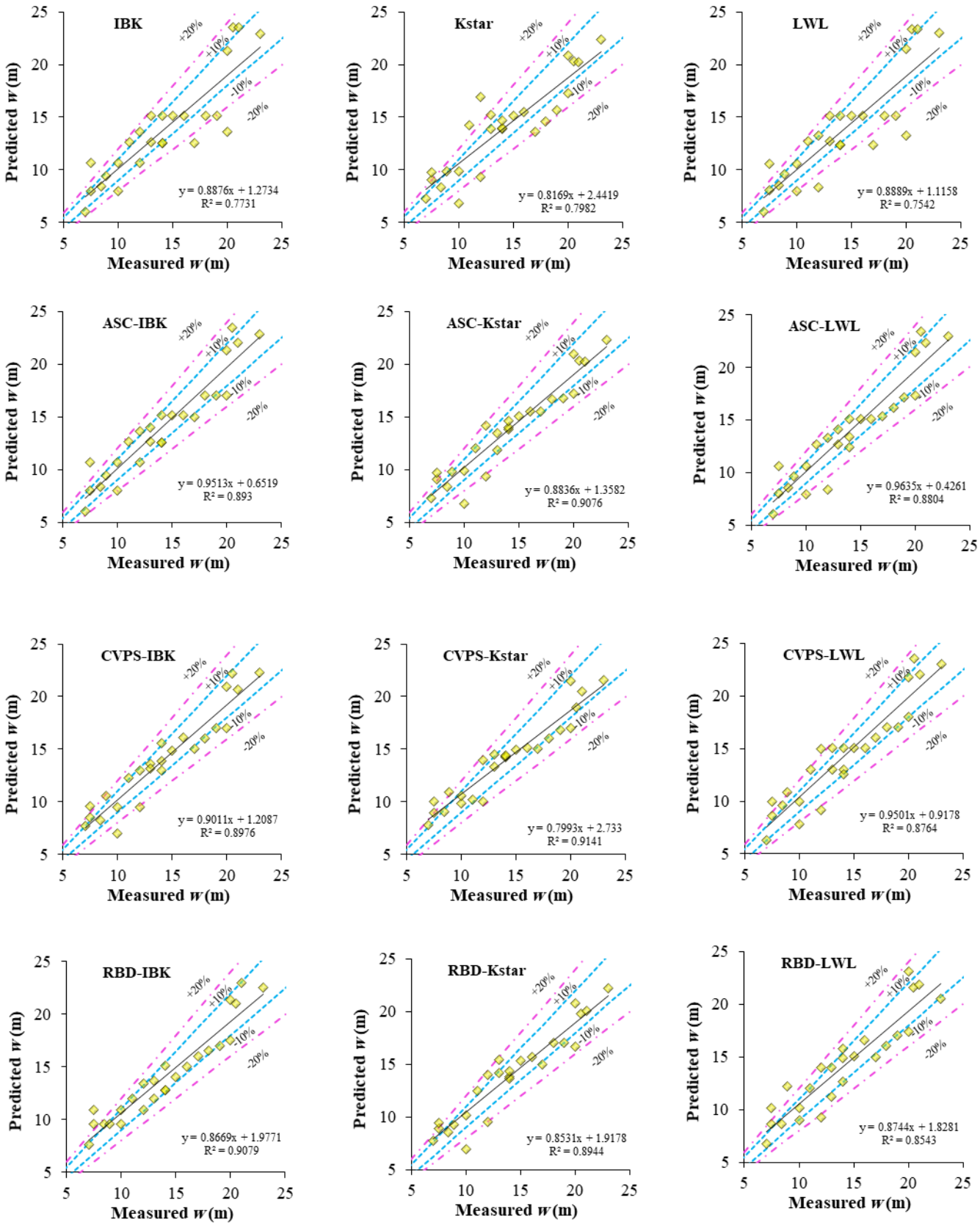
440

441

442

443

444


IBK
y = 0.8876x + 1.2734
R² = 0.7731
Kstar
y = 0.8169x + 2.4419
R² = 0.7982
LWL
y = 0.8889x + 1.1158
R² = 0.7542
ASC-IBK
y = 0.9513x + 0.6519
R² = 0.893
ASC-Kstar
y = 0.8836x + 1.3582
R² = 0.9076
ASC-LWL
y = 0.9635x + 0.4261
R² = 0.8804
CVPS-IBK
y = 0.9011x + 1.2087
R² = 0.8976
CVPS-Kstar
y = 0.7993x + 2.733
R² = 0.9141
CVPS-LWL
y = 0.9501x + 0.9178
R² = 0.8764
RBD-IBK
y = 0.8669x + 1.9771
R² = 0.9079
RBD-Kstar
y = 0.8531x + 1.9178
R² = 0.8944
RBD-LWL
y = 0.8744x + 1.8281
R² = 0.8543
+20%
+10%
-10%
-20%
Predicted w (m)
Measured w (m)

445

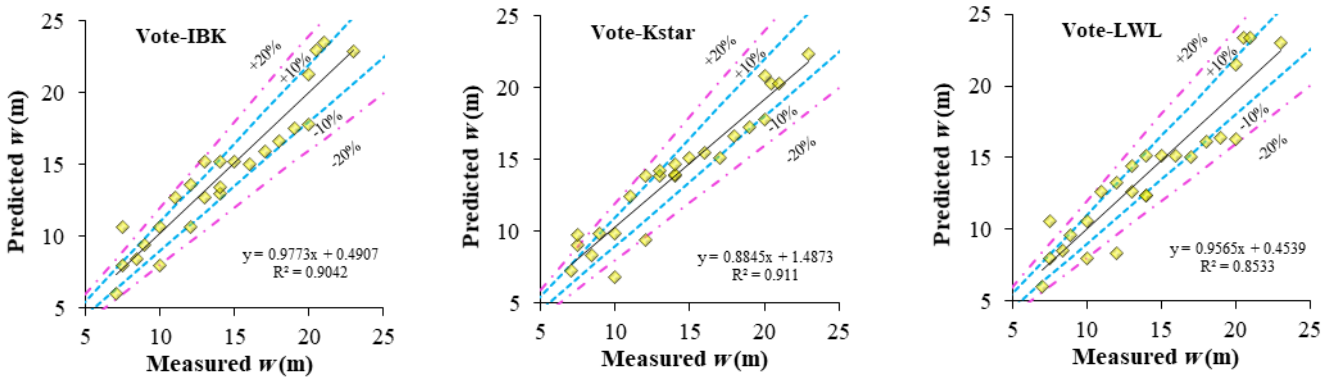


Fig 6. Scatter plot of measured versus predicted water-surface width *w*.

