Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-020-11335-5

RESEARCH ARTICLE

Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction

Sarita Gajbhiye Meshram[1,2] · Mir Jafar Sadegh Safari[3] · Khabat Khosravi[4] · Chandrashekhar Meshram[5]

Received: 5 July 2020 / Accepted: 20 October 2020

Abstract
Suspended sediment load is a substantial portion of the total sediment load in rivers and plays a vital role in determination of the service life of the downstream dam. To this end, estimation models are needed to compute suspended sediment load in rivers. The application of artificial intelligence (AI) techniques has become popular in water resources engineering for solving complex problems such as sediment transport modeling. In this study, novel integrative intelligence models coupled with iterative classifier optimizer (ICO) are proposed to compute suspended sediment load in Simga station in Seonath river basin, Chhattisgarh State, India. The proposed models are hybridization of the random forest (RF) and pace regression (PR) models with the iterative classifier optimizer (ICO) algorithm to develop ICO-RF and ICO-PR hybrid models. The recommended models are established using the discharge and sediment daily data spanning a 35-year period (1980–2015). The accuracy of the developed models is examined in terms of error; by root mean square error (*RMSE*) and mean absolute error (*MAE*); and based on a correlation index of determination coefficient (R^2). The proposed novel hybrid models of ICO-RF and ICO-PR have been found to be more precise than their stand-alone counterparts of RF and PR. Overall, ICO-RF models delivered better accuracy than their alternatives. The results of this analysis tend to claim the appropriateness of the implemented methodology for precise modeling of the suspended sediment load in rivers.

Keywords Hybrid technique · Iterative classifier optimizer · Pace regression · Random forest · River · Suspended sediment load

Introduction

The hydrological modeling of sediment, river stream and rainfall–overflow connection are significant to offer a design insight for the water resources management projects in practice (Firat and Gungor 2009). Sediment transport modeling is required for issues in the outline of transport of sediment in channels, ponds and bays, stable stations and dams, repositories of dams, protection of fish, effect of watershed administration, and ecological effect valuation (Cigizoglu 2004). In the field of computational hydrology, sediment and water quality modeling is a challenging task (Kisi et al. 2009). Sediment load has been estimated using traditionally method such as experimental relations, numerical reproductions, materially grounded models, remote sensing (RS) and geographic information systems (GIS) practices (Gajbhiye et al. 2015).

Responsible Editor: Marcus Schulz

✉ Sarita Gajbhiye Meshram
saritagmeshram@tdtu.edu.vn

Mir Jafar Sadegh Safari
jafar.safari@yasar.edu.tr

Khabat Khosravi
Khabat.Khosravi@gmail.com

Chandrashekhar Meshram
cs_meshram@rediffmail.com

[1] Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[2] Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[3] Department of Civil Engineering, Yaşar University, Izmir, Turkey

[4] Department of Watershed Management Engineering, Sari Agricultural Science and Natural Resources University, Sari, Iran

[5] Department of Post Graduate Studies and Research in Mathematics, Jayawanti Haksar Government Post Graduation College, College of Chhindwara University, Chhindwara, Betul, India

Published online: 30 October 2020

Reservoir sedimentation is the main concern, since the reservoir capacity is reduced to a large amount each year (Iraji et al. 2020). Surveying by traditional method for reservoir evaluation is a time-and money-consuming task. Due to various problems like population growth, agricultural activities, deforestation and poor soil conservation practices, the sedimentation has become a major problem in Indian Reservoirs. Global sediment production is estimated at around 15 × 10^{16} Ton/year, according to an estimate. The Indian subcontinent river carries around 6 billion tonnes of sediment per year. The soil erosion issue predominates over around 53% of India's total land area (Narayana and Ram Babu 1983). Currently about 40,000 major reservoirs are used worldwide for water storage, power generation, flood control etc. About 0.5 and 1% of the total storage capacity of these existing reservoirs is lost each year as a result of sedimentation, and 300–400 new dams are needed to be installed per year only to sustain the current total storage (White 2001).

