Bedload transport rate prediction: Application of novel hybrid data mining techniques

Khabat Khosravi[1], James R. Cooper[2], Prasad Daggupati[1], Binh Thai Pham[3], Dieu Tien Bui*[4,5]

1. School of Engineering, University of Guelph, Ontario, Canada.

kkhosrav@uoguelph.ca (K.K) pdaggupa@uoguelph.ca (P.D)

2. School of Environmental Sciences, University of Liverpool, Liverpool, UK.

James.Cooper@liverpool.ac.uk

3. Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

phamthaibinh2@duytan.edu.vn

4. Geographic Information Science Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam.

5. Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Corresponding author: Dieu Tien Bui (buitiendieu@tdtu.edu.vn)

Abstract

The accurate prediction of bedload transport in gravel-bed rivers remains a significant challenge in river science. However the potential for data mining algorithms to provide models of bedload transport have yet to be explored. This study provides the first quantification of the predictive power of a range of standalone and hybrid data mining models. Using bedload transport data collected in laboratory flume experiments, the performance of four types of recently developed standalone data mining techniques - the M5P, random tree (RT), random forest (RF) and the reduced error pruning tree (REPT) - are assessed, along with four types of hybrid algorithms trained with a Bagging (BA) data mining algorithm (BA-M5P, BA-RF, BA-RT and BA-REPT). The main findings are four-fold. First, the BA-M5P model had the highest prediction power (R^2

= 0.943; *RMSE* = 0.061 kg m^{-1} s^{-1}; *MAE* =0.040 kg m^{-1} s^{-1}; *NSE* = 0.945; *PBIAS*= -1.60) followed by M5P, BA-RT, RT, BA-RF, RF, BA-REPT, and REPT. All models displayed 'very good' performance except the BA-REPT and REPT model, which were 'satisfactory'. Second, the M5P, BA-RT, and RT models underestimated, and the BA-M5P, BA-RF, RF, BA-REPT and REPT models overestimated, bedload transport rates. Third, flow velocity had the most significant impact on bedload transport rate (*PCC* = 0.760) followed by shear stress (*PCC* = 0.709), discharge (*PCC* = 0.668), bed shear velocity (*PCC* = 0.663), bed slope (*PCC* = 0.490), flow depth (*PCC* = 0.303), median sediment diameter (*PCC* = 0.247), and relative roughness (*PCC* = 0.003). Fourth, the maximum depth of tree was the most sensitive operator in decision tree-based algorithms, and batch size, number of execution slots and number of decimal places did not have any impact on model' prediction power. Overall the results revealed that hybrid data mining techniques provide more accurate predictions of bedload transport rate than standalone data mining models. In particular, M5P models, trained with a Bagging data mining algorithm, have great potential to produce robust predictions of bedload transport in gravel-bed rivers.

Keywords: bedload, flume experiment, data mining, river, artificial intelligence

Table A. Summary of abbreviation and symbols

Full name	**Abbreviation**
Random tree	RT
M5 Prime	M5P
Random forest	RF
Reduced error pruning tree	REPT
Bagging	BA
Pearson correlation coefficient	PCC
Artificial neural network	ANN

Adaptive Neural Fuzzy Inference System	ANFIS
Support Vector Machines	SVM
Genetic programming	GP
Fuzzy logic	FL
Feed-forward neural network-extreme learning machine	FFNN-ELM
Logistic model tree	LMT
Naïve Bayes Trees	NBT
Instance-Based K-nearest Neighbours	IBK
Hydrologic Engineering Center's River Analysis System	HEC-RAS
Median sediment grain diameter	d_{50}
Sediment diameter	D
Flow discharge	Q
Flow velocity	V
Flow depth	Y
Bed slope	S
Relative roughness	RR
Shear stress	τ
Shear velocity	V^*
Water density	ρ
Gravitational acceleration	g
Hydraulic radius	R
Bedload sediment transport rate	q
Mass of collected sediment in trap	G
Flume width	b
Sampling duration	T
Root Mean Square Error	$RMSE$
Coefficient of determination	R^2

Mean Absolute Error	*MSE*
Nash-Sutcliffe Efficiency	*NSE*
Percent bias	*PBIAS*
Observed values	X_o
Predicted values	X_e
Mean observed values	$\overline{X}_0$
Mean predicted values	$\overline{X}_e$
Standard deviation	*SD*
Multivariate adaptive regression splines	MARS
Standard deviation reduction	*SDR*
Plane rooted decision tree with m nodes	*T*
Class of a plane rooted tree	*C*
Random tree sets	C_m
Weight function	$\omega(T)$
Out-degree	*k*
Nonnegative numbers	φ
Generating function	*a(t)*

1. Introduction

Bedload transport, particularly of coarser sediments, is one of the main drivers for the morphological change of gravel-bed rivers. Thus the quantification of bedload transport rate is of paramount importance for river engineers and fluvial geomorphologists interested in river management (Wilcock, 1998) and landscape evolution (Howard, 1994). In the field, the measurement of bedload transport is challenging and often expensive, especially during flooding, and is associated with estimates of rate with a high uncertainty (Mao, 2012; Graf,

1971). Laboratory flume experiments are more commonly used because of the ability to carefully control boundary conditions and to perform more precise measurements, allowing bedload transport formulas to be developed over a range of flow and bed conditions (e.g. Einstein, 1950; Engelund and Hansen, 1967; Meyer-Peter and Muller, 1948; Wilcox and Crowe, 2003). However these experimental investigations have their disadvantages: (1) they can be costly and time-consuming; (2) they are a simplification of a natural gravel-bed river (e.g. use of equilibrium sediment transport conditions and steady and uniform flows); (3) are scaled versions of a natural system, and thus problems exist in trying to correctly scale flow and sediment properties; and (4) the magnitude of transport that can be reproduced is limited. One consequence is that bedload transport formulas that are developed from flume experiments can be associated with a high degree of predictive uncertainty (Mao, 2012).

A suite of empirical, mathematical and numerical approaches have been developed for bedload transport prediction, and these approaches have their weaknesses. Extensive data is required to build, calibrate and validate these models, particularly if the model is process-based. Furthermore there is much complexity and difficulty in performing model parameterization and calibration if the physics is not well understood, which is often the case because of the non-linear dynamics of bedload transport (Hamel et al., 2017; Kisi et al., 2012). One example is in the use of the most popular and widely used model in river science, the Hydrologic Engineering Center's River Analysis System (HEC-RAS) model. Although this model has been used successfully implemented in numerous studies, it suffers from the need for approximations and simplifications that can introduce errors in the prediction of bedload transport: (1) the flow is often modeled using a depth-averaged approach; (2) the law of the wall formula is fed with an average shear stress value; (3) transport capacity equations are used to predict sediment transport under the assumption of unlimited sediment supply conditions; and (4) the determination of the correct value of some of the model parameters, such as the active layer

depth and the Manning roughness coefficient, is problematic without detailed observations of the river being modeled (HEC-RAS, 2008; Ghafouri Azar et al., 2012; Mustafa et al., 2017; Shahiri et al. 2016). Considering these challenges, alternative approaches to the use of empirical, mathematical and numerical methods should be explored.

Recently the advent of artificial intelligence algorithms, based on machine learning and data mining techniques, are providing new insights in multiple areas of science, including water resources and geoscience. These algorithms attempt to deduce the optimal relationship between the inputs (i.e. significant conditioning factors) and the target (i.e. output parameters) mainly operating as a black-box type, non-linear, statistical model (Yaseen et al., 2017). Most phenomena within a watershed, including sediment transport, are relatively complex and they cannot be predicted easily (Khosravi et al., 2018a). Thus, in most situations, the applied model must be versatile, flexible and possess a non-linear modeling structure. Artificial intelligence algorithms meet these requirements.

Ebtehaj and Bonakdari (2013) applied artificial neural network (ANN) algorithms for predicting sediment transport in sewers, revealing that ANN had a higher prediction power than existing empirical transport formulas. Similar results have also been found within other areas of hydrology and hydraulics (Melesse et al., 2011; Kisi et al, 2016). However ANN algorithms have poor prediction power when the range of the testing dataset is outside of the range of the training data (Melesse et al., 2011; Kisi et al., 2012), and they require a long-term dataset to achieve a reasonable result. Thus, to solve this weakness, ANN algorithms have been ensembled with fuzzy logic (FL) algorithms to create Adaptive Neural Fuzzy Inference System (ANFIS) models. Ebtehaj and Bonakdari (2014) used such a model for predicting sediment transport in sewers, showing that ANFIS models had greater accuracy than empirical sediment transport rate equations. However, similar to ANN, ANFIS algorithms suffer from one important disadvantage; the lack of a systematic approach in the design of fuzzy rules and in

the choice of membership functions variables (Tien bui et al., 2016; Khosravi et al., 2018b). Kisi et al. (2012) compared the predictions of daily suspended sediment load made by ANFIS, ANN, Support Vector Machines (SVM) models and a genetic programming (GP) model in Cumberland River in the U.S. They revealed that GP provided more accurate predictions than the ANFIS, ANN and SVM models. Thus, meta-heuristic (or evolutionary) algorithms (e.g. particle swarm optimization, whale optimization algorithm) have been hybridized with ANFIS algorithms to overcome this weakness and improve the performance of ANFIS models. For example, ANFIS-meta-heuristic hybrid models have been applied to the prediction of groundwater potential mapping (Khosravi et al., 2018b; Termeh et al., 2019; Chen et al., 2019), flood susceptibility mapping (Tien Bui et al., 2018a,b) and sediment transport rate prediction (Qasem et al., 2017), revealing that this hybrid algorithm has a higher prediction power than ANFIS.

The use of other standalone and hybrid algorithms has also been explored in the field of sediment transport. For example, Ebtehaj et al. (2016a) applied a hybrid model of feed-forward neural network-extreme learning machine (FFNN-ELM) for open-channel sediment transport, revealing that this model outperformed GP and empirical sediment transport models. Similarly, they found wavelet-support vector machine (SVM-Wavelet) algorithms had a better prediction performance than SVM and existing empirical equations (Ebtehaj et al., 2016b). In the modeling of daily dissolved oxygen concentration in three US rivers, Haddam and Kisi (2018) applied least square SVM, multivariate adaptive regression splines (MARS), and M5 Model Tree (M5T) algorithms. Their study showed the dissolved oxygen concentrations were successfully predicted using all three models and that the best performing model differed from one measurement station to another.

Recently, new artificial intelligence algorithms, notably data mining algorithms, have been developed and applied in the fields of hydrology and hydraulics. These techniques do not seek

to explain the physical processes and mathematical reasoning for changes in environmental behavior but to recognize statistical patterns, both expected and unexpected, within data. These patterns can highlight environmental relationships in space and time that may unveil critical details about behavior, reveal previously unsuspected relationships, or mitigate uncertainty in estimates. Thus these types of techniques are at their most beneficial in situations when process-based models cannot be applied (e.g. lack of understanding of underlying physics of the process) or that suffer from inadequacies due to the limitation of data. Therefore artificial intelligence methods mean that some parameters which are difficult or expensive to measurement, such as bedload transport, could be easily predicted using other factors that are more readily available, such as discharge and bed slope. Such methods could be particularly attractive to developing nations where extensive measurement networks do not exist and there can be a lack of highly-skilled end-users to build and run more complex process-based models.

Recently some data mining algorithms, including random forest (RF), logistic model tree (LMT) and Naïve Bayes Trees (NBT) algorithms, have been applied for flood susceptibility mapping (Khosravi et al. 2018c), groundwater vulnerability assessments (Khosravi et al. 2018d), and landslide susceptibility mapping (Pham et al. 2018). The performance, accuracy, and reliability of these methods for spatial mapping have been proven in these fields. However these algorithms are rarely used for prediction and forecasting, not only for bedload transport rate prediction, but more generally in the field of geosciences. Only a few examples of their use exist, including the application of RF and Random Tree (RT) for solar radiation prediction (Sherafati et al. 2019), the Reduced Error Pruning Tree (REPT), Instance-Based K-nearest Neighbours (IBK) and M5P Model Tree techniques for suspended sediment load prediction (Khosravi et al. 2018a) and modeling dissolved oxygen concentration in rivers (Heddam and Kisi 2018). Thus a significant gap exists in the application of novel data mining algorithms for

bedload transport prediction 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) to experimental measure the bedload transport rate under uniform flow conditions; (2) produce predictions of bedload transport rate using novel data mining techniques, namely the M5P Model Tree, random tree (RT), random forest (RF) and the reduced error pruning tree (REPT), along with four types of hybrid algorithms trained with a Bagging (BA) data mining algorithm (BA-M5P, BA-RF, BA-RT and BA-REPT); (3) compare the predictive power of these data-driven models; and (4) perform a sensitivity analysis of the driving variables used in each model. This study is the first to apply a diverse range of data-mining models to the prediction of bedload transport. The research offers new insight into which data mining algorithms offer the potential to provide relatively cheap and fast predictions of bedload transport in poorly monitored rivers, where understanding of the physical processes at play may not be well understood.