Box plots of measured and predicted hydraulic geometry dimensions shows that the hybrid models ASC-IBK, CVPS-IBK and Vote-IBK predicted the maximum and third quartile depth well, and the IBK standalone algorithm was reasonably accurate in predicting the maximum. Kstar was the only model to predict the median depth well. In terms of the first quartile, IBK, LWL, ASC-LWL, CVPS-IBK, CVPS-LWL, Vote-IBK, and Vote-LWL were the most accurate, and the LWL, ASC-LWL, CVPS-LWL, Vote-Kstar, Vote-LWL, CVPS-Kstar, ASC-Kstar and Kstar model were the most accurate for the minimum channel depth.

All algorithms predicted the maximum and third quartile slope well (Figure 7(b)), but only RBD-LWL, Vote-IBK and Vote-LWL were able to predict median slope accurately. All algorithms provided good estimates of the first quartile, except the standalone algorithms and RBD-LWL. The minimum slope was well reproduced by the LWL, ASC-LWL and Vote-LWL models.

In contrast, none of the algorithms were able to predict maximum and third quartile width accurately. But Vote-Kstar, RBD-LWL and RBD-Kstar models predicted median values very well, and RBD-LWL and CVPS-Kstar did likewise for the third quartile.

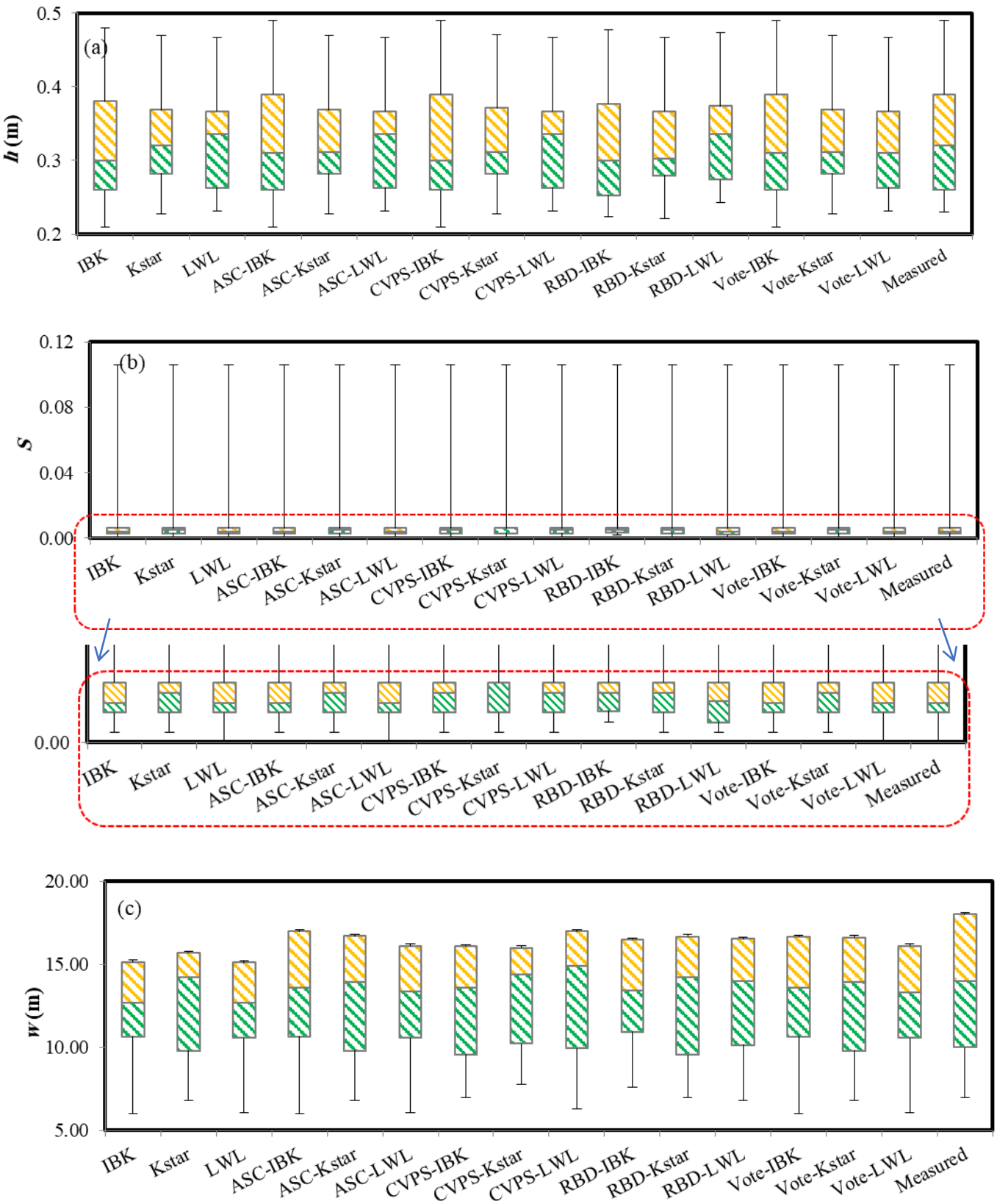

(a)
h (m)
0.5
0.4
0.3
0.2
(b)
S
0.12
0.08
0.04
0.00
0.00
(c)
w (m)
20.00
15.00
10.00
5.00
IBK
Kstar
LWL
ASC-IBK
ASC-Kstar
ASC-LWL
CVPS-IBK
CVPS-Kstar
CVPS-LWL
RBD-IBK
RBD-Kstar
RBD-LWL
Vote-IBK
Vote-Kstar
Vote-LWL
Measured

Fig 7. Box plot of measured and predicted hydraulic geometry: (a) flow depth, (b) longitudinal water surface slope and (c) water surface width.

Since the coefficient of determination R^2 is standardised for differences between the mean and variance of measured and predicted values, this metric is sensitive to outliers and should not be used for model evaluation alone (Legates and McCabe, 1999; Shiri and Kisi, 2012). Thus other evaluation metrics were considered and are shown in Table 3. The metrics of model performance reveal that Vote-Kstar algorithm had the highest prediction power for depth (*RMSE* = 0.0292 m *MAE* = 0.0241, *NSE* = 0.872) followed by RBD-LWL (*RMSE* = 0.0304 m, *MAE* = 0.0229 m, *NSE* = 0.862) and CVPS-Kstar (*RMSE* = 0.0317 m, *MAE* = 0.0251 m, *NSE* = 0.850) (Table 4). The best performing model (Vote-Kstar) had 49.5 %, 7.8 % and 19.2 % higher prediction capability than the IBK, Kstar and LWL standalone algorithms, based on the *NSE* metric. According to the *NSE* values, IBK model had an 'acceptable' performance, ASC-IBK had a 'satisfactory' performance, LWL, ASC-LWL, CVPS-IBK and CVPS-LWL had 'good' prediction power, and the rest of algorithms had 'very good' performance.