The growing population and per capita consumption means that demand for water storage in reservoirs is rising, despite the increasing usage of alternative water sources and more productive water usage. Morris et al. (2008) predict that sedimentation would have depleted more than 30% of the world's reservoir capacity by the mid-twenty-first century.

India has invested heavily in developing these vital infrastructure needs that helped to improve the use and management of India's limited water resources. India ranks third worldwide in terms of the number of large dams, behind China and the USA with 5254 large dams built, and about 447 large dams under construction. As of March 2017, the live storage capacity of completed large dams amounted to approximately 283 billion cubic meters (BCM), i.e., 37% of the total useable surface water resources 690 BCM in India.

Systematic strategies and policies are required to reduce the adverse effects of sedimentation and extend reservoir existence. In designing sound sediment management strategies and policies, the ability to estimate the rate of watershed surface erosion, sediment transport, scouring and deposition in a river system, and sediment deposition and distribution within a reservoir is important. An addition to the planning and formulation of policies is the effective use of latest available technology like remote sensing (SRS), geographic information system (GIS), and soft computing techniques to measure the reservoir sedimentation.

The hydrologic conditions change spatio-temporally, and the challenges emerging in resolution of their special possessions have stimulated the engagement of black box models in the deferred sediment valuations. Black box models come in two types, that is, linear and non-linear. Artificial intelligence (AI) methods are normally used in the forming of non-linear system performance. The artificial intelligence techniques have attracted interest as modeling tools that have been applied to derive historical data to forecast future knowledge about a specific parameter over the last several decades. The artificial intelligence techniques have been adopted in many studies in the sense of hydrological problems such as rainfall-runoff modeling (Asadi et al. 2013; Tayebiyan et al. 2016; Juan et al. 2017; Tao et al. 2018; Mirabbasi et al. 2019; Safari et al. 2020), streamflow estimation (Besaw et al. 2010; Mehr et al. 2015; Fathian et al. 2019; Meshram et al. 2019b), reservoir inflow forecasting (Coulibaly et al. 2000; Sattari et al. 2012), water quality modeling (Khalil and Ouarda Taha 2011; Bui et al. 2020), prediction of evapotranspiration (Huo et al. 2012; Xiong et al. 2016; Khosravi et al. 2019), and sediment transport modeling (Yadav et al. 2017, 2018; Meshram et al. 2019a, 2020; Kargar et al. 2019; Safari et al. 2019; Safari 2020; Khosravi et al. 2020).

Due to the non-linear behavior of the suspended sediment problem and stochastic nature of the sediment particle movement in the flow, conventional computational methods may fail for accurate suspended sediment load prediction. To this end, AI approaches have been commonly implemented for sediment transport modeling in rivers (Nourani et al. 2016; Kisi and Yaseen 2019). Applied AI techniques for suspended sediment load prediction can be classified as stand-alone and hybrid algorithms. As examples of application of stand-alone algorithms, Tayfur (2002), Alp and Cigizoglu (2007), and Mustafa et al. (2012) investigated the efficiency of artificial neural networks (ANN) for suspended load prediction. Satisfactory performances of genetic algorithm (GA), neuro-fuzzy (NF), neural differential evolution (NDE), least square support vector regression (LSSVR), support vector machine (SVM), multivariate adaptive regression spline (MARS), and classification and regression tree (CART) as stand-alone models for suspended sediment load prediction were reported by Altunkaynak (2009), Rajaee et al. (2009), Kisi (2010), Kumar et al. (2016), Nourani et al. (2016), Yilmaz et al. (2018), and Choubin et al. (2018), respectively. Hybrid models may be implemented for suspended sediment transport modeling to improve the computational performance of the stand-alone models. For instance, Shiri and Kisi et al. (2012), Ramezani et al. (2015), and Zounemat-Kermani (2016) implemented wavelet-gene expression programming (W-GEP), ANN-social-based algorithm (ANN-SBA), and ANN-particle swarm optimization (ANN-PSO), respectively, for river suspended sediment modeling. Chen and Chau (2016) and Meshram et al. (2018) applied feed-forward ANN-based hybrid models of double feed forward neural network (HDFNN) and feed-forward neuron network particle swarm optimization gravitational search algorithm (FNN-PSOGSA) for the same purpose. Recently, alternative novel approaches of bagging-M5P, W-ANN, ANFIS-bat algorithm (ANFIS-BA), W-M5, and evolutionary fuzzy (EF) were suggested for suspended sediment load prediction by Khosravi et al. (2018), Sharghi et al. (2019), Ehteram et al. (2019), Nourani et al. (2019), and Kisi and Yaseen (2019), respectively. Despite relevant literature review showing that the random forest (RF) and pace regression (PR) models were rarely used for the modeling of suspended sediment load, their hybridized version integrated with an optimization algorithm is very rare in the literature.