2. Materials and Methods

2.1-Flume setup and experimental procedure

A total of 72 bedload transport experiments have been carried out in a 12 m long and a 0.5 m wide and deep tilting flume (Fig 1). A tailgate at the downstream end was adjusted to create uniform flow conditions over the mobile section of the flume, informed by water depth measurements made using two mechanical point gauges and three ultrasonic sensors. A 4 m upstream and 2.8 m downstream section of the flume was artificially roughened with the same-sized gravel as the mobile section to prevent upstream scour, promote the development of fully

developed flow within the mobile section of the flume and to reduce the backwater effect of the tailgate. The central 5 m section of the flume contained screeded, loose sediment with a thickness equal to 5-6 d_{50} (d_{50} is the median grain diameter of the sediment). A bedload trap, 0.5 m wide and 0.2 m long, was installed at the downstream end of the flume to sample the transported sediment through time. Four types of rounded and naturally-shaped, uniform-sized sediment were investigated (d_{50} of 5.17, 10.35, 14, and 20.7 mm) with a specific gravity of 2.39, 2.38, 2.90, and 2.55 respectively.

Before each experiment, the slope was set, tailgate raised, the pump turned on, and the flow slowly allowed to fill the flume without any disturbance to the bed. The tailgate was then opened, the flow discharge set, and after the establishment of uniform flow, sediment transport sampling commenced. The time of each experiment varied from 1 to 30 minutes, and the frequency of bedload sampling varied from several seconds to several minutes; both were dependent upon the bedload transport rate – the higher the rate, the higher the sampling frequency and the lower the sampling duration. The flow discharge (Q), velocity (V), depth (Y), bed slope (S) and sediment diameter (D) were measured for each experiment and these parameters were used to calculate relative roughness ($RR = D/Y$), shear stress ($\tau = \rho gRS$) and shear velocity ($V^* = \sqrt{\tau / \rho}$), where ρ is the water density, g is the gravitational acceleration and R is the hydraulic radius. The collected bed sediment was dried and weighed for calculation of the bedload sediment transport rate (q) in kg m^{-1} s^{-1} as follows:

$$q = G / bT \tag{1}$$

where G is the mass of collected sediment [kg], b is the flume width [0.5 m] and T is the sampling duration [s]. More information about the flume set-up and sediment sampling can be found in Khosravi et al. (2020).

Fig. 1. Experimental flume set-up (not to scale)

2.2-Development and application of data mining techniques

2.2.1. Sample size

72 experiments were undertaken; 50 experiments were used for model building and 22 were used for model validation. There is no universal guideline for the training-testing ratio, but a ratio of 70:30 is most commonly used in spatial and time series modeling (Khosravi et al., 2018a, d; Pham et al., 2017). Given the experimental dataset is relatively small, this approach was combined with a 10-fold cross validation technique. This technique is summarized in Fig 2. Each experimental dataset was used 10 times to provide 720 sets of data. In each iteration a different section of the dataset was considered as training and testing dataset.

Fig. 2. 10- fold cross validation technique.

2.2.2-Preparation of dataset

Eight factors, known to have a strong correlation with bedload transport rate, were used to perform the data-driven modeling: d_{50}, S, Y, V, Q, RR, V^* and τ. The correlation of these variables with bedload transport was investigated using the Pearson correlation coefficient (*PCC*). A total of 10 input variable combinations were constructed and investigated. These were assessed to determine the input combination that produced the most accurate prediction of bedload transport rate.

2.2.3-Model description

2.2.3.1-Random forest (RF)

The RF is a flexible, nonparametric, ensemble learning technique first developed by Breiman (2001), and is a hybrid procedure between a decision tree and a regression. (Breiman et al., 1984). The model is commonly used for both classification and regression problems; for example in drought forecasting (Deo et al., 2017), vegetation mapping prediction for changes in climate (Iverson et al., 2004) and soil moisture prediction (Prasad et al., 2018a, b).

In RF, each decision tree is constructed by selecting randomly a subset of a sample, selecting variables from a training dataset by using a deterministic algorithm (Mutanga et al., 2012; Deo et al., 2017), and using a random bootstrap sample for the training dataset to build multiple trees (Breiman et al., 1984). The RF algorithm is trained by means of several steps: (1) a bootstrap sample is drawn from the training data; (2) a decision tree is grown for each bootstrap sample by selecting the best split among the subset selected randomly from all the features, and the tree is then grown to the maximum size with no pruning back; (3) these aforementioned steps are repeated until a sufficiently large number of trees are created (Mutanga et al., 2012). The general structure of a RF model is shown in Fig 3.

Fig. 3. Random forest algorithm structure (Rodriguez-Galiano et al., 2016)

2.2.3.2 M5P

The M5P (also known as M5 Tree) model is a well-known piecewise linear tree-based model used to predict continuous variables and was first introduced by Quinlan (1992). Recent applications of M5P models can be found in several studies, such as in the prediction of dissolved oxygen (Heddam and Kisi, 2018) and suspended sediment load (Khosravi et al. 2018a). M5P is a flexible algorithm because the decision tree constructed by M5P can have multivariate linear models (Zhan et al., 2011). The M5P tree is developed through three main

steps: (1) constructing the tree; (2) pruning the tree; and (3) smoothing the tree. In the process of growing the tree using the M5P, the standard deviation reduction (*SDR*) is maximized to achieve the best model performance. The SDR is expressed as follows (Zhan et al., 2011):

$$SDR = SD(E) - \sum_i \frac{|E_i|}{|E|} x SD(E_i) \quad (2)$$

where E is defined as the set of cases, E_i is defined as the ith subset of cases which result from splitting the tree, $SD(E)$ is defined as the standard deviation of E, and $SD(E_i)$ is defined as the standard deviation of E_i.

The tree pruning step is started after the tree is constructed to eliminate undesired sub-trees. The purpose is to avoid data over-fitting problems that occur during the construction of the tree (over-fitting problems arise when the model is very accurate with the training dataset but fails with the testing dataset). In this pruning step, the attributes are reduced one by one to minimize the estimated error. The smoothing step is started after the tree pruning step and is performed to compensate for the sharp discontinuities between adjacent linear models at the leaves of the pruned tree (Wang and Witten, 1997). This step is achieved by using the leaf model to compute the predicted value, which is then filtered along the path back to the root node (Wang and Witten, 1997).

2.2.3.3 Reduced Error Pruning Tree (REPT)

The REPT model is well-known as a fast decision tree method that constructs a decision tree to reduce the error in the prediction (Mohamed et al., 2012). First, the model utilizes the regression tree logic to create multiple trees in various iterations (Jayanthi and Sasikala, 2013). Second, the model chooses the best tree (the one with the least error) from multiple trees. Third, the Reduced Error Pruning (REP) technique is used to prevent over-fitting problems. Finally, the

algorithm handles missing values using a $C_{4.5}$ algorithm, and sorts the values of numerical attributes using the embedded method.

The REPT algorithm uses a stopping criterion (the sum of squared errors) to build a tree with maximum information gain. The stopping criterion is expressed as follows (Quinlan, 1987):

$$S = \sum_{Eeleaves(RT)} q_c U_c \tag{3}$$

where q_c is defined as the class prediction and U_c is the leaf-within variance.

2.2.3.4 Random Trees (RT)

The RT model is formed by a stochastic process and builds the decision trees on a random subset of columns. The RT works in a similar manner to traditional decision trees but has one key exception; only a random subset of attributes is available for each split of the training dataset. The algorithm is a fast and flexible tree learner and has been applied to solve a broad range of problems, such as in philology (Najock and Heyde, 1982) and medicine (Busch et al., 2009).

Let T be a plane rooted decision tree with *m* nodes, referred to as a family tree, in which the profile of the tree might be described by the number of the nodes or the number of the leaves, . Suppose that *C* is a class of a plane rooted tree, and each $T \in C$, the size |*T*| by the number of nodes T includes a weight function $\omega(T)$, expressed as follows (Drmota and Gittenberger, 1997):

$$\omega(T) = \prod_{k \geq 0} \alpha_k^{m_k(T)} \tag{4}$$

where $(\alpha_k; k \geq 0)$ is defined as the non-negative numbers and $m_k(T)$ is defined as the number of the nodes $\upsilon \in T$ with out-degree k. Thereafter, set $a_m = \sum_{T:|T|=m} \omega(T)$ then the corresponding

generating function $a(t)=\sum_{m\geq 0} a_m t^m$ must satisfy the functional equation as follows (Drmota and Gittenberger, 1997):

$$a(t)=t\varphi\left(a(t)\right) \tag{5}$$

where $\varphi(x)=\sum_{k\geq 0}\varphi_k x^k$. In the final step, sets $C_m=\left\{T\in C:|T|=m\right\}$ are equipped with the probability distribution caused by the weight function $\omega(T)$ (Drmota and Gittenberger, 1997).

2.2.3.5 Bagging (BA)

BA is one of the most effective ensemble methods to solve classification and regression problems. The method is able to weaken the defects of component learners and raise the recognition rate of unstable classifiers. Thus it can enhance the predictive capability of the weak learners (Breiman, 1996). In the BA algorithm, the training process is carried out through three main steps: (1) selecting randomly and independently the data from the primary training dataset. This step is repeated several times to create a certain number of sub-datasets; (2) designating the base learning algorithm to train the various sub-datasets, and gain the sequence of predictive function; and (3) vote for the outcomes and select the final outcome with the most votes (Bauer and Kohavi, 1999). The BA method has been applied to improve many base learners such as trees (Mert et al., 2014), support vector machines (Pham et al., 2018) and Naïve Bayes trees (Pham and Prakash, 2017). In this study, the BA has been used to train the M5P, RF, RT, and REPT base learners for bedload transport rate prediction. The general structure of a Bagging model is shown in Fig 4.

Fig. 4. Bagging algorithm structure (Khosravi et al. 2018b)

2.3. Sensitivity analysis

There are two main steps in prediction using AI algorithms: (i) determination of the best input variable combination; and (ii) identifying the operator's optimum values. Each combination of input variables has a different impact on the modeled result, and thus the most effective input combination should be determined. There are no optimum operator values which work globally for model calibration. Hence, to enhance the prediction power of each algorithm, these values need to be 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 best input variable combination and optimum operator values were achieved by minimizing the Root Mean Square Error (*RMSE*) using trial and error during the testing phase. Also a sensitivity analysis was carried out to identify which input variables and model operators had the greatest effect on the predicted transport rate.

2.4. Model evaluation

The five most commonly used metrics for assessing the performance of models were used: coefficient of determination (R^2), *RMSE*, Mean Absolute Error (*MAE*), Nash-Sutcliffe Efficiency (*NSE*) and percent bias (*PBIAS*). These metrics were calculated as follows (Moriasi et al. (2007; Dawson et al. 2006; Legates et al. (1999):

$$R^2 = \left(\frac{\sum_{i=1}^{n}(X_o - \overline{X}_o)(X_e - \overline{X}_e)}{\sqrt{\sum_{i=1}^{n}(X_o - \overline{X}_o)^2 \sum_{i=1}^{n}(X_e - \overline{X}_e)^2}}\right)^2 \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_e - X_o)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_e - X_o\right| \tag{8}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(X_e - X_o)^2}{\sum_{i=1}^{n}(X_o - \overline{X}_o)^2} \tag{9}$$

$$PBIAS = \left(\frac{\sum_{i=1}^{n}(X_o - X_e)}{\sum_{i=1}^{n}X_e}\right) * 100 \tag{10}$$

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 1.

Table.1. Performance classification of the model evaluation metrics

For a visual analysis and assessment of the applied models, Taylor diagrams and box-plots were used. One distinct advantage of the Taylor diagram is that it benefits from the use of the two most common correlation statistics: correlation coefficient (*CC*) and standard deviation (*SD*) (Taylor, 2001). The closer the predicted value to the observed value in terms of the *CC* and *SD*, the higher the prediction capability (Sigaroodi et al. 2014). The advantage of a box-plot is that it can show how well a model predicts extreme, median and quartile values.

3. Results

3.1. Determination of the best input variable combination

The *PCC* values in Table 2 show that flow velocity had the highest impact on bedload transport rate (*PCC* = 0.760), followed by shear stress (*PCC* = 0.709), flow discharge (*PCC* = 0.668), shear velocity (*PCC* = 0.663), bed slope (*PCC* = 0.490), flow depth (*PCC* = 0.303), d_{50} (*PCC* = 0.247), and relative roughness (*PCC* = 0.0033).

Table 2. Pearson correlation coefficient between input variables and bedload transport rate.

Based on these *PCC* values, ten different input combinations were constructed and investigated:

1. $BL=f(Q)$
2. $BL=f(V)$
3. $BL=f(V, \tau)$
4. $BL=f(V, \tau, Q)$
5. $BL=f(V, \tau, Q, V^*)$
6. $BL=f(V, \tau, Q, V^*, S)$
7. $BL=f(V, \tau, Q, V^*, S, Y)$
8. $BL=f(V, \tau, Q, V^*, S, Y, d_{50})$
9. $BL=f(V, \tau, Q, V^*, S, Y, d_{50}, RR)$
10. $BL=f(V, \tau, V^*, S, Y, d_{50}, RR)$

This approach starts with the variable with the highest *PCC* (*Q*) and then variables with lower *PCC's* are added into the combination until the variable with the lowest *PCC* (*RR*) is finally added. In each model, all 10 combinations were used in the training and testing phases. In the testing phase *RMSE* and *PCC* values were calculated to determine the optimal combination for building the final version of the model.