Differing results were found in the prediction of slope. The ASC-Kstar algorithm (*RMSE* = 0.001 m, *MAE*= 0.0008 m, and *NSE* = 0.904) outperformed other algorithms, followed by Vote-Kstar (*RMSE* = 0.001, *MAE* = 0.0008 m, *NSE* = 0.902), RBD-Kstar (*RMSE* = 0.0011 m, *MAE*= 0.0007 m, *NSE* = 0.897). In terms of *NSE*, ASC-Kstar, as the most accurate model, had 27.0 %, 13.8 % and 29.3 % higher prediction power than the IBK, Kstar and LWL standalone models respectively. LWL had an 'acceptable' performance, IBK had a 'good' performance, and other algorithms had a 'very good' prediction power.

Vote-Kstar outperformed all algorithms (*RMSE* = 1.373 m, *MAE* = 1.059 m, *NSE* = 0.909) for the estimation of width, as also observed for depth,, followed by RBD-IBK (*RMSE* = 1.401 m, *MAE* = 1.206 m, a *NSE* = 0.905), ASC-Kstar (*RMSE* = 1.418 m, *MAE*= 1.065 m, *NSE* = 0.903). In terms of *NSE*, the Vote-Kstar model had about 17.4 %, 12.4 % and 20.8 % higher performance than the standalone models. LWL had 'good' prediction and other algorithms had 'very good' performance.

515 According to the *PBIAS* metric, all developed algorithms under-estimated depth except RBD-Kstar and
516 RBD-LWL models, over-estimated slope except IBK and Kstar, and under-estimated width except CVPS-
517 LWL, RBD-IBK, RBD-LWL and Vote-IBK.

518 All model performance metrics reveal that although hybridisation enhances the prediction power of
519 standalone algorithms, the level of enhancement and overall performance of hybridised algorithms were
520 strongly dependent upon the choice of standalone algorithm. For instance, in the prediction of depth, the
521 use of Vote to hybridise IBK increased the *NSE* by 72 % but by just 9 % in the case of Kstar. But despite
522 this increase, the standalone Kstar algorithm (*NSE* = 0.80) still had a higher performance than the hybrid
523 Vote-IBK model (*NSE* = 0.76).

524

525 Table 4. Evaluation of model performance

Variable	Models	R^2	*RMSE* (m)	*MAE* (m)	*NSE*	*PBIAS* (%)	Rank based on *NSE*	Percentage lower performance than the best model according to NSE
h	IBK	0.62	0.06	0.04	0.44	2.61	13	49.46
	Kstar	0.81	0.04	0.03	0.80	0.56	6	7.98
	LWL	0.72	0.04	0.03	0.71	3.06	10	19.22
	ASC-IBK	0.70	0.05	0.03	0.59	1.72	12	32.16
	ASC-Kstar	0.84	0.03	0.03	0.84	0.78	4	4.12
	ASC-LWL	0.72	0.04	0.03	0.71	3.06	10	19.22
	CVPS-IBK	0.75	0.05	0.03	0.68	2.40	11	22.08
	CVPS-Kstar	0.85	0.03	0.03	0.85	0.75	3	2.60
	CVPS-LWL	0.72	0.04	0.03	0.71	3.06	10	19.22
	RBD-IBK	0.78	0.04	0.03	0.75	4.16	9	13.60
	RBD-Kstar	0.83	0.03	0.03	0.84	-0.17	5	4.89
	RBD-LWL	0.86	0.03	0.02	0.86	-0.20	2	1.46
	Vote-IBK	0.80	0.04	0.03	0.76	1.90	8	13.02
	Vote-Kstar	0.88	0.03	0.02	0.87	1.62	1	------
	Vote-LWL	0.79	0.04	0.03	0.78	2.87	7	11.76

	Models	R^2	*RMSE*	*MAE*	*NSE*	*PBIAS* (%)	Rank based on *NSE*	Lower performance than the best model (%) according to

								NSE
	IBK	0.66	0.0019	0.0010	0.6608	0.8333	14	26.991
	Kstar	0.79	0.0016	0.0010	0.7797	2.5000	13	13.827
	LWL	0.64	0.0020	0.0010	0.6399	-0.8333	15	29.314
	ASC-IBK	0.86	0.0012	0.0008	0.8608	-1.7500	11	4.867
	ASC-Kstar	0.91	0.0010	0.0008	0.9042	-2.0583	1	------
	ASC-LWL	0.88	0.0012	0.0007	0.8776	-4.1667	9	2.986
	CVPS-IBK	0.89	0.0011	0.0009	0.8862	-4.0833	5	1.991
S	CVPS-Kstar	0.86	0.0013	0.0010	0.8566	-5.8333	12	5.309
	CVPS-LWL	0.87	0.0012	0.0008	0.8674	-4.9167	10	4.092
	RBD-IBK	0.91	0.0012	0.0009	0.8796	-9.3750	7	2.765
	RBD-Kstar	0.90	0.0011	0.0007	0.8972	-4.2500	3	1.32
	RBD-LWL	0.91	0.0011	0.0008	0.8906	-3.0833	4	1.548
	Vote-IBK	0.88	0.0011	0.0007	0.8802	-0.4167	6	2.654
	Vote-Kstar	0.90	0.0010	0.0008	0.9021	-1.6667	2	0.221
	Vote-LWL	0.88	0.0012	0.0007	0.8785	-4.5000	8	2.876

	Models	R^2	*RMSE* (m)	*MAE* (m)	*NSE*	PBIAS (%)	Rank based on *NSE*	Lower performance than the best model (%) according to NSE
	IBK	0.77	2.27	1.71	0.75	2.19	14	17.38
	Kstar	0.79	2.06	1.50	0.80	0.96	13	12.43
	LWL	0.75	2.41	1.82	0.72	3.18	15	20.79
	ASC-IBK	0.89	1.52	1.24	0.89	0.24	8	2.20
	ASC-Kstar	0.90	1.42	1.07	0.90	1.99	3	0.66
	ASC-LWL	0.88	1.63	1.32	0.87	0.62	9	4.07
	CVPS-IBK	0.89	1.47	1.17	0.90	1.30	6	1.43
w	CVPS-Kstar	0.91	1.45	1.18	0.90	0.65	4	1.10
	CVPS-LWL	0.87	1.66	1.38	0.87	-1.53	10	5.39
	RBD-IBK	0.90	1.40	1.21	0.91	-0.74	2	0.44
	RBD-Kstar	0.89	1.50	1.18	0.89	1.06	7	1.98
	RBD-LWL	0.85	1.74	1.45	0.85	-0.43	11	6.16
	Vote-IBK	0.90	1.46	1.21	0.90	-1.22	5	1.32
	Vote-Kstar	0.91	1.37	1.06	0.91	0.98	1	------
	Vote-LWL	0.85	1.83	1.49	0.84	1.13	12	7.70

4. Discussion

4.1 Effect of input variables on model prediction performance

530 The combination of input variables had a strong effect on model prediction power, confirming that the
531 determination of the optimum combination is one of the most significant steps in producing an accurate
532 data mining model. For example, the best input combination for the prediction of flow depth using the
533 Vote-LWL model had ~51 % higher prediction performance (in terms of *NSE*) than the worst input
534 combination. The optimum input variable combination was different from one model to another one,
535 resulting from the different structure of each model, particularly in terms of their flexibility, computing
536 capability and complexity. Thus a range of different input variable combinations must be considered in
537 the optimisation of data mining models.