Summary and a basic description of the RF algorithm can be found at Hastie et al. (2009), Verikas et al. (2011), Biau and Scornet (2016), Shirzad and Safari et al. (2019), and Safari et al. (2020). Regression with RF can be applied for forecasting purposes of the time series. Representative applications can be found with varying success in earth science studies including engineering (Herrera et al. 2010; Dudek 2014) and environmental and geophysical sciences (Chen et al. 2011; Naing and Htike 2015), with varying performances. Small datasets are often used in these applications; therefore, the results cannot be generalized. It can develop various advanced models that are focused on regression. For example, Wang (2000) proposed pace approach, on the basis of a technique similar to an empirical Bayes method. Pace regression (PR) is a linear regression approach that its outperformance on alternative linear methods was demonstrated, especially for problems having higher effective variables (Wang and Witten 1999). PR involves a form of collection of features; thus, not all features are included in the models result.

For the best author's information, there was no recorded work for the RF and PR model integrated with ICO for suspended sediment forecasting. The goal of the current study is to integrate the stand-alone RF and PR model with ICO to create robust smart models for forecasting suspended sediment load. For the purpose of validating the predictive accuracy of ICO-RF and ICO-PR models, the recorded data of Chhattisgarh State in India for the period of 1980 to 2015 was tested against the stand-alone RF and PR model for predicting daily suspended sediment load outcomes.

Materials and methods

Study area and modeling data

The Seonath river basin in Chhattisgarh State (India) is River Mahanadi's longest tributary sub-basin, covering 25% of the Mahanadi region area. The river crosses a length of 380 km. The basin is situated between latitude 20° 16′ N to 22° 41′ N and longitudes 80° 25′ E to 82° 35′ E. The normal elevation of basin is 329 m above MSL with minimum and maximum elevation of 204 m and 1058 m, individually (Fig. 1).

Most of the tributaries of Seonath River get dried by mid-winter season, and both rural and urban areas are subjected to severe water crisis during the summer season due to erratic and skewed nature of rainfall. The river basin experiences a sub-humid type of climate. The geographical factors such as distance from the sea and altitude have influenced the basin climate. The mean annual rainfall in the basin varies from 1005 to 1255 mm. The major part of rainfall occurs only within three monsoon months (July–September). It experiences higher humidity levels during monsoon season. The summer season prevails from April to middle of June. The climatic condition during summer is hot and gusts of dry wind blow; the temperature varies from 40 to 45.5 °C. The mean daily maximum temperature varies from 42 to 45.5 °C for the hottest month of May. During winter the temperature varies between 10 and 25 °C.

The main soil types found in the basin are sandy clay and silt loam. Agriculture is the main occupation of people in this sub-basin. There are two cropping seasons, namely monsoon (kharif) season from mid-June to October and post-monsoon (rabi) season from November to middle of April. Rice is the major crop of monsoon season covering 94% of the cultivated basin area. During rabi season, wheat, summer paddy, pulses, and oilseed are grown.

Daily data used in this study includes discharge (m^3/s) and suspended sediment load (ton/day) obtained from the Simga station for the period of 1980 to 2015. Among the 35 years of data, 75% discharge and suspended sediment load were utilized for the model development/calibration, and the rest 25% were employed to test/validate the model performance. Figure 2 displays the time series of the entire data that was implemented for Simga station. Table 1 lists the statistical parameters for the results.