Table 3 shows that, due to the different structures of each model, the optimal input variable combination differs between the models. Input combination 10, in which all variables except *Q* were considered, was the best input combination for RF, RT and BA models, but input combination 8, in which all variables except *RR* were considered, was the best for M5P and REPT models. Adding *RR* to the input combination caused an increase the error in most of the cases (M5P, REPT, RT models), but with the RF and BA models it produced a better performance (compare the input combinations of 8 and 9). Including *Q* in the input combination caused an increase in model error (compare the input combinations of 3 and 4; this results explainswhy input combination 10 was created with all input variables except *Q*).

The input combinations constructed with variables with high *PCC* values (3, 4, 5, and 6), did not produce a good agreement between observed and predicted values. Indeed by adding d_{50} into a combination, which had a low *PCC*, the prediction power of the models increased significantly (compare the input combinations of 7 and 8). A comparison of input combinations 8-10 with 2-6 shows overall that including variables with low *PCC* (such as *RR*) and removing variables with high *PCC* improved prediction performance,. Thus, these results confirm that input variable selection must be carried out on a trial and error basis.

Table 3. Determination of the best input variable combination during the testing phase.

3.2. Model performance and sensitivity analysis

3.2.1. RF model

Twelve operators were considered for the RF model and their optimum values were obtained using a trial and error approach. These operators were bag size percent (percentage of the training set used), batch size (preferred number of instances to process if batch prediction is being performed), maximum depth of tree, number of decimal places used in the output of numbers in the model, number of execution slots used for constructing the ensemble, number of features (randomly chosen attributes), number of iterations to be performed, seed number, break ties randomly when several attributes look equally good, calculation of out of bag error, de bug and do not check capabilities (classifier may output additional information to the console, if set). The optimum values for these operators were 100, 100, 8, 2, 2, 0, 100, 1, no, no, no and no, respectively.

The sensitivity analysis shows that the maximum depth of tree had the biggest impact on the *RMSE* of the RF model (Fig. 5), followed by bag size percent, number of features, number of seeds, number of iterations, batch size and number of execution slots, respectively.

Fig. 5. Sensitivity analysis and identification of optimum operator values for the RF model

3.2.2. M5P model

Six main operators were established in the structure of the M5P model. These parameters were batch size, minimum number of instances to allow at a leaf node, number of decimal places used in the output of numbers in the model, build regression tree (whether to generate a

regression tree/rule instead of a model tree/rule), do not check capabilities and unpruned (whether unpruned tree/rule is to be generated). The optimum values for these operators were 100, 4, 2, no, no, and no, respectively. The sensitivity analysis shows that none of these operators had an impact on the predictive capability of the M5P model, and thus the default values were used (Fig. 6).

Fig. 6. Sensitivity analysis and identification of optimum operator values for the M5P model

3.2.3. REPT model

Eight main operators were considered in the REPT model: batch size, initial count (initial class value count), maximum depth, minimum number (the minimum total weight of the instance in a leaf), minimum variance probability (the minimum proportion of variance on all the data that need to be presented at a node in order for splitting to be performed in a regression tree), number of decimal places, number of folds (the amount of data used for back-fitting) and number of seeds. The optimum values for these operators were 100, 1, -1, 2, 0.001, 2, 3, and 1 respectively.

All of the operators except the initial count and batch size had a noticeable impact on the predictive power of the REPT model (Fig. 7). The most sensitive operator was the maximum depth of tree (same as observed with the RF model).

Fig. 7. Sensitivity analysis and identification of optimum operator values for the REPT model: (a) operators, (b) minimum variance probability, and (c) maximum depth

3.2.4. RT model

Eight operators were used in the RT model: K value (sets the number of randomly chosen attributes), batch size, maximum depth, minimum number (the minimum total weight of the instance in a leaf), minimum variance probability, number of decimal places, number of folds and number of seeds. The optimum values for these operators were 2, 100, -1, 1, 0.001, 1, 0 and 3, respectively (Fig. 8).

All the operators, except the batch size and number of decimal places, had a noticeable effect on model performance, (Fig. 8a). The maximum depth of tree and minimum variance probability had the most significant impact (Fig. 8b and c).

Fig. 8. Sensitivity analysis and identification of optimum operator values for the RT model: (a) some operators, (b) minimum variance probability, and (c) maximum depth.

3.2.5. BA model

Six operators were used in the BA model: bag size percent, batch size, number of decimal places, number of execution slots, number of iterations number of seeds. The optimum values for these operators were 20, 100, 2, 0, 12 and 0 respectively (Fig. 9). Of these six operators, only the number of iterations (most sensitive operator) and the number of seeds had a significant effect on the RMSE of the BA model.

Fig. 9. Sensitivity analysis and identification of optimum operator values for the bagging model

3.3. Model performance assessment

After the determination of the most effective input variable combination and the optimum operator values, each algorithm was trained by a training dataset and evaluated by a testing dataset. Since the models were built by a training dataset, this evaluation can only show how well the constructed model fits the testing dataset, and cannot be used for model validation (Khosravi et al, 2016; Chen et al, 2019).

An assessment of the predictive capability of the eight developed models o is shown in Table 4. The R^2 values show that the BA-M5P model had the highest prediction power (0.943) followed by the M5P (0.932), BA-RT (0.910), RT (0.890), BA-RF (0.833), RF (0.784), BA-REPT (0.596), and REPT (0.570). According to the classification of performance for this metric (Legates and McCabe, 1999; Moriasi et al, 2007; Ayele et al., 2017), all models had a 'very good' performance except the BA-REPT and REPT models which had 'satisfactory' performance. Since R^2 is standardized for differences between the mean and variance of observed and predicted values, it 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. In terms of *RMSE*, *MAE,* and *NSE*, the BA-M5P was superior to the other models. Using the *NSE* values, all of the applied models had a 'very good performance' except the BA-REPT and REPT which were 'satisfactory' and 'acceptable' respectively. According to the *PBIAS* values, all the models had a 'very good' performance except BA-RT, RT, and REPT. The M5P and BA-RT, and RT models underestimated (shown by positive *PBIAS* values) and the BA-M5P, BA-RF, RF, BA-REPT, and REPT models overestimated the bedload transport rate.

Table 5. Evaluation of model performance

A Taylor plot of model performance shows that the BA-M5P model had the highest performance because the predicted standard deviation of bedload transport rate is the closest to the standard deviation of the observed data, and the correlation is also the highest (Fig. 10). Considering all the evaluation metrics together, the BA-M5P and REPT moidels had the highest and the lowest prediction capability.

Fig. 10. Taylor plot of model performance

A further comparison between observed and predicted bedload transport rate is shown in Figures 11 and 12. Both figures confirm that the BA-M5P model predicts the bedload rate more accurately than the other models. Generally, the results show that the prediction power of the hybrid algorithms mostly depends on the base algorithm (i.e. M5P, RF, and so on). For example the incorporation of the Bagging algorithm increased the prediction power of REPT, but the BA-REPT algorithm still had a lower prediction power than M5P, RF, and RT as standalone algorithms.

Fig. 11. Observed and predicted bedload transport rate in the testing phase: a) M5P, b) RF, c) RT, d) REPT, e) BA-M5P, f) BA-RF, BA-RT, and BA-REPT models.

Fig. 12. Scatter plots of observed versus predicted bedload transport rate in the testing phase: a) M5P, b) RF, c) RT, d) REPT, e) BA-M5P, f) BA-RF, BA-RT, and BA-REPT models.

The box plots of bedload transport rates (Fig. 13) show that the BA-RT and RT models produced a perfect match for the maximum observed maximum transport rate (1.06 kg m^{-1} s^{-1}), and BA-REPT and REPT (1.05 kg m^{-1} s^{-1}) and BA-M5P (1.03 kg m^{-1} s^{-1}) produced a close match. The other algorithms could not predict the rate accurately. The same result was observed for predicting the minimum observed transport rate. These two results reveal that, although M5P is overall the most accurate for the testing dataset, it cannot predict extreme values well. However when trained with a Bagging data mining algorithm, its performance in predicting extreme values is much improved. The BA-RF and RF models predict the observed values in the third quartile better than the other models, the BA-RT and RT better predict the median values, and the BA-REPT and REPT models better predict values in the first quartile.

Fig. 13. Box plots of observed and predicted bedload transport rates.

4. Discussion

The determination of the best input variable combination and optimum operator values is one of the most significant steps in producing an accurate data mining model. Some researchers have determined the best input combination using a Principal Component Analysis approach or according to the strongest *PCC* (Barzegar et al, 2016a,b). However the current paper show these approaches might not be the best to take. Due to nonlinearity between variables, the variables with low *PCC* enhanced the prediction power of the models, and the most effective combination varied from only model to another. Thus a range of different input variable combinations must be considered in the optimization of data mining models.

The M5P, RF, RT, and REPT standalone models had contrasting prediction performance. Given the same bedload dataset was used to test performance, this contrast results from a difference

in each model's structure (Loh, 2011), particularly their flexibility, computing capability, complexity, and ability to reduce over-fitting (Kisi et al. 2019). The hybrid models built using a Bagging algorithm, in most of the cases, had a higher prediction power than standalone models because hybrid models are more flexible than standalone models and have a nonlinear structure (De'ath and Fabricius, 2000). These two model properties are particularly important in the prediction of bedload transport because of the nonlinearity between variables. In particular the Bagging algorithm benefits from ensemble learning in its structure (multiple weak learners) which outperforms a single strong learner. This learning helps to reduce variance and avoid the over-fitting problem caused by the use of a bootstrap procedure.

Given this is the first study to examine the prediction performance of these data mining algorithms for the prediction of bedload transport rate, no direct comparisons exist. However, the improved performance of hybrid Bagging models conforms with previous tests of data mining algorithms in other fields (Khosravi et al, 2018a; Khozani et al, 2019; Sherafti et al, 2019; Ghorbani et al., 2017; Yaseen et al., 2017). Furthermore, other studies, in the prediction of water quality (nitrate, ammonia and phosphate) (Shkurin et al. 2015), suspended sediment load and (Francke et al. 2008) and solar radiation (Sun et al, 2016; Sherafati et al, 2019), have also found RF and RT algorithms have very good prediction power.

The M5P model was the best performing standalone and hybrid model because it is a decision tree based algorithm that does not have a hidden layer in its structure (Kisi et al., 2012), and thus can learn efficiently and does not require any assumptions about the type of data distribution. The main advantage of the M5P algorithm over other decision tree algorithms is that M5P has an ability to handle data without any vagueness, as well as handle very large datasets with a large number of dimensions and attributes (Quinlan, 1992). Also, the M5P model benefits from model trees, that are much smaller than regression trees and have proven to be more accurate (Quinlan, 1992). Although RF and RT models have some advantages, such as being simple, flexible and

easy to use, they have one severe limitation; the construction of a large number of trees can make the algorithm slow and ineffective for real-time predictions, and a more accurate prediction requires more trees. In addition, the RF modle suffers from poor performance for regression subjects, especially when the range of testing data is out of the range of the training dataset.

Overall, the results show that M5P models, especially those trained with a Bagging data mining algorithm, have great potential to produce robust predictions of bedload transport in gravel-bed rivers. Such models could be particularly useful in data-poor watersheds, especially in developing nations where technical skills and understanding of the processes occurring in the watershed may be lacking. The M5P models could potentially be used alone or replace process-based models because they represent well the highly stochastic behavior of sediment transport and are inexpensive to build and run. This type of data-driven model could also complement existing process-based models in well-gaged watersheds to recognize patterns within collected data that could unveil critical details about behavior, reveal previously unsuspected environmental relationships, or mitigate uncertainty in model estimates. Future studies should consider the performance of these algorithms in the prediction of bedload transport in more complex conditions than those studied here, such as with poorly-sorted sediments, water-worked beds that mimic better the surface topographies of natural coarse-grained rivers (Cooper and Tait, 2009), unsteady flows and in non-equilibrium transport conditions in the case of an upstream sediment supply (Mao et al., 2012).

5. Conclusions

The accurate prediction of bedload transport rate is vital for understanding gravel-bed river morphodynamics. Due to the non-linear and chaotic behavior of bedload transport in a river, data mining and machine learning algorithms have great potential to produce accurate

predictions of bedload transport rate. Using bedload transport data collected in laboratory flume experiments, this p study tested this potential for the first time by examining the prediction power of standalone M5P, random tree (RT), random forest (RF) and reduced error pruning tree (REPT) models, as well as these models trained with a Bagging (BA) algorithm: BA-M5P, BA-RF, BA-RT and BA-REPT. The main findings were as follows:

(1) A test of model performance showed that the BA-M5P model had the highest prediction power followed by M5P, BA-RT, RT, BA-RF, RF, BA-REPT, and REPT. All models displayed 'very good' performance except the BA-REPT and REPT model, which were 'satisfactory'.