538 To determine this optimum input combination, this paper used a manual approach, building and testing
539 numerous input combinations. Others have used Principal Component Analysis (*PCA*) (e.g. Barzegar *et*
540 *al*., 2017) or a gamma test (e.g. Ahmadi *et al*., 2015) of the input and output data to determine just one
541 input combination automatically. Determining the optimum combination manually can produce models
542 with a higher prediction performance. For example, in the prediction of fluoride concentration in
543 groundwater Khosravi *et al*. (2019a) built eight different input combinations, and Barzegar *et al*. (2017),
544 using the same dataset, applied *PCA* to extract the best input combination. The manually derived input
545 variable combination produced a 27.5 % higher prediction performance (in terms of *NSE*) than the one
546 extracted by PCA, highlighting the need to first conduct a sensitivity analysis to establish the range of
547 input combinations that need to be considered manually.

548 According to the findings of this paper, excluding Shields stress from the input combination in the
549 prediction of flow depth caused a 49.2 % decrease in model prediction power, and was the most
550 influential variable on model prediction performance, followed by Q (28.0 %) and d_{50} (21.4 %) (Figure
551 8). Very similar results were found for longitudinal slope; Shields stress caused a 54.0 % change in
552 prediction power and was the most effective parameter, followed by d_{50} (22.3 %) and Q (7.7 %), in line
553 with previous results (Julien and Wargadalam, 1995; Afzalimehr *et al*., 2010; Gholami *et al*., 2017;

554 Shaghaghi *et al*., 2018). Omitting flow discharge as an input variable increased model prediction
555 performance, showing that Shields stress and d_{50} are only required to predict accurately slope.

556 In the prediction of top width, d_{50} caused a 15.4 % change in model prediction power, and was the most
557 effective parameter, followed by Shields stress (13.3 %) and discharge (1.41 %). When compared to the
558 effect on the prediction of depth and slope, d_{50} had a much lower impact on width, resulting from the low
559 correlation between d_{50} and width in alluvial rivers. For example, width more strongly depends on the
560 characteristics of the bank material, such as the percentage of bank vegetation growth (e.g. Hey and
561 Thorne, 1986; Bettess *et al*., 1988; Gholami *et al*., 2017) than of the bed materials.

562 These results on the most effective parameters are intuitive, given the strong correlation between
563 sediment transport rate - and thus channel form - and Shields stress, and the weaker correlations with Q
564 and d_{50} (Julien and Wargadalam 1995). For example, Q only has an in-direct influence on sediment
565 transport through its correlation with Shields stress. In other words, two channels, one wide and one
566 narrow, or one shallow and one steep, with the same d_{50} can experience the same Q but different Shields
567 stress and thus sediment transport rates. However, contrasting results on the most influential factors on
568 depth and width have been found. For example, numerous studies have found discharge to be the most
569 important factor, followed by Shields stress and d_{50} (Afzalimehr et al. 2009, 2010; Bray 1982; Hey and
570 Thorne 1986a). Abdelhaleem *et al*. (2016) showed that as well as flow discharge other controlling
571 factorssuch as flow velocity must be incorporated to increase prediction accuracy. Thus, the most
572 effective input parameter is not constant and differs from one river to another according to morphology,
573 such as the presence of bedforms, large woody debris, vegetation and changes in channel planform.
574 Therefore the models presented here, which are statistical in nature, apply only to the three rivers
575 considered and river swith similar conditions, and should not be applied universally.

576
577

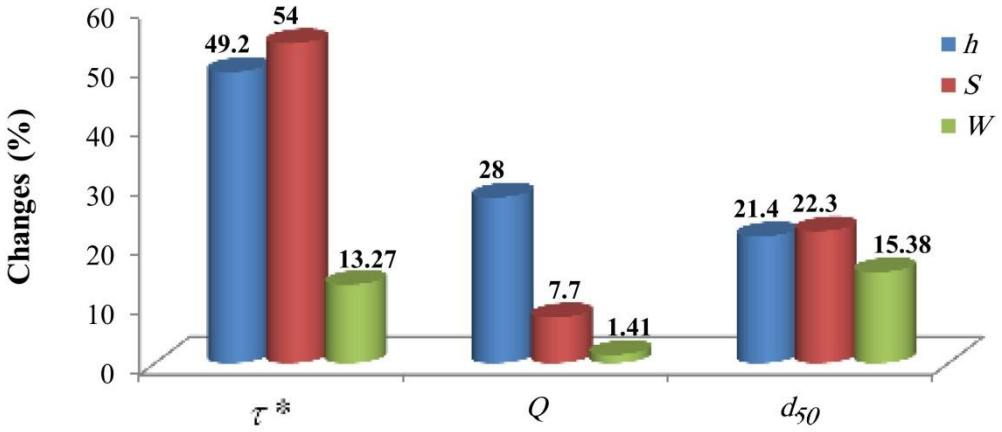


Fig 8. The percentage change in model *RMSE* with each input variable

4.2 Comparison in model prediction performance between empirical, traditional and advanced data mining models

The Afzalimehr et al. (2010) dataset used in this study provides a unique opportunity to compare the performance of empirical equations, traditional machine learning algorithms with the newly developed, advanced data mining models directly: Afzalimehr et al. (2010) tested the performance of empirical and NLR models using this dataset, and Shaghaghi et al. (2018a,b) used the dataset to test traditional data mining models (a hybrid model GS-GMDH, and two standalone models, GEP, Multivariate Adaptive Regression Splines (MARS), Least Square Support Vector Regression (LSSVR) and NLR). Figure 9 shows the results of this comparison in performance, revealing that, the newly developed advanced data mining models outperformed the NLR and traditional data mining models in most of the cases. For example, in modelling depth, the newly developed an ASC-Kstar model produced the lowest *RMSE* value of 0.03 m using all the variables (Q, d_{50} and τ^*) as input, comparing favourably to the results for the GS-GMDH (*RMSE* = 0.24 GEP (*RMSE* = 0.18 m), MARS (*RMSE* = 0.07 m) and NLR (*RMSE* = 0.07 m) models.