Random forest

Breiman (2001) initially developed the random forest (RF) model based on a variation of the decision tree classifiers (Breiman and Cutler 2004). RF is a set of methods of learning which can be used for regression and classification. The basic principle of the methodology of the random forest is the construction of a forest of random trees that are generated through randomizing the spilt at every decision tree node. RF integrates the robustness of several individual trees to create a more accurate model applying an ensemble approach (Jayech and Mahjoub 2011; Goeschel 2016).

A number of studies explored RF's application in engineering applications and demonstrated its viability in prediction processes (Rudžianskaitė-Kvaraciejienė et al. 2015; Yaseen et al. 2019a, 2019b; Shirzad and Safari 2019). Under the bootstrapping method, data is selected randomly and independently during the training phase to develop the RF model, and data not involved in the selection process is referred to as "out-of-bag" (Catani et al. 2013). Owing to the large number of trees, over-fitting does not occur in the RF algorithm and the choice of the correct type of random variables leads to precise classification. RF contain several parameters, such as number of trees, minimum gain, and maximum tree depth, that need to be optimized.

In order to calculate the output of $\hat{f}_{rf}^{B}(x)$ in input x, RF model is fitted for each bootstrap samples of $b = 1, 2, 3, \ldots,$

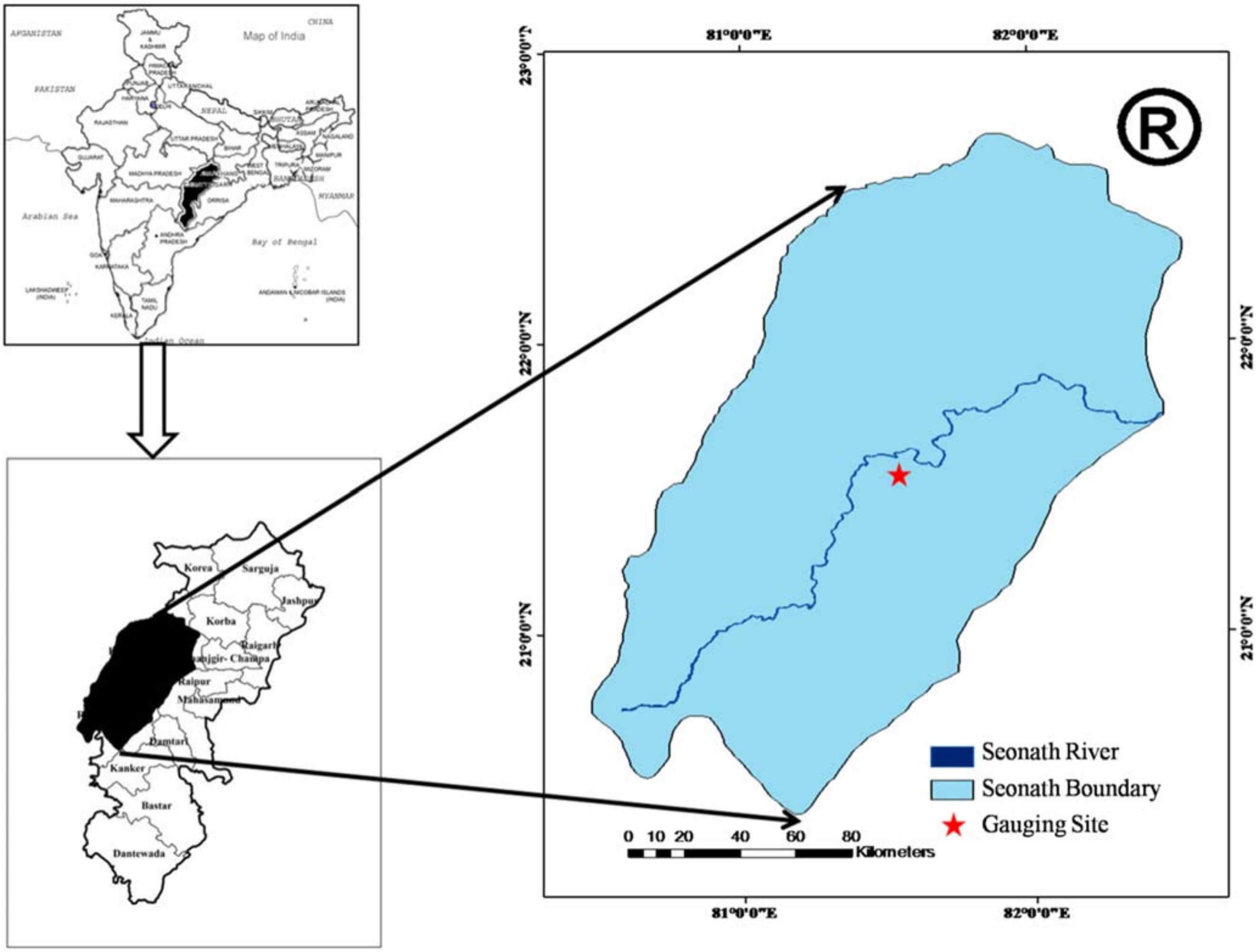

Fig. 1 Index map of Seonath river basin (study area)

B (Hastie et al. 2009). Output of the RF model through construction of random forest tree T_b on the bootstrapped data is computed by:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{1}$$

RF is a hyper-parameter algorithm and this is the main drawback of RF model.

Pace regression