(2) The M5P, BA-RT, and RT models underestimated, and the BA-M5P, BA-RF, RF, BA-REPT and REPT models overestimated bedload transport rates prediction.

(3) A sensitivity analysis revealed that flow velocity had the biggest impact on the bedload transport rate followed by shear stress, flow discharge, bed shear velocity, bed slope, flow depth, median sediment diameter and relative roughness.

(4) The input combination that included all variables except relative roughness was found to be provide the best predictive capability for the M5P and REPT models, while the input combination without flow discharge provided the best accuracy for the RF, RT and BA models.

(5) The maximum depth of tree was the most sensitive operator in decision tree-based algorithms and batch size, number of execution slots and number of decimal places did not have any impact on the model prediction power.

Overall the results revealed that hybrid data mining techniques provide more accurate predictions of bedload transport rate than standalone data mining models. These models could

be especially important in data-poor catchments, particularly in developing nations, where technical skills and understanding of the processes occurring in the catchment may be lacking. In this case, understanding more about the potential for data mining algorithms to provide relatively cheap and fast predictions of non-linear processes represents a vital research frontier for river scientists.

Acknowledgments

We would like to thank the editor (Prof. András Bárdossy) associate editor and three anonymous reviewers for their invaluable comments and suggestions which improved the quality of the paper. The authors are grateful to the Iranian Ministry of Science, Research and Technology for the grant to carry out this research. Also, the authors wish to thank the University of Guilan (Faculty of Engineering), Dr. Amir H.N Chegini and Mr. Nader Izadpanah for their cooperation.

References

Ayele, G.T., Teshale, E.Z., Yu, B., Rutherfurd, I.D., Jeong, J. 2017. Stream flow and Sediment Yield Prediction for Watershed Prioritization in the Upper Blue Nile River Basin, Ethiopia. Water, 9,782; doi:10.3390/w9100782.

Barzegar, R., Moghaddam, A. A., Adamowski, J., and Fijani, E. 2016a. Comparison of machine learning models for predicting fluoride contamination in groundwater. Stochastic Environmental Research and Risk Assessment, 31(10), 2705-2718.

Barzegar, R., Adamowski, J., Moghaddam, A. A. 2016b. Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran. Stochastic Environmental Research and Risk Assessment, 30:1797–1819

Bauer, E., Kohavi, R. 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, pp. 105-139.

Breiman, L. 1996. Bagging predictors, Machine Learning 24, pp. 123-140.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Chapman and Hall, London.

Busch JR, Ferrari PA, Flesia AG, Fraiman R, Grynberg SP, Leonardi F. 2009. Testing statistical hypothesis on random trees and applications to the protein classification problem. J. Appl. Statist., 3: 542-563.

Chen, W., Panahi, M., Khosravi, K., Pourghasemi, HR., Rezaei, F. 2019. Spatial prediction of groundwater potentiality using ANFIS ensembled with teaching-learning-based and biogeography-based optimization. Journal of Hydrology, 572, 435-448

Cooper, J. R., & Tait, S. J. 2010. Examining the physical components of boundary shear stress for water-worked gravel deposits. Earth Surface Processes and Landforms, 35(10), 1240-1246. doi:10.1002/esp.2020.

Deo, R., C, Kisi, O., Singh, V.P. 2017. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model, Atmos. Res. 184, pp. 149–175.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Drmota, M., Gittenberger, B. 1997. On the profile of random trees. Random Structures Algorithms, 10: 421-451.

Ebtehaj, I and Bonakdari, H. 2013. Evaluation of Sediment Transport in Sewer using Artificial Neural Network. Engineering Applications of Computational Fluid Mechanics. 7, 3: 382-392

Ebtehaj, I and Bonakdari, H. 2014. Performance Evaluation of Adaptive Neural Fuzzy Inference System for Sediment Transport in Sewers. Water Resources Management, 28, 13:4765–4779.

Ebtehaj, I., Binakdari, H., Shamshirband, Mohamadi, K. 2016b. A combined support vector machine-wavelet transform model for prediction of sediment transport in sewer. Flow Measurement and Instrumentation, 47:19-27.

Ebtehaj, I., Binakdari, H., Shamshirband, S. 2016a. Extreme learning machine assessment for estimating sediment transport in open channels. Engineering with Computers, 32 (4):691-704.

Einstein, H. A. (1950). The bed-load function for sediment transportation in open channel flows (No. 1026). Washington, D.C.: US Department of Agriculture.

Engelund, F., & Hansen, E. 1967. Monograph on sediment transport in alluvial streams. Copenhagen: Teknisk forlag.

Francke, T., L´opez-Tarazon, J.A. and Schroder, B. 2008. Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. Hydro. Process., 22: 4892–4904.

Ghafouri Azar, M., Namaee, M.R., Rostami, M. 2012. Evaluating a numerical model to simulate the variation of river bed due to a mining pit based on the experimental data. Asian J. of Appl. Sci., 5(3):154-163.

Ghorbani, M.A., Deo, R.C., Yaseen, Z.M., H. Kashani, M., Mohammadi, B., 2017. Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. Theoretical and Applied Climatology. doi:10.1007/s00704-017-2244-0

Graf, W. H. 1971. Hydraulics of Sediment Transport. New York: McGraw-Hill.

Hamel P, Falinski K, Sharp R, et al. 2017. Sediment delivery modeling in practice: Comparing the effects of watershed characteristics and data resolution across hydroclimatic regions. Sci Total Environ. doi: 10.1016/j. Sci. Tot. Env.,12.103.

Heddam, S., Kisi, O. 2018. Modelling Daily Dissolved Oxygen Concentration Using Least Square Support Vector Machine, Multivariate Adaptive Regression Splines and M5 model Tree. J. of Hydro., , https://doi.org/10.1016/j.jhydrol.2018.02.061.

Howard, H.C. 2008. River Morphology and River Channel Changes. Transactions of Tianjin University, 14:254-262.

Iverson, L., Prasad, A., Liaw, A. 2004. New machine learning tools for predictive vegetation mapping after climate change: Bagging and Random Forest perform better than regression tree analysis. Proceedings, UK-International Association for Landscape Ecology, Cirencester, UK, 317-320.

Jayanthi S, Sasikala S. 2013. Reptree classifier for identifying link spam in web search engines. J. Soft. Comput., 3: 498-505.

Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Tien Bui, D., 2018c. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ., https://doi.org/10.1016/j.scitotenv.2018.01.266

Khosravi, K., Mao, L., Kisi, O., Yaseen, Z., 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.

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 & Earth System Sciences, 22 (9).

Khosravi, K., Sartaj, M., Tsai, FTC., Singh, VP., Kazakis, N., Melesse and et al. 2018d. A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment. Science of the Total Environment, 642, 1032-1049.

Khosravi, K., Chegini, A.H.N., Cooper, J.R., Daggupati, P., Binns, A.D., Mao, L. 2019. Uniform and graded bed-load sediment transport in a degrading channel with non-equilibrium conditions. International Journal of Sediment Research (In press). https://www.sciencedirect.com/science/article/pii/S1001627918302750.

Khozani, Z., Khosravi, K., Pham, B., Kløve, B., Wan Mother, H., Yaseen, Z. 2019. Determination of compound channel apparent shear stress: application of novel data mining models. Journal of Hydroinformatics, 21 (5): 798-811.

Kisi, O., Dailr, A.H., Cimen, M., Shiri, J., 2012. Suspended sediment modeling using genetic programming and soft computing techniques. J. Hydro., 450, 48–58

Kisi, O., Genc, O., Dinc, S., Zounemat-Kermani, M. 2016. Daily pan evaporation modeling using chi-squared automatic interaction detector, neural networks, classification and regression tree. Comput. Electr. Agric., 122: 112–117

Kisi, O., Heddam, S., Yaseen, Z.M., 2019. The implementation of univariable scheme-based air temperature for solar radiation prediction: New development of dynamic evolving neural-fuzzy inference system model. Applied Energy, 241, 184–195.

Legates, D.R., McCabe, G.J. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Res. Res., 35: 233–241.

Loh W-Y. 2011. Classification and regression trees. WIREs Data Min. Knowl. Discov.. doi:10.1016/0169-7439(91)80113-5.

Mao, L. 2012. The effect of hydrographs on bed load transport and bed sediment spatial arrangement. Journal of Geophysical Research, 117, 374–386

Mao, L., Cooper, J. R., & Frostick, L. E. 2012. Grain size and topographical differences between static and mobile armour layers. Earth Surface Processes and Landforms, 36(10), 1321-1334. doi:10.1002/esp.2156

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. Agric. Water Manag.. https://doi.org/10.1016/j.agwat.2010.12.012

Mert A, Kılıç N, Akan A. 2014. Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats. Neural .Comput. Appl., 24: 317-326.

Meyer-Peter, E., & Müller, R. 1948. Formulas for bed-load transport. Proc. 2nd congress of IAHR, Stockholm, Sweden, 39-64.

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. Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on, IEEE, pp. 392-397.

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. https://doi.org/10.13031/2013.23153.

Mustafa, A.S., Sulaiman, S.O., Al_Alwani, K.M. 2017. Application of HEC-RAS Model to Predict Sediment Transport for Euphrates River from Haditha to Heet. J. Eng. Sci., 20 (3):570-577.

Mutanga, O., Adam, E., Cho, M.A. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int. J. Appl. Earth Observ.Geoinfo., 18: 399-406.

Najock D, Heyde C. O. 1982. The number of terminal vertices in certain random trees with an application to stemma construction in philology. J. Appl.Prob. ; 19: 675-680.

Pham, B., Prakhsh, I., Khosravi, K., Chapi, K., Trinh, P, Hosseini, V. 2018. A comparison of Support Vector Machines and Bayesian algorithms for landslide susceptibility modelling. Geocarto International, 34:13, 1385-1407, doi: 10.1080/10106049.2018.1489422

Pham, B.T., Khosravi, K., Prakhsh, I. 2017. Application and comparison of decision tree-based machine learning methods in landside susceptibility assessment at Pauri Garhwal Area, Uttarakhand, India. Environ. Processes, 4 (3):711–730

Pham, B.T., Prakash, I. 2018. Bagging based Support Vector Machines for spatial prediction of landslides. Environ. Earth Sci., 77-146.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T. 2018a. Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. Soil Tillage Res., 181: 63-81.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T. 2018b. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition, Geoderma, 330:136-161.

Qasem, S.N., Ebtehaj, I., Madavar, H. 2017. Optimizing ANFIS for sediment transport in open channels using different evolutionary algorithms. Journal of Applied Research in Water and Wastewater, 4:290-298.

Quinlan, J.R. 1992. Learning with continuous classes. 5th Australian joint conference on artificial intelligence, Singapore, 343-348.

Quinlan, J.R., 1987. Simplifying decision trees. Int. J. Man Mach. Stud., 27, 221–234.

Rodriguez-Galiano , V., Sanchez-Castillo, M., Dash, J., Atkinson, P., and Ojeda-Zujar, J. 2016. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. Biogeosciences, 13(11):3305-3317

Shahiri, P., Noori, M., Heydari, M., Rashidi, M. 2016. Floodplain Zoning Simulation by Using HEC-RAS and CCHE2D Models in the Sungai Maka River. Air, Soil. Water Res., 9:55-62.

Sharafati, A., Khosravi, K., Khosravinia, P., Ahmed, K., Salman, S., Yaseen, Z. 2019. The potential of novel data mining models for global solar radiation prediction. International Journal of Environmental Science and Technology, 1-18

Shiri J, Kişi Ö. 2012. Estimation of Daily Suspended Sediment Load by Using Wavelet Conjunction Models. J. Hydrol. Eng., 17: 986–1000. doi: 10.1061/(ASCE)HE.1943-5584.0000535

Shkurin, A. 2015. Water quality analysis using machine learning algorithms. Bachelor's Thesis in Environmental Engineering, MAMK University of applied science, 54 pp.

Sigaroodi, SK., Chen, Q., Ebrahimi, S., et al. 2014. Long-term precipitation forecast for drought relief using atmospheric circulation factors: A study on the Maharloo basin in Iran. Hydrol Earth Syst. Sci., doi: 10.5194/hess-18-1995-2014

Sun, H., Gui, D., Yan, B., Liu, L., Liao, W., Zhu, Y., Lu, C., Zhao, N. 2016. Assessing the potential of random forest method for estimating solar radiation using air pollution index. Energy Convers. Manag., 119:121-129.

Taylor, K.E. 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos., 106:7183–7192. doi: 10.1029/2000JD900719

Termeh, V., Khosravi,K., Sartaj, M., Keesstra, SD., Tsai, FTC., R Dijksma and Pham, B. 2019. Optimization of an adaptive neuro-fuzzy inference system for groundwater potential mapping. Hydrogeology Journal, 1-24.

Tien Bui, D., Khosravi, K., Li, S., Shahabi, H., Panahi, M., Singh, V., Chapi, K., 2018a. New hybrids of anfis with several optimization algorithms for flood susceptibility modeling. Water, 10(9),121.