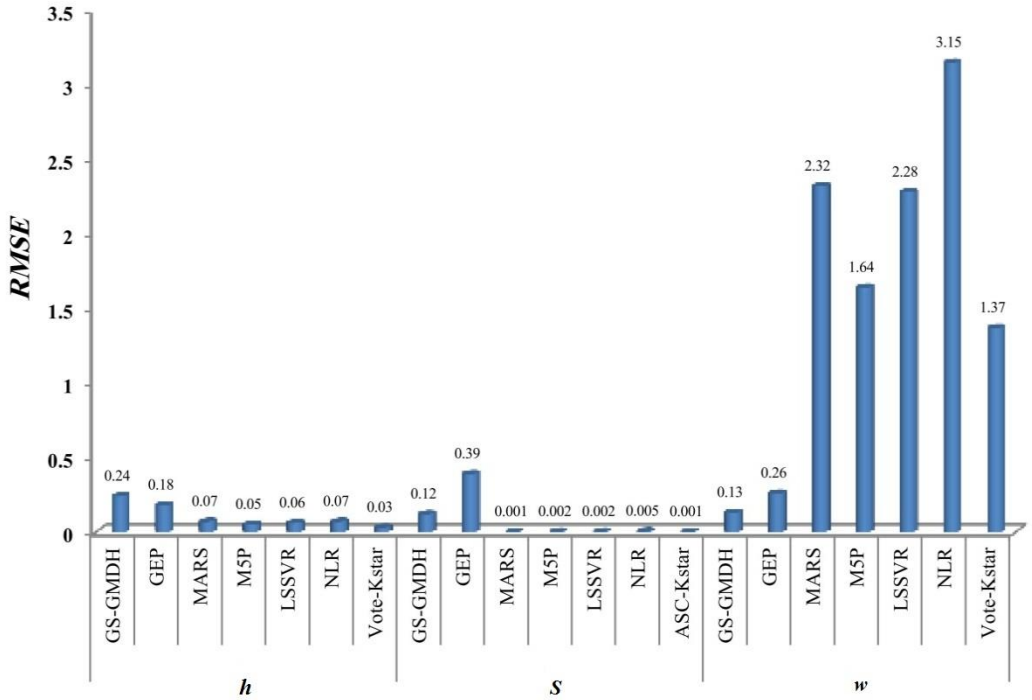


Fig 9. Comparison between model performance of the present study with literature in terms of depth (*h*), slope and width variables.

These comparisons reveal that the new hybrid models proposed in this study are more flexible and accurate than traditional machine learning standalone and hybrid models in most of the cases. The reasons are three-fold. First traditional models are neuron based and need to be optimised to get high prediction power, especially in the determination of the weights of membership function. Advanced data mining models such as tree, lazy, and rule-based models do not have this weakness. Second, NLR models are regression based models and due to their simple structure, are not capable of predicting complicated phenomena accurately. Finally hybridisation improves the performance of standalone models because the process develops a coupled model with higher flexibility, which is proven to better reproduce complex, nonlinear processes that are at play in rivers (Khosravi *et al*. 2020).

4.3 Applying advanced data mining models to forecast stable channel geometry

608 The choice of the 'best' predictive model is most often a compromise between model prediction accuracy
609 and model complexity, with the later, in data mining models, most closely related to the data input
610 requirements. In some data mining models, the highest accuracy has been achieved using all input
611 variables (e.g. Bui *et al.*, 2020b; Khosravi *et al*., 2020), whilst in others, the best prediction power has
612 been obtained with a less complex model using fewer input variables (e.g. Sheikh Khozani *et al.,* 2017b;
613 2019). The major advantages of the data mining models developed in this paper are their simplicity, and
614 their ease and inexpensive to build and run, unlike theoretical and numerical models, whilst providing
615 little compromise on model performance. In other words, a number of the advanced hybrid data mining
616 models provided very good prediction performance for depth and width based on just three input
617 parameters, and for slope based on just two. In stable channel design, predictions of channel geometry are
618 often constrained by the availability of channel data, making less complex models more desirable. Thus
619 the results reveal that these models have great potential for use in stable channel design in data poor
620 catchments, especially in developing nations where technical modelling skills and understanding of the
621 hydraulic and sediment processes occurring in the river system may be lacking.

622 The major disadvantages of these types of model however are two-fold. First, like all statistical methods,
623 the developed models only relate directly to the rivers being considered, and thus their application to
624 other rivers may prove inappropriate. Future studies should apply the developed models to rivers with
625 differing morphologies to discover whether this is the case. Second, due to their 'black-box' structure,
626 they provide poor explanatory power, and thus are unable to extract understanding on the physics that
627 determine hydraulic geometry.

628 With these considerations in mind, the use of data mining techniques may not simply lie in predicting
629 stable channel parameters, but integrating these techniques into process-based models to help identify and
630 optimise model parameters and mitigate uncertainty in model estimates (e.g. Vojinovic *et al*., 2013),d
631 help recognise patterns within observational data to unveil critical details about behavior, and possibly
632 reveal new environmental relationships. Future studies should seek to explore this potential.

633 This study has only considered three controlling parameters. Where data is available, future studies
634 should consider other factors in data mining models, such as flow velocity, relative roughness, suspended
635 sediment load, and bed load transport rate, vegetation form, channel planform, channel roughness, Froude
636 and Reynolds number, and sediment composition (e.g. Abernethy, 2000; Davidson and Hey, 2011),
637 helping to determine the most influential parameters on stable hydraulic geometry and why they vary
638 between rivers.

639

640 ## 5. Conclusion

641
642 Using at-a station field data, this paper has quantified, for the first time, the potential of advanced data
643 mining algorithms to provide accurate predictions of stable hydraulic geometry. Predictions of mean flow
644 depth, top-width and longitudinal slope were made using three standalone data mining techniques -
645 Instance-based Learning (IBK), KStar, Locally Weighted Learning (LWL) - along with 12 types of novel
646 hybrid algorithms in which the standalone models were trained with Vote, Attribute Selection
647 Committees (ASC), Regression by Discretization (RBD), and Cross-validation Parameter Selection
648 (CVPS) algorithms. A comparison was made of the predictive power of these data-driven models, and a
649 sensitivity analysis of three driving variables (discharge, median bed grain diameter and Shields stress)
650 was performed. The main findings were as follows:

651 1- Shield stress was the most effective variable on flow depth and slope prediction; excluding it as
652 an input variable to models caused a 49.2 % and 54 % increase in relative error. Median sediment
653 size had the greatest effect on width prediction power, and excluding this parameter caused a 15.4
654 % increase in relative error. Overall, Shield stress parameter was the most effective parameter on
655 all geometry dimensions.
656 2- The hybrid data mining models had a higher prediction power than standalone models, empirical
657 equations and traditional machine learning algorithms because the hybrid models were more

flexible and thus could better reproduce the nonlinear interactions between input variables and hydraulic geometry. In particular, Vote-Kstar model had the highest prediction capability for depth and width prediction, and ASC-Kstar for slope,

3- According to Nash-Sutcliffe Efficiency values, the IBK model had an acceptable performance, ASC-IBK a satisfactory performance, LWL, ASC-LWL, CVPS-IBK and CVPS-LWL a good prediction power and the rest of the algorithms had a very good performance in flow depth prediction. In estimating slope, LWL had an acceptable performance, IBK a good performance and all other algorithms had a very good prediction accuracy. LWL had a good prediction power in width prediction and all other algorithms had a very good performance.