Pace regression (PR) approach was first introduced by Wang (2000) for linear-fitting problems. The fundamental principles of PR are based upon Robbins' (1964) empirical Bayes methodology. An asymptotic normality property for maximum likelihood estimation (MLE) is applied to convert the initial variables to dummy variables. A nonparametric mixture approximation is developed for the measured quantities of these dummy variables, and in the end, an empirical Bayes approach is applied to minimize the Kullback–Leibler distance.

Considering the methodological approach of Bayes, given independent samples of $x_{1,}.........x_k$ from $\mathcal{F}(x_i\ ;\theta_i\ )$ distributions, where the values of θ_i may differ completely with respect to one another, it is recognized that the MLE provided from the $\mathcal{F}(x;\theta)$ joint distribution is a vector, with each entry being a univariate MLE; for instance, if $\mathcal{F}(x_i\ ;\theta_i)$, is the normal distribution with the mean θ_i, then $\widehat{\theta}$ =*x*. The MLE calculator is lower than the empirical one:

$$\widetilde{\theta}_{\iota}^{\mathcal{EB}} = \frac{\int \theta \mathfrak{f}(x_i;\theta) d\mathcal{G}_{k}(\theta)}{\int \mathfrak{f}(x_i;\theta) d\mathcal{G}_{k}(\theta)} \tag{2}$$

where $\mathfrak{f}(x_i;\theta_i)$, is the probability density function in proportion to $\mathcal{F}(x_i\ ;\theta_i)$ that is inferior in which predicted squared error $\mathcal{E}_{\mathfrak{f}(x)}\left\|\hat{\theta}-\theta\right\|^2$ is not minimized with respect to the estimator $\widetilde{\theta}(x)$, where $\theta_{1,}..........\theta_{\|}$ are independent and distributed equivalently from $\mathcal{G}(\theta)$, where $\mathcal{G}$ is the distribution of the function $\mathfrak{f}_{\mathcal{G}}(x) = \int \mathfrak{f}(x;\theta)\, d\mathcal{G}$, and $\mathcal{G}_{\|}$ is a consistent calculator of $\mathcal{G}$ provided the mixture sample of *x*.