Tien Bui, D., Panahi, M., Shahabi, H., Singh, V., Shirzadi, A., Chapi, K., Khosravi, K. 2018b. Novel hybrid evolutionary algorithms for spatial prediction of floods. Scientific Reports 8 (1), 15364

Tien Bui, D., Pradhan, B., Nampak, H., Bui, Q.-T., Tran, Q.-A., Nguyen, Q.-P., 2016. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. J. Hydrol., https://doi.org/10.1016/j.jhydrol.2016.06.027

Wang, Y., Witten, I.H., 1997. Inducing model trees for continuous classes. Proceedings of the Ninth European Conference on Machine Learning, pp. 128-137.

Wilcock, P.R. 1998. Two-fraction model of initial sediment motion in gravel-bed rivers. Sci., 280(5362), 410-412.

Wilcock, P.R., Crowe, J.C. 2003. Surface-based transport model for mixed-size sediment. J. Hydral. Eng., 129(2), 120-128.

Yaseen, Z.M., Ebtehaj, I., Bonakdari, H., C.Deo, R., Danandeh Mehr, A., Wan Mohtar, H., Diop, L., El-Shafie, A., P. Singh, V. 2017. Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. J. Hydro., http://dx.doi.org/10.1016/j.jhydrol.2017.09.007.

Zhan C, Gan A, Hadi M. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. IEEE transactions on intelligent transportation systems 2011; 12: 1549-1557.

Highlights

- Eight standalone and hybrid data mining algorithms were used to predict bedload transport rate in a laboratory flume
- Flow velocity had the largest impact on the bed load transport rate
- BA-M5P hybrid algorithm had the highest prediction performance
- All data mining models displayed ‘very good’ prediction performance except the BA-REPT and REPT models

Table 1. Performance classification of the model evaluation metrics

Objective function	Value range	Performance classification	References
R^2	$0.7 < R^2 < 1$	Very good	Moriasi et al. (2007); Ayele et al. (2017)
	$0.6 < R^2 < 0.7$	Good	
	$0.5 < R^2 < 0.6$	Satisfactory	
	$R^2 < 0.5$	Unsatisfactory	
RMSE		The lower the *RMSE*, the better the model performance	Dawson et al. (2006)
MAE		The low 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	

Table 2. Pearson correlation coefficient between input variables and bed load transport rate

Variable	d_{50}	S	Y	V	Q	RR	V^*	τ
PCC	0.247	0.490	0.303	0.760	0.668	0.003	0.663	0.709

Table 3. Determination of the best input variable combination during the testing phase

Models	Evaluation criteria	Input 1	Input 2	Input 3	Input 4	Input 5	Input 6	Input 7	Input 8	Input 9	Input 10
RF	*RMSE* (kg m^{-1} s^{-1})	0.223	0.222	0.254	0.239	0.256	0.223	0.224	0.160	0.140	0.126
	PCC	0.551	0.602	0.855	0.615	0.587	0.663	0.660	0.801	0.846	0.878
M5P	*RMSE* (kg m^{-1} s^{-1})	0.209	0.188	0.189	0.191	0.186	0.191	0.186	0.077	0.078	0.078
	PCC	0.617	0.710	0.714	0.704	0.723	0.706	0.723	0.957	0.955	0.955
REPT	*RMSE* (kg m^{-1} s^{-1})	0.218	0.230	0.285	0.285	0.285	0.285	0.285	0.194	0.197	0.198
	PCC	0.566	0.640	0.485	0.485	0.485	0.485	0.485	0.755	0.749	0.746
RT	*RMSE* (kg m^{-1} s^{-1})	0.26	0.25	0.337	0.335	0.340	0.212	0.231	0.135	0.172	0.144
	PCC	0.501	0.53	0.521	0.526	0.508	0.726	0.536	0.865	0.810	0.873
BA	*RMSE* (kg m^{-1} s^{-1})	0.220	0.21	0.205	0.208	0.208	0.208	0.216	0.106	0.080	0.068
	PCC	0.630	0.582	0.625	0.612	0.612	0.613	0.571	0.878	0.923	0.960

Table 4. Evaluation of model performance

Models	R^2	*RMSE* (kg m^{-1} s^{-1})	*MAE* (kg m^{-1} s^{-1})	*NSE*	*PBIAS*
BA-M5P	0.943	0.061	0.040	0.945	-1.60
M5P	0.932	0.077	0.055	0.914	7.30
BA-RF	0.833	0.107	0.070	0.832	-2.58
RF	0.784	0.122	0.080	0.782	-4.18
BA-RT	0.910	0.080	0.057	0.9055	11.24
RT	0.890	0.090	0.060	0.881	12.08

BA-REPT	0.596	0.183	0.101	0.510	-9.40
REPT	0.570	0.194	0.110	0.449	-11.35

Conflicts of Interest: The authors declare no conflict of interest

Conceptualization: Khabat Khosravi, James R. Cooper, Prasad Daggupati. **Data curation:** Khabat Khosravi, Prasad Daggupati. **Methodology:** Khabat Khosravi, James R. Cooper, Prasad Daggupati, Binh Thai Pham, Dieu Tien Bui. **Writing—original draft preparation**: Khabat Khosravi, Prasad Daggupati, Binh Thai Pham. **Writing—review and editing:** James R. Cooper, Dieu Tien Bui.

Figure Captions

Fig. 1. Experimental flume set-up (not to scale)

Fig 2. k fold cross validation technique (https://en.wikipedia.org/wiki/Cross-validation_(statistics))

Fig. 3. Random forest algorithm structure (Rodriguez-Galiano et al. 2016)

Fig. 4. Bagging algorithm structure (Khosravi et al. 2018b)

Fig. 5. Sensitivity analysis and identification of optimum operator values for the RF model

Fig. 6. Sensitivity analysis and identification of optimum operator values for the M5P model

Fig. 7. Sensitivity analysis and identification of optimum operator values for the REPT model: (a) operators, (b) minimum variance probability, and (c) maximum depth

Fig. 8. Sensitivity analysis and identification of optimum parameter values for the RT model: (a) operators, (b) minimum variance probability, and (c) maximum depth

Fig. 9. Sensitivity analysis and identification of optimum operator values for the bagging model

Fig. 10. Taylor plot of model performance

Fig. 11. bserved and predicted bed load transport rate in the testing phase: a) M5P, b) RF, c) RT, d) REPT, e) BA-M5P, f) BA-RF, BA-RT, and BA-REPT models.

Fig. 12. Scatter plots of observed versus predicted bed load transport rate in the testing phase: a) M5P, b) RF, c) RT, d) REPT, e) BA-M5P, f) BA-RF, BA-RT, and BA-REPT models.

Fig. 13. Box plots of observed and predicted bedload transport rate

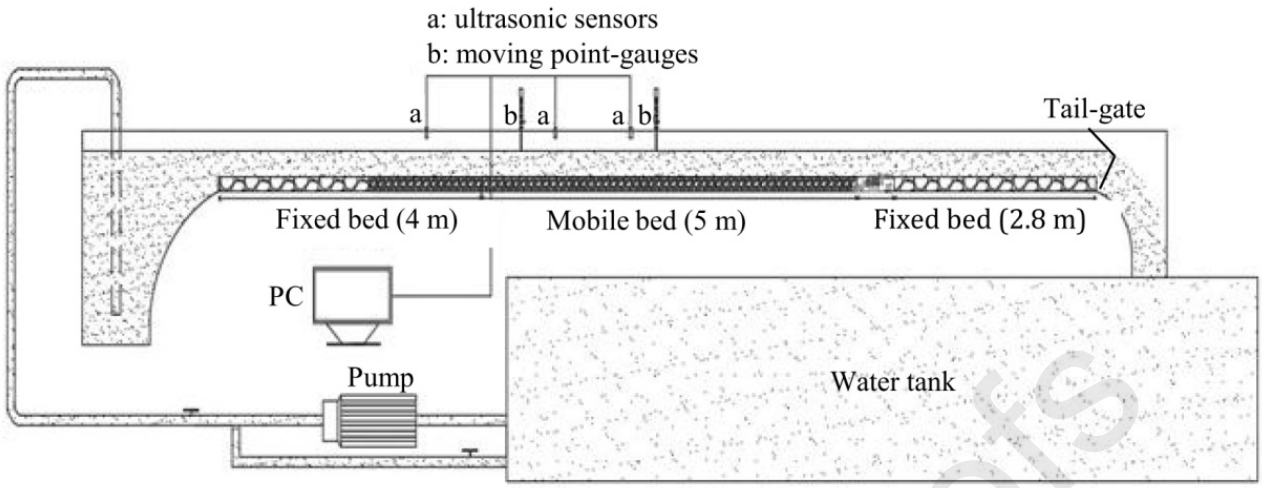


Fig.1. Experimental flume set-up (not to scale)

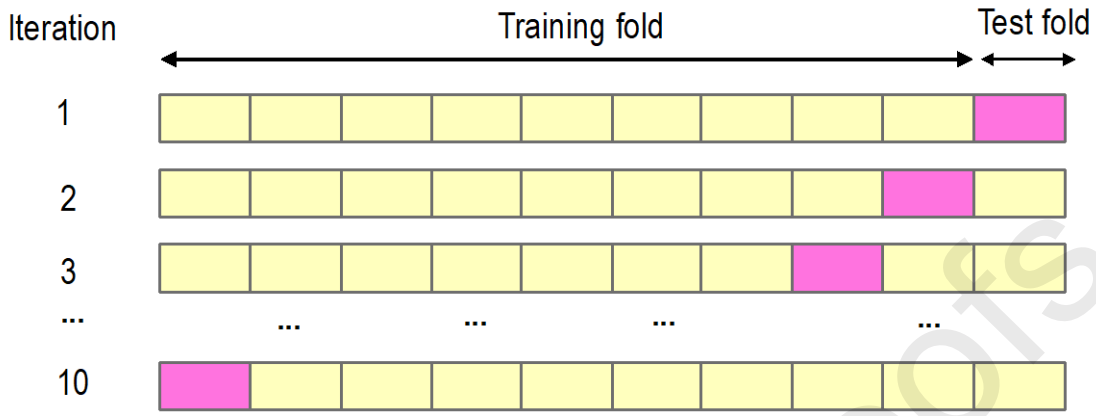


Fig.2. 10- fold cross validation technique

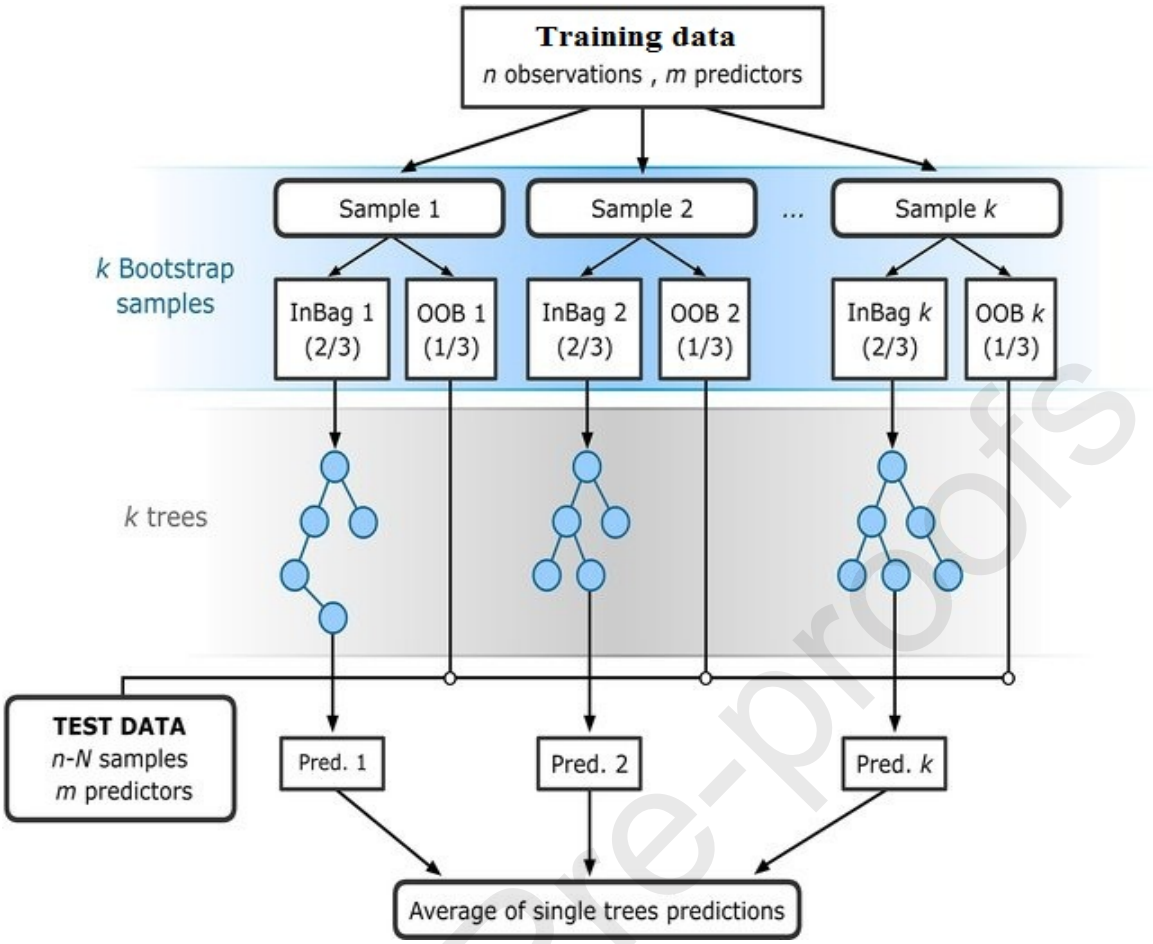


Fig. 3. Random forest algorithm structure (Rodriguez-Galiano et al. 2016)