The strength of these hybrid algorithms lies in their ease to implement, use of a small number of input variables, and being inexpensive to build and run-in comparison to theoretical and numerical models, whilst providing little compromise on model performance. Together, these findings reveal that hybrid data mining models have great potential for use in stable channel design, especially in situations when understanding of the physical processes at play may not be well understood. Thus, understanding more about this potential for different river conditions and input variables represents a vital research avenue.

Authorship contribution statement

Khabat Khosravi: Conceptualization, formal analysis, writing original draft (result and discussion), review & editing, **Zohreh Sheikh Khozani**: Data collection and writing original draft (Introduction, methodology and models description), **James R. Cooper**: writing, review & editing.

Authors' Note: The authors do not have any conflicts of interest or financial disclosures to report.

Software availability

Software Name

683 Waikato Environment for Knowledge Analysis (WEKA) software.

684

685 **Availability**

686 Weka is open-source software and has been written in Java and developed at the University of Waikato,
687 New Zealand. It is free software licensed under the GNU General Public License. Software and
688 documentation (user manual and training material) are freely available at:
689 https://www.cs.waikato.ac.nz/ml/weka/downloading.html

690

691 # References

692 Abdelhaleem, F. S., Amin, A. M., & Ibraheem, M. M. (2016). Updated regime equations for alluvial
693 Egyptian canals. *Alexandria Engineering Journal*, *55*(1), 505–512.
694 https://doi.org/10.1016/j.aej.2015.12.011
695 Abernethy B, R. I. (2000). The effect of riparian tree roots on the mass-stability of riverbanks. *Earth*
696 *Surface Processes and Landforms*, *25*(9), 921–937. https://doi.org/10.1002/1096-
697 9837(200008)25:9<921::AID-ESP93>3.0.CO;2-7
698 Afzalimehr, H., Abdolhosseini, M., & Singh, V. P. (2010). Hydraulic geometry relations for stable
699 channel design. *Journal of Hydrologic Engineering*, *15*(10), 859–864.
700 https://doi.org/10.1061/(ASCE)HE.1943-5584.0000260
701 Afzalimehr, H., Singh, V. P., & Abdolhosseini, M. (2009). Effect of nonuniformity of flow on hydraulic
702 geometry relations. *Journal of Hydrologic Engineering*, *14*(9), 1028–1034.
703 https://doi.org/10.1061/(ASCE)HE.1943-5584.0000095
704 Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy
705 systems: A comparison of support vector regression, random forest, extra trees and regression trees.
706 *Journal of Cleaner Production*, *203*, 810–821. https://doi.org/10.1016/j.jclepro.2018.08.207
707 Ahmadi, A., Han, D., Lafdani, E. K., & Moridi, A. (2015). Input selection for long-lead precipitation
708 prediction using large-scale climate variables: A case study. *Journal of Hydroinformatics*, *17*(1),
709 114–129. https://doi.org/10.2166/hydro.2014.138
710 Anastasakis, L., & Mort, N. (2001). *The development of self-organization techniques in modelling: a*
711 *review of the group method of data handling (GMDH)*. *gmdhsoftware.com*. United Kingdom.
712 Antar, M. A., Elassiouti, I., & Allam, M. N. (2006). Rainfall-runoff modelling using artificial neural

networks technique: A Blue Nile catchment case study. *Hydrological Processes*, *20*(5), 1201–1216. https://doi.org/10.1002/hyp.5932

Arif, M., Ishihara, T., & Inooka, H. (2001). Incorporation of experience in iterative learning controllers using locally weighted learning. *Automatica*, *37*(6), 881–888. https://doi.org/10.1016/S0005-1098(01)00030-9

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Locally Weighted Learning. *Artificial Intelligence Review*, *11*(1–5), 11–73. https://doi.org/10.1007/978-94-017-2053-3_2

Ayele, G. T., Teshale, E. Z., Yu, B., Rutherfurd, I. D., & Jeong, J. (2017). Streamflow and sediment yield prediction for watershed prioritization in the upper Blue Nile river basin, Ethiopia. *Water (Switzerland)*, *9*(10), 782. https://doi.org/10.3390/w9100782

Barzegar, R., Asghari Moghaddam, A., Adamowski, J., & Fijani, E. (2017). Comparison of machine learning models for predicting fluoride contamination in groundwater. *Stochastic Environmental Research and Risk Assessment*, *31*(10), 2705–2718. https://doi.org/10.1007/s00477-016-1338-z

Blench, T. (1969). *Mobile-bed fluviology*. Alberta Press,Edmonton, Canada. https://doi.org/10.1016/0022-1694(70)90091-0

Blench, Thomas. (1952). Regime theory for self-formed sediment-bearing channels. *Transactions of the American Society of Civil Engineers*, *117*(1), 383–400.

Bose, N. K. (1936). Silt movement and design of channels. In *Punjab Eng Congr*. Punjab, India.

Bray, D. I. (1982). Regime equations for gravel-bed rivers. In *Gravel bed rivers: Fluvial processes, engineering and management, R. D. Hey, J. C. Bathurst, and C. R. Thorne, eds., Wiley* (pp. 517–552). Chichester, U.K.

Bui, D. T., Khosravi, K., Karimi, M., Busico, G., Khozani, Z. S., Nguyen, H., et al. (2020). Enhancing nitrate and strontium concentration prediction in groundwater by using new data mining algorithm. *Science of the Total Environment*, *715*, 136836. https://doi.org/10.1016/j.scitotenv.2020.136836

Bui, D. T., Khosravi, K., Li, S., Shahabi, H., Panahi, M., Singh, V. P., et al. (2018). New hybrids of ANFIS with several optimization algorithms for flood susceptibility modeling. *Water (Switzerland)*, *10*(9). https://doi.org/10.3390/w10091210

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, *721*. https://doi.org/10.1016/j.scitotenv.2020.137612

Chang, H. H. (1980). Stable alluvial canal design. *Journal of the Hydraulics Division, ASCE*, *106*(HY5, Proc. Paper, 15420), 873–891.

Chen, W., Panahi, M., & Pourghasemi, H. R. (2017). Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm

(GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling. *Catena*, *157*, 310–324. https://doi.org/10.1016/j.catena.2017.05.034

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., & Kløve, B. (2018). River suspended sediment modelling using the CART model: A comparative study of machine learning techniques. *Science of the Total Environment*, *615*, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293

Cleary, J. G., & Trigg, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. In *Machine Learning Proceedings 1995* (pp. 108–114). https://doi.org/10.1016/b978-1-55860-377-6.50022-0

Cuest Cordoba, G. A., Tuhovčák, L., & Tauš, M. (2014). Using artificial neural network models to assess water quality in water distribution networks. In *Procedia Engineering* (Vol. 70, pp. 399–408). Elsevier Ltd. https://doi.org/10.1016/j.proeng.2014.02.045

Davidson, S. K., & Hey, R. D. (2011). Regime equations for natural meandering cobble- and gravel-bed rivers. *Journal of Hydraulic Engineering*, *137*(9), 894–910. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000408

Dawson, C. W., Abrahart, R. J., & See, L. M. (2007). HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling and Software*, *22*(7), 1034–1052. https://doi.org/10.1016/j.envsoft.2006.06.008

Deshpande, V., & Kumar, B. (2012). Review and assessment of the theories of stable alluvial channel design. *Water Resources*, *39*(4), 481–487. https://doi.org/10.1134/S0097807812040033

Dietterich, T. G. (1997). Machine learning research_ four current directions. *AI Magazine*, *18*(4), 97–1.