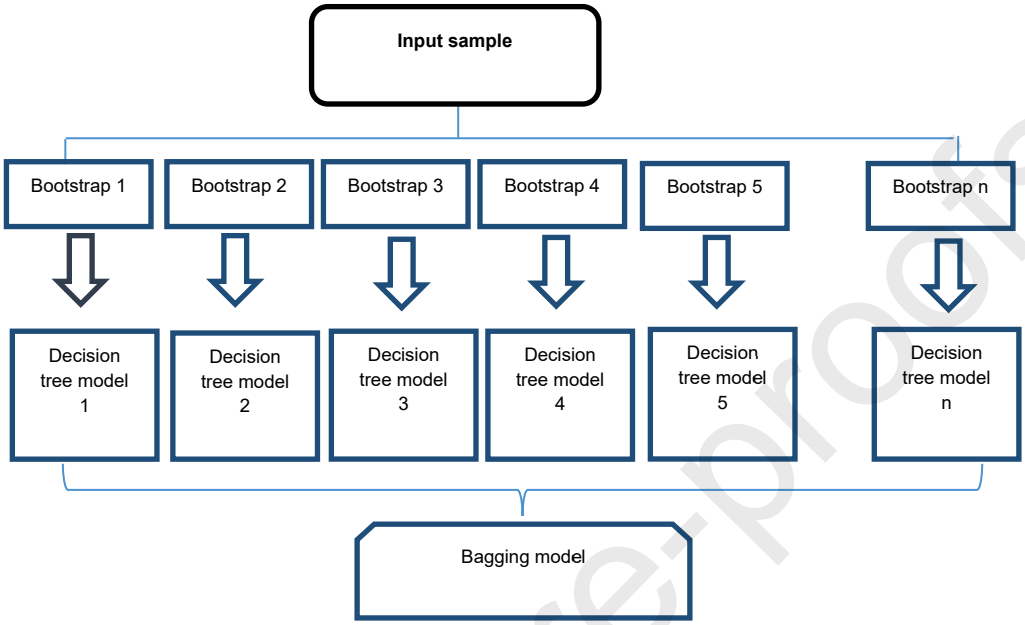


Fig. 4. Bagging algorithm structure (Khosravi et al. 2018b)

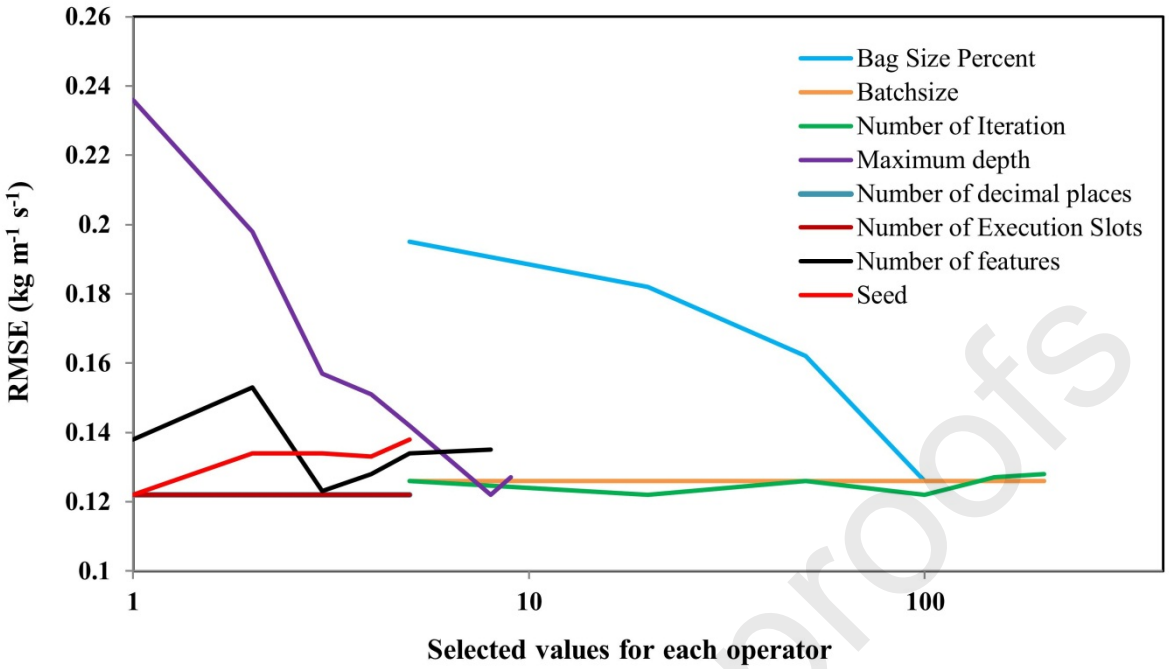


Fig. 5. Sensitivity analysis and identification of optimum operator values for the RF model

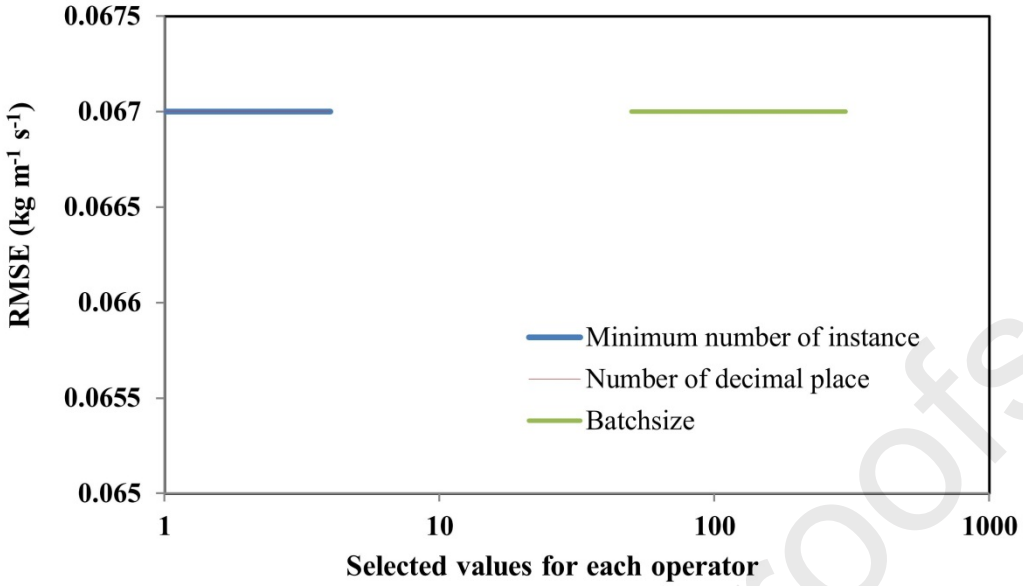


Fig. 6. Sensitivity analysis and identification of optimum operator values for the M5P model

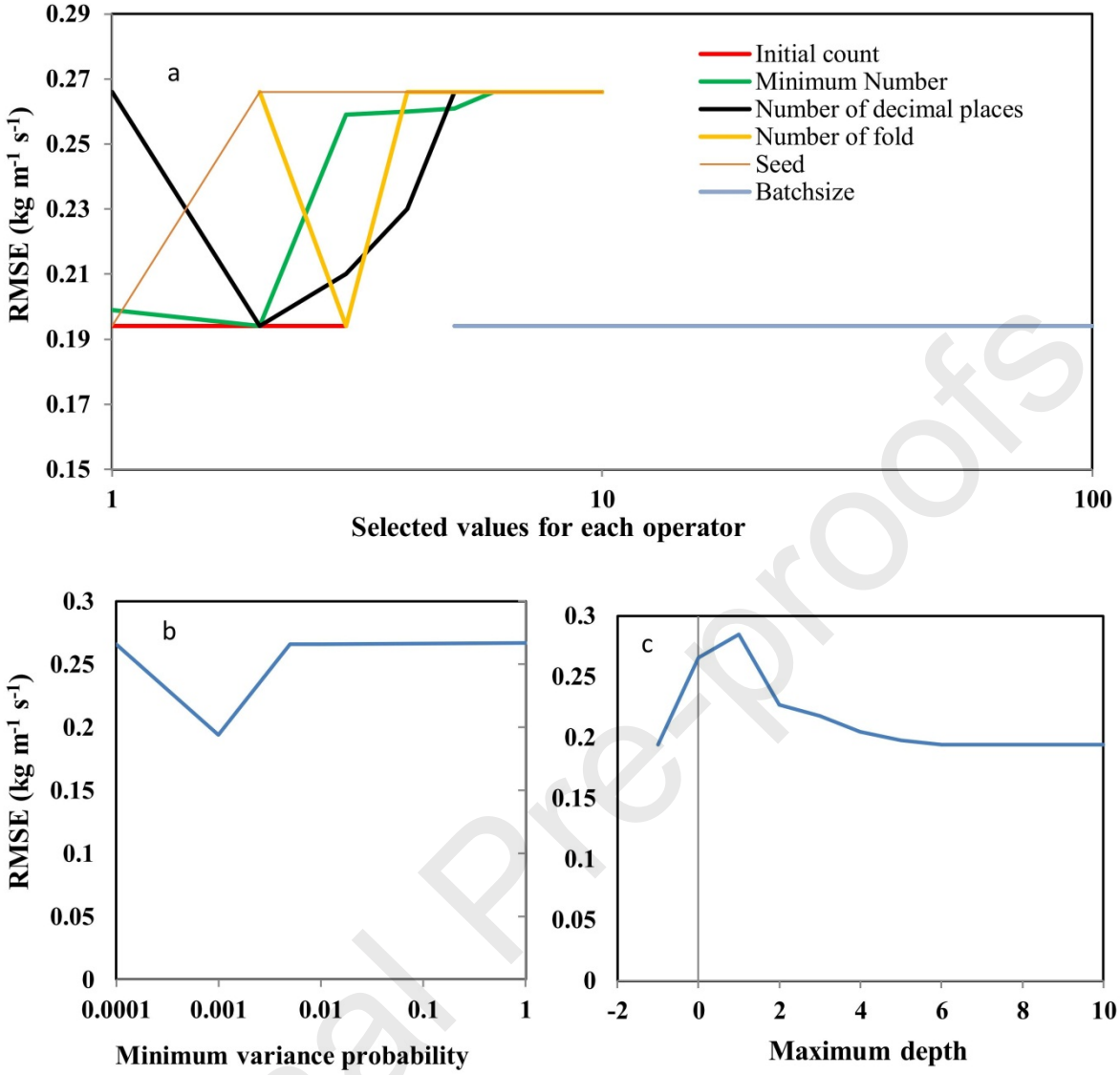


Fig. 7. Sensitivity analysis and identification of optimum operator values for the REPT model: (a) operators, (b) minimum variance probability, and (c) maximum depth