Eaton, B. C., & Church, M. (2007). Predicting downstream hydraulic geometry: A test of rational regime theory. *Journal of Geophysical Research: Earth Surface*, *112*(3), 3025. https://doi.org/10.1029/2006JF000734

Ferguson, R. I. (1986). Hydraulics and hydraulic geometry. *Progress in Physical Geography*, *10*(1), 1–31. https://doi.org/10.1177/030913338601000101

Ferreira, C. (2001). Gene Expression Programming: a New Adaptive Algorithm for Solving Problems. *Complex System*, *13*(2), 87–129.

Ferreira, C. (2002). Genetic representation and genetic neutrality in gene expression programming. *Advances in Complex Systems*, *05*(04), 389–408. https://doi.org/10.1142/s0219525902000626

Frank, E., & Bouckaert, R. R. (2009). Conditional density estimation with class probability estimators. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5828 LNAI, pp. 65–81). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-05224-8_7

Garg, T., & Khurana, S. S. (2014). Comparison of classification techniques for intrusion detection dataset

using WEKA. In *International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2014*. https://doi.org/10.1109/ICRAIE.2014.6909184

Gholami, A., Bonakdari, H., Ebtehaj, I., Shaghaghi, S., & Khoshbin, F. (2017). Developing an expert group method of data handling system for predicting the geometry of a stable channel with a gravel bed. *Earth Surface Processes and Landforms*, *42*(10), 1460–1471. https://doi.org/10.1002/esp.4104

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. In *Pattern Recognition Letters* (Vol. 27, pp. 294–300). North-Holland. https://doi.org/10.1016/j.patrec.2005.08.011

Gleason, C. J. (2015). Hydraulic geometry of natural rivers: A review and future directions. *Progress in Physical Geography*, *39*(3), 337–360. https://doi.org/10.1177/0309133314567584

Hastie, T., & Loader, C. (1993). Local regression: Automatic Kernel carpentry. *Statistical Science*, *8*(2), 120–129. https://doi.org/10.1214/ss/1177011002

Henderson, F. M. (1961). Stability of alluvial channels. *Journal of the Hydraulics Division*, *87*(6), 109–138.

Hey, R. D., & Thorne, C. R. (1986a). Stable channels with mobile gravel beds. *Journal of Hydraulic Engineering*, *112*(8), 671–689. https://doi.org/10.1061/(ASCE)0733-9429(1986)112:8(671)

Hey, R. D., & Thorne, C. R. (1986b). Stable Channels with Mobile Gravel Beds. *Journal of Hydraulic Engineering*, *112*(8), 671–689. https://doi.org/10.1061/(asce)0733-9429(1986)112:8(671)

Hooshyaripor, F., Tahershamsi, A., & Golian, S. (2014). Application of copula method and neural networks for predicting peak outflow from breached embankments. *Journal of Hydro-Environment Research*, *8*(3), 292–303. https://doi.org/10.1016/j.jher.2013.11.004

Huang, H. Q., & Nanson, G. C. (1998). The influence of bank strength on channel geometry: an integrated analysis of some observations. *Earth Surface Processes and Landforms*, *23*(10), 865–876. https://doi.org/10.1002/(SICI)1096-9837(199810)23:10<865::AID-ESP903>3.0.CO;2-3

Julien, P. Y., & Wargadalam, J. (1995). Alluvial channel geometry: Theory and applications. *Journal of Hydraulic Engineering*, *121*(4), 312–325. https://doi.org/10.1061/(ASCE)0733-9429(1995)121:4(312)

Khadangi, E., Madvar, H. R., & Kiani, H. (2009). Application of artificial neural networks in establishing regime channel relationships. In *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*. https://doi.org/10.1109/IC4.2009.4909224

Khosravi, K., Barzegar, R., Miraki, S., Adamowski, J., Daggupati, P., Alizadeh, M. R., et al. (2019). Stochastic Modeling of Groundwater Fluoride Contamination: Introducing Lazy Learners. *Groundwater*, gwat.12963. https://doi.org/10.1111/gwat.12963

Khosravi, K., Cooper, J. R., Daggupati, P., Thai Pham, B., & Bui, D. T. (2020). Bedload transport rate

prediction: Application of novel hybrid data mining techniques. *Journal of Hydrology*, *585*, 124774. https://doi.org/10.1016/j.jhydrol.2020.124774

Khosravi, K., Daggupati, P., Alami, M. T., Awadh, S. M., Ghareb, M. I., Panahi, M., et al. (2019). Meteorological data mining and hybrid data-intelligence models for reference evaporation simulation: A case study in Iraq. *Computers and Electronics in Agriculture*, *167*. https://doi.org/10.1016/j.compag.2019.105041

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

Lane, E. W. (1957). A study of the shape of channels formed by natural streams flowing in erodible material. *U.S. Army Corps of Engineers Report*, (9), 141. https://www.worldcat.org/title/study-of-the-shape-of-channels-formed-by-natural-streams-flowing-in-erodible-material/oclc/6671104. Accessed 26 June 2021

Legates, D. R., & McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, *35*(1), 233–241.

Leopold, L., & Wolman, M. (1957). River channel patterns: braided, meandering, and straight. *USGS Professional Paper*, *282-B*, 51.

Mehta, D., Yadav, S., & Anal, S. (2013). Geomorphic channel design and analysis using HEC-RAS hydraulic design functions. *Journal of Global Analysis*, *2*(4), 90–93.

Millar, R. G. (2005). Theoretical regime equations for mobile gravel-bed rivers with stable banks. *Geomorphology*, *64*(3–4), 207–220. https://doi.org/10.1016/j.geomorph.2004.07.001

Mislan, Haviluddin, Hardwinarto, S., Sumaryono, & Aipassa, M. (2015). Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia. In *Procedia Computer Science* (Vol. 59, pp. 142–151). Elsevier. https://doi.org/10.1016/j.procs.2015.07.528

Mohamed, H. I. (2013). Design of alluvial Egyptian irrigation canals using artificial neural networks method. *Ain Shams Engineering Journal*, *4*(2), 163–171. https://doi.org/10.1016/j.asej.2012.08.009

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, 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*(3), 885–900. https://doi.org/10.13031/2013.23153

Noori, R., Deng, Z., Kiaghadi, A., & Kachoosangi, F. T. (2016). How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers? *Journal of Hydraulic Engineering*, *142*(1). https://doi.org/10.1061/(ASCE)HY.1943-7900.0001062

Parhami, B. (1994). Voting Algorithms. *IEEE Transactions on Reliability*, *43*(4), 617–629.

https://doi.org/10.1109/24.370218

Parker, G., Wilcock, P. R., Paola, C., Dietrich, W. E., & Pitlick, J. (2007). Physical basis for quasi-universal relations describing bankfull hydraulic geometry of single-thread gravel bed rivers. *Journal of Geophysical Research: Earth Surface*, *112*(4). https://doi.org/10.1029/2006JF000549

Reyes, O., Cano, A., Fardoun, H. M., & Ventura, S. (2018). A locally weighted learning method based on a data gravitation model for multi-target regression. *International Journal of Computational Intelligence Systems*, *11*(1), 282–295. https://doi.org/10.2991/ijcis.11.1.22

Robinson, C. (1998). *Multi-objective optimisation of polynomial models for time series prediction using genetic algorithms and neural networks.* University of Sheffield, UK.

Shaghaghi, S., Bonakdari, H., Gholami, A., Kisi, O., Shiri, J., Binns, A. D., & Gharabaghi, B. (2018). Stable alluvial channel design using evolutionary neural networks. *Journal of Hydrology*, *566*, 770–782. https://doi.org/10.1016/j.jhydrol.2018.09.057

Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., et al. (2020). Predicting Standardized Streamflow index for hydrological drought using machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 339–350. https://doi.org/10.1080/19942060.2020.1715844

Sheikh Khozani, Z., Bonakdari, H., & Ebtehaj, I. (2017). An analysis of shear stress distribution in circular channels with sediment deposition based on Gene Expression Programming. *International Journal of Sediment Research*, *32*(4), 575–584. https://doi.org/10.1016/J.IJSRC.2017.04.004

Sheikh Khozani, Z., Bonakdari, H., & Zaji, A. H. (2017). Estimating the shear stress distribution in circular channels based on the randomized neural network technique. *Applied Soft Computing*, *58*, 441–448. https://doi.org/10.1016/j.asoc.2017.05.024

Sheikh Khozani, Z., Khosravi, K., Pham, B. T., Kløve, B., Wan Mohtar, W. H. M., & Yaseen, Z. M. (2019). Determination of compound channel apparent shear stress: application of novel data mining models. *Journal of Hydroinformatics*, *21*(5), 798–811. https://doi.org/10.2166/hydro.2019.037

Shelley, J., & Parr, A. D. (2009). Using HEC-RAS hydraulic design functions for geomorphic channel design and analysis. In *Proceedings of World Environmental and Water Resources Congress 2009 - World Environmental and Water Resources Congress 2009: Great Rivers* (Vol. 342, pp. 3722–3731). Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/41036(342)374

Singh, V. P., & Zhang, L. (2008). At-a-station hydraulic geometry relations, 1: Theoretical development. *Hydrological Processes*, *22*(2), 189–215. https://doi.org/10.1002/hyp.6411

Sterling, M., & Knight, D. (2002). An attempt at using the entropy approach to predict the transverse distribution of boundary shear stress in open channel flow. *Stochastic environmental research and risk assessment*, *16*(2), 127–142.

Stevens, M. A., & Nordin, C. F. (1987). Critique of the regime theory for alluvial channels. *Journal of Hydraulic Engineering*, *113*(11), 1359–1380. https://doi.org/10.1061/(ASCE)0733-9429(1987)113:11(1359)

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, *10*, 1040–1053.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Societ. *Journal of the Royal Statistical Society*, *36*(2), 111–147. https://doi.org/10.2307/2984809

Taheri, K., Shahabi, H., Chapi, K., Shirzadi, A., Gutiérrez, F., & Khosravi, K. (2019). Sinkhole susceptibility mapping: A comparison between Bayes-based machine learning algorithms. *Land Degradation and Development*, *30*(7), 730–745. https://doi.org/10.1002/ldr.3255

Tahershamsi, A., Majdzade Tabatabai, M. R., & Shirkhani, R. (2012). An evaluation model of artificial neural network to predict stable width in gravel bed rivers. *International Journal of Environmental Science and Technology*, *9*(2), 333–342. https://doi.org/10.1007/s13762-012-0036-8

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. Part F1288, pp. 847–855). Association for Computing Machinery. https://doi.org/10.1145/2487575.2487629

Vojinovic, Z., Abebe, Y. A., Ranasinghe, R., Vacher, A., Martens, P., Mandl, D. J., et al. (2013). A machine learning approach for estimation of shallow water depths from optical satellite images and sonar measurements. In *Journal of Hydroinformatics* (Vol. 15, pp. 1408–1424). IWA Publishing. https://doi.org/10.2166/hydro.2013.234

Wan Mohtar, W. H. M., Afan, H., El-Shafie, A., Bong, C. H. J., & Ab. Ghani, A. (2018). Influence of bed deposit in the prediction of incipient sediment motion in sewers using artificial neural networks. *Urban Water Journal*, *15*(4), 296–302. https://doi.org/10.1080/1573062X.2018.1455880

Wang, Z., Xing, H., Li, T., Yang, Y., Qu, R., & Pan, Y. (2016). A modified ant colony optimization algorithm for network coding resource minimization. *IEEE Transactions on Evolutionary Computation*, *20*(3), 325–342. https://doi.org/10.1109/TEVC.2015.2457437

White, W. R. (1982). Analytical approach to river regime. *Journal of the Hydraulics Division*, *108*(10), 1179–1193.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. https://doi.org/10.1016/c2009-0-19715-5

Wolman, M. G. (1954). A method of sampling coarse river-bed material. *Eos, Transactions American Geophysical Union*, *35*(6), 951–956. https://doi.org/10.1029/TR035i006p00951

Wu, C. H., Lin, I. S., Wei, M. L., & Cheng, T. Y. (2013). Target position estimation by genetic expression programming for mobile robots with vision sensors. *IEEE Transactions on Instrumentation and Measurement*, *62*(12), 3218–3230. https://doi.org/10.1109/TIM.2013.2272173

Zounemat-Kermani, M., Seo, Y., Kim, S., Ghorbani, M. A., Samadianfard, S., Naghshara, S., et al. (2019). Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida. *Applied Sciences (Switzerland)*, *9*(12). https://doi.org/10.3390/app9122534

Supplementary material

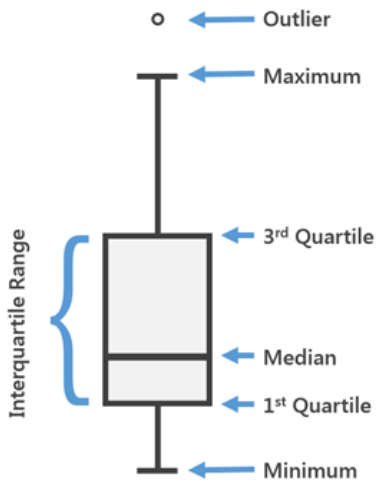


Fig A. Box-plot and its component in details