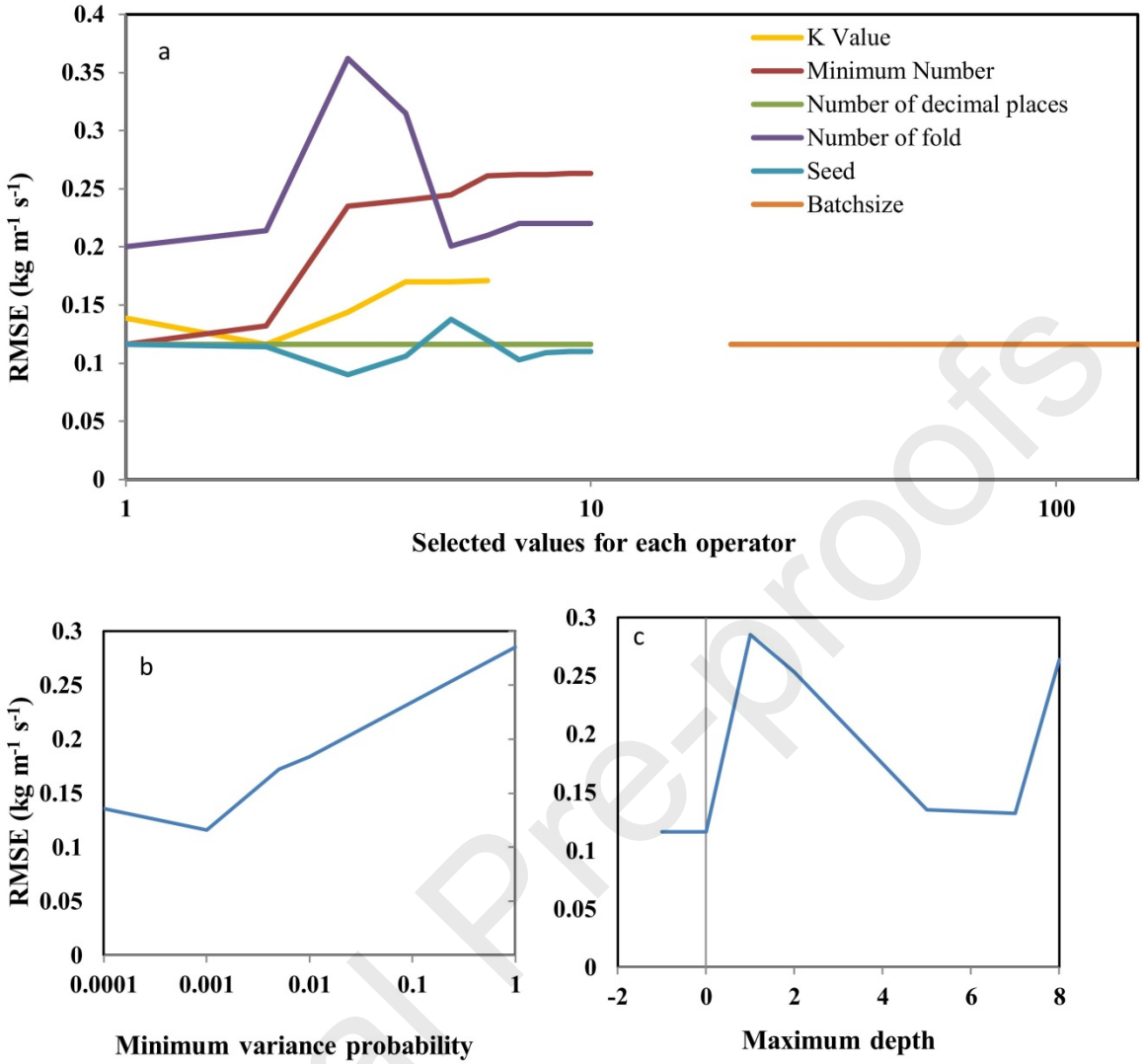


Fig. 8. Sensitivity analysis and identification of optimum operator values for the RT model: (a) operators, (b) minimum variance probability, and (c) maximum depth.

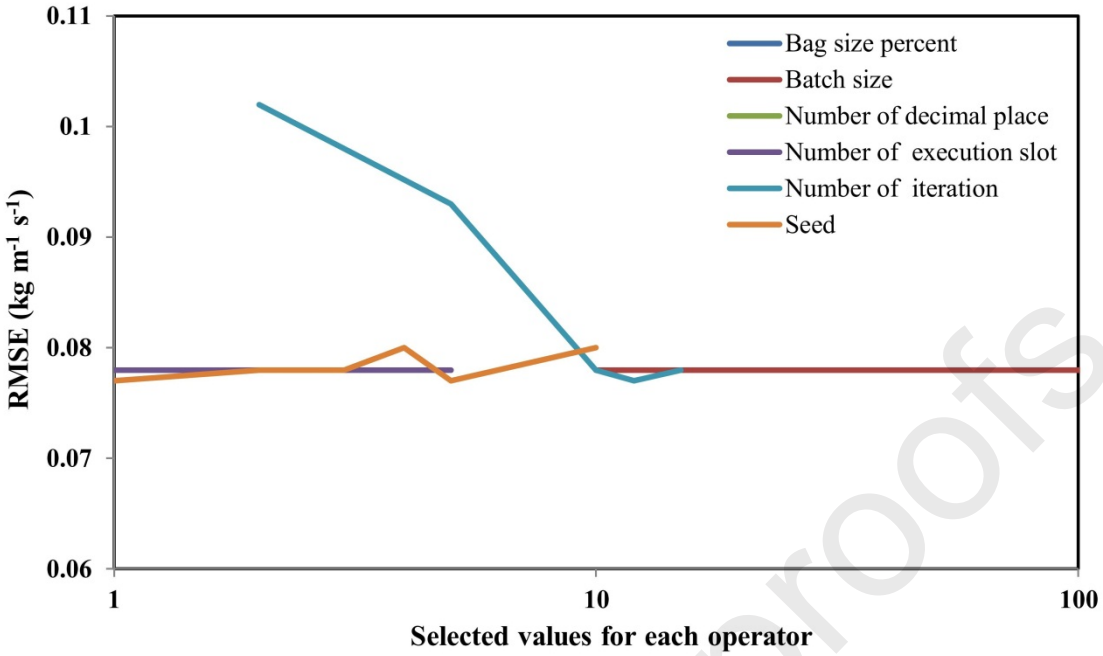


Fig. 9. Sensitivity analysis and identification of optimum operator values for the Bagging model

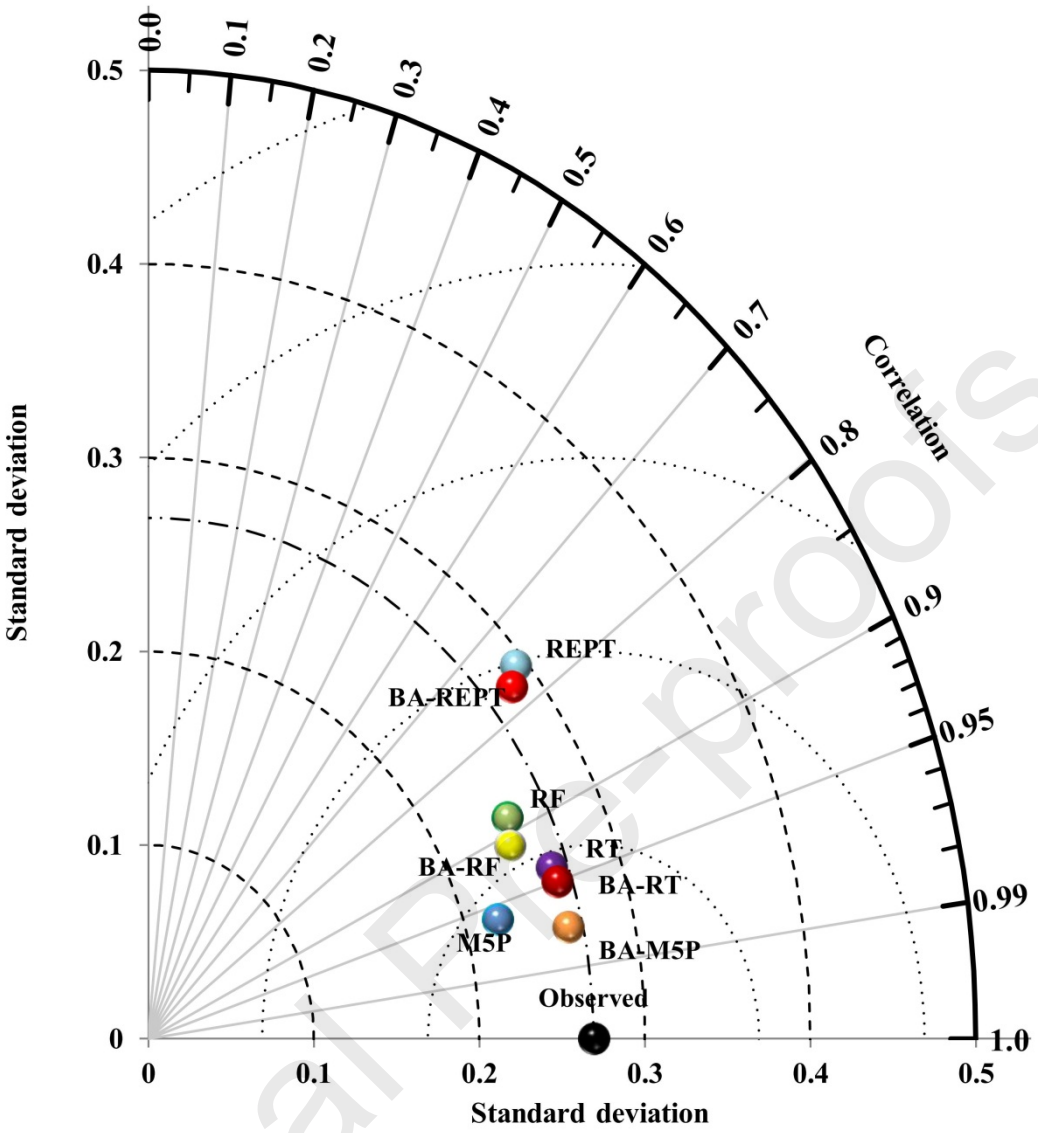
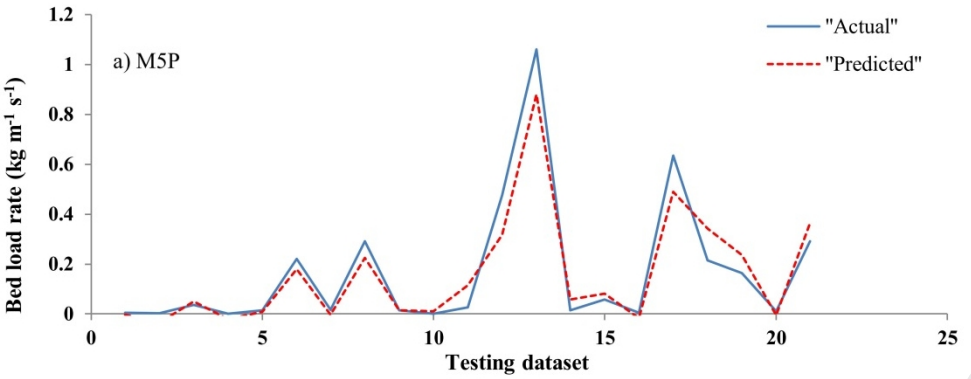
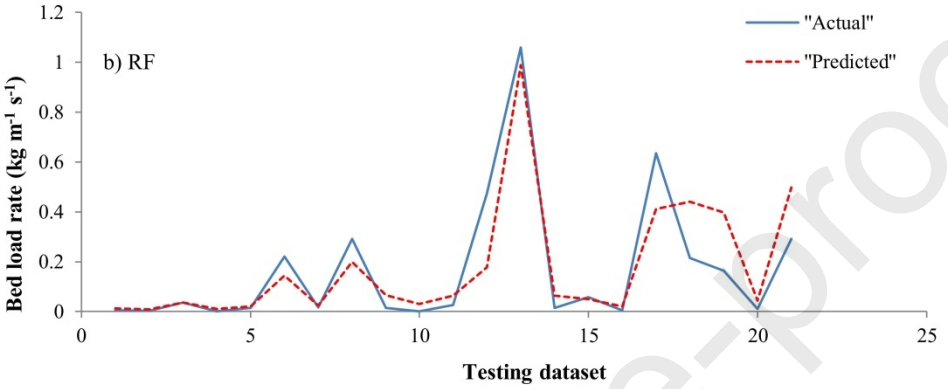
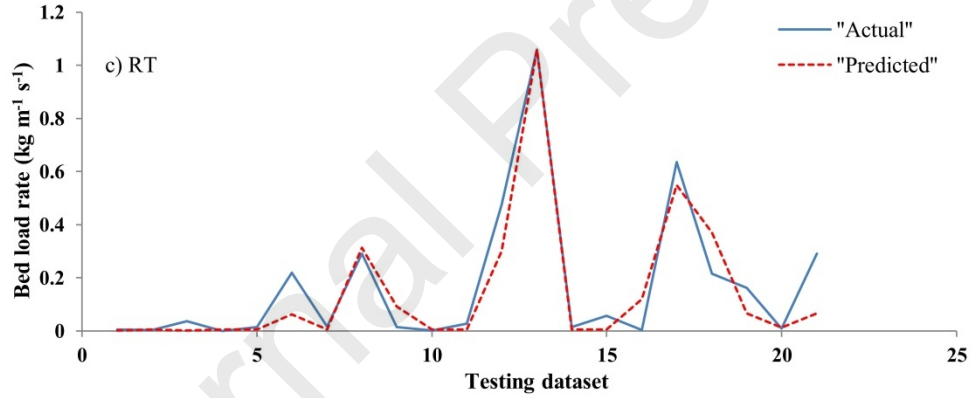
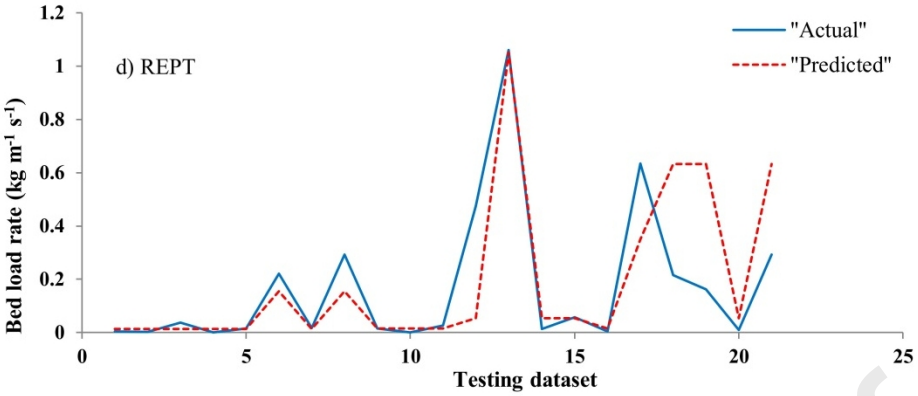


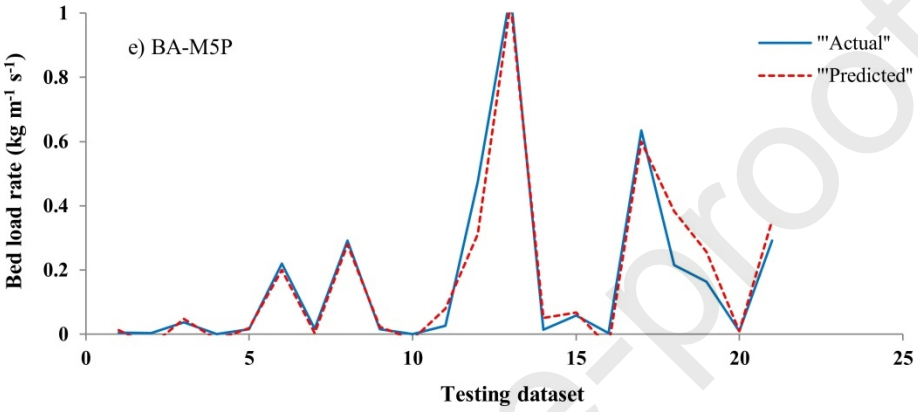
Fig. 10. Taylor plot of model performance

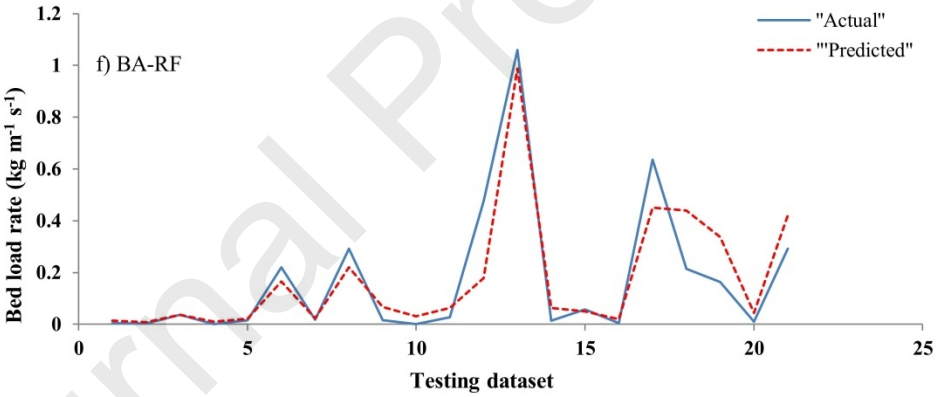












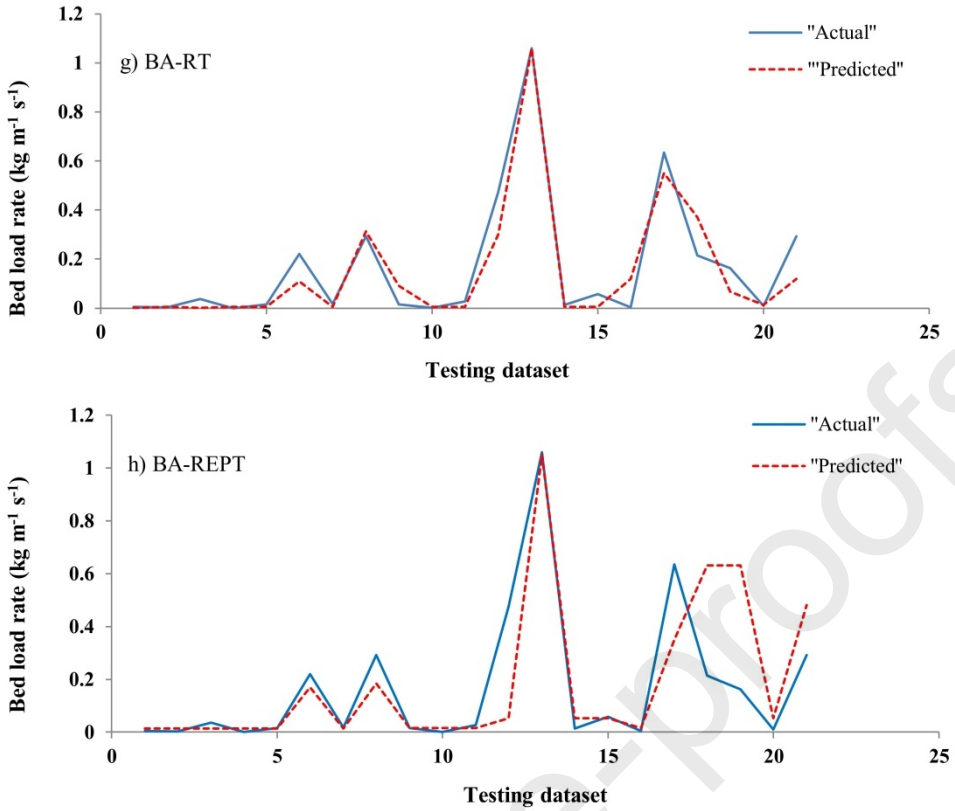
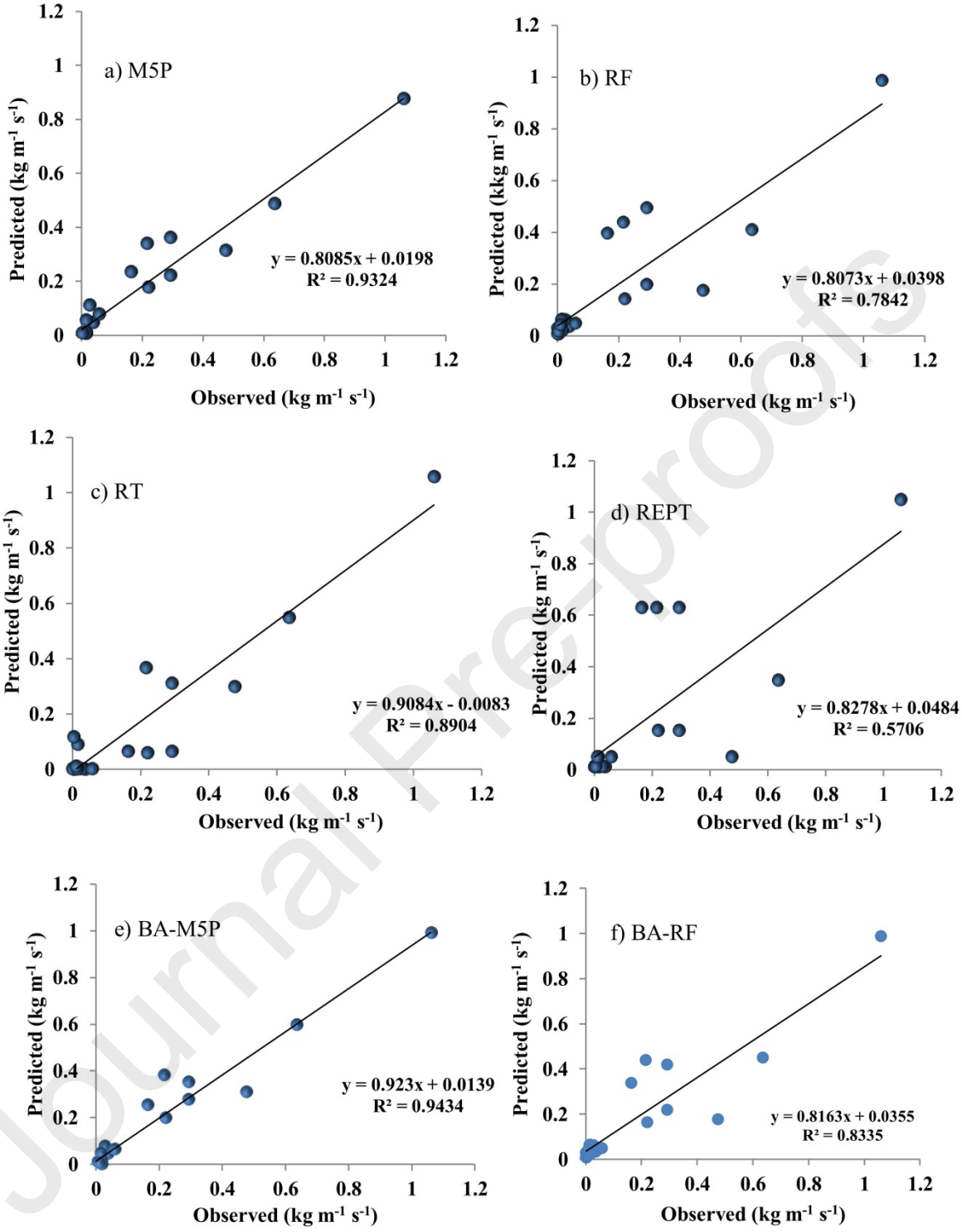


Fig. 11. Observed and predicted bed load transport rate in the testing phase: a) M5P, b) RF, c) RT, d) REPT, e) Bagging-M5P, f) BA-RF, BA-RT, and BA-REPT models.



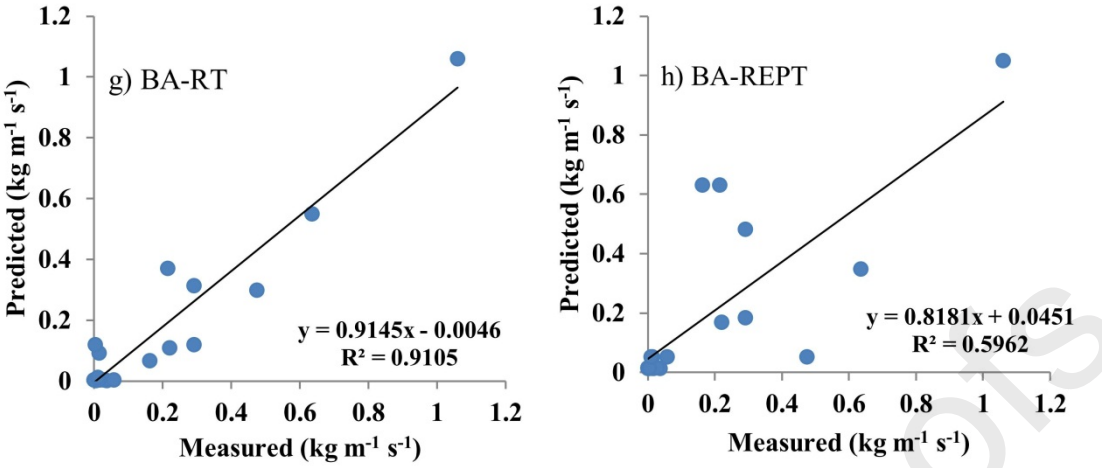


Fig. 12. Scatter plots of observed verus predicted bed load transport rate in the testing phase: a) M5P, b) RF, c) RT, d) REPT, e) BA-M5P, f) BA-RF, BA-RT, and BA-REPT models. The solid line denotes the line of best-fit.

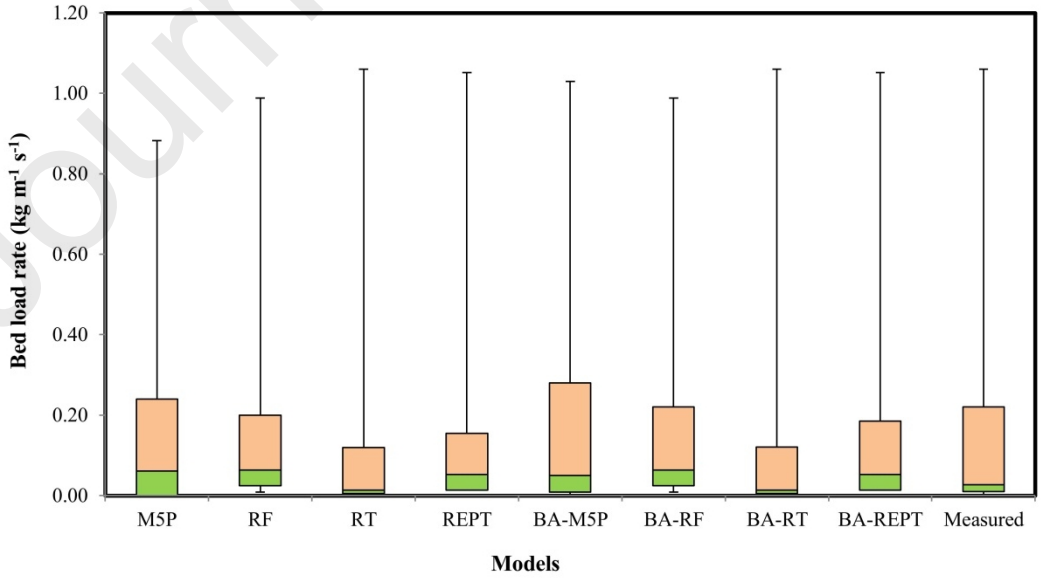


Fig. 13. Box plots of observed and predicted bedload transport rates.