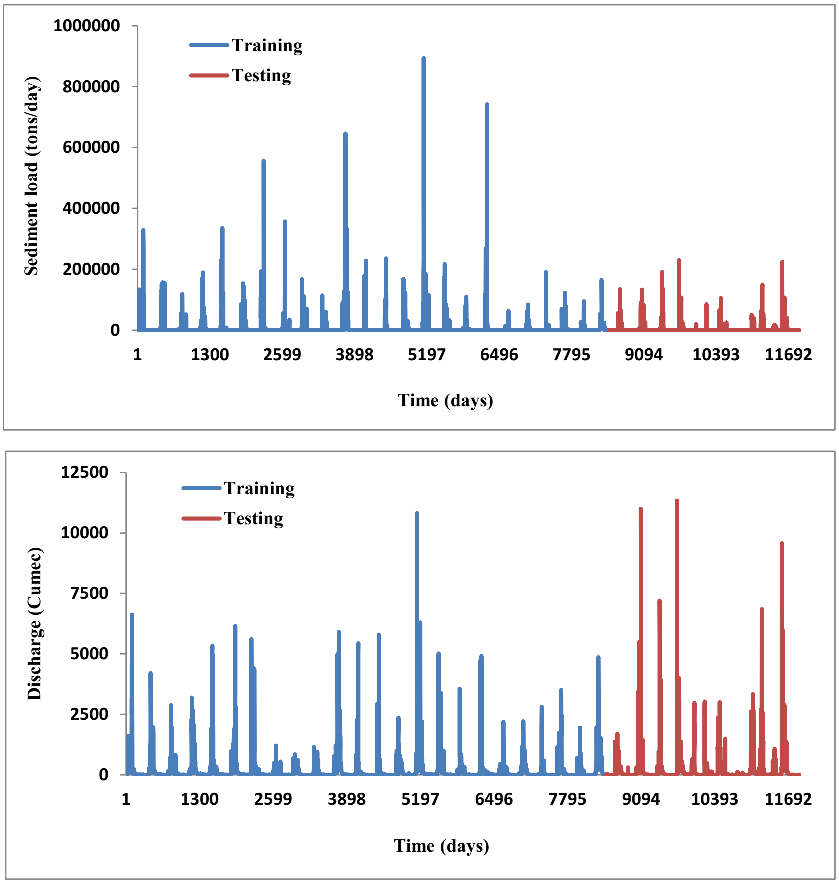

Fig. 2 Time series of observed data (discharge and sediment) used for training and testing stages

Iterative classifier optimizer

Iterative classifier optimizer (ICO) uses cross-validation and optimizes the number of iteration for the given classifier; it is capable of handling missing, nominal, binary classes and attributes like numeric, nominal, binary, empty nominal (Omondi and Rajapakse 2010). Through the optimization procedure of ICO algorithm, after developing the model, comparing the observed and measured values, the model performance is examined and, then, the obtained information are introduced to the model for tuning the outputs. The main objective of the hybridization is to enhance the prediction accuracy of the stand-alone RF and PR algorithms. As stated previously, RF algorithm suffers from determination of the optimal hyper-parameter and in this study the RF and PR are integrated with ICO for improving the results and develop robust algorithms. It is already reported that each tree in a RF model can grow incorrectly and reduced the prediction accuracy of the model (Adnan et al. 2019). Number of trees grown and number of predictors sampled for splitting at each

Table 1 Statistics of the data

Parameters	X_{min}	X_{mean}	X_{max}	Standard deviation	Variation coefficient
Entire data					
Discharge (m^3/s)	0.230	251.7989	11,331.68	685.7905	272.3565
Suspended sediment load (ton/day)	0.081	6812.378	892,862.4	30,038.83	440.9449
Training					
Discharge (m^3/s)	0.014249	157.7602	10,821	509.0265	322.6583
Suspended sediment load (ton/day)	0	4917.394	892,862.4	27,337.24	555.9294
Testing					
Discharge (m^3/s)	0	191.5584	11,331.68	681.4086	355.7184
Suspended sediment load (ton/day)	0	3053.609	229,393.1	14,209.63	465.3388

node are two operators from these hyper-parameter which significantly affect the RF prediction power. To this end, ICO algorithm was implemented to determine the best subset of features in RF model to enhance the result. Figure 3 shows flowchart of RF integrated with ICO.

Performance criteria

Three statistical indices of root mean square error (*RMSE*), mean absolute error (*MAE*), and determination coefficient (R^2) were utilized for performance examination of stand-alone RF and PR, as well as the hybrid ICO-RF and ICO-PR models for modeling suspended sediment loads. *RMSE*, *MAE*, and R^2 can be expressed respectively by:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{n}} \tag{3}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|x_i-y_i| \tag{4}$$

$$R^2=\left(\frac{1}{n}\times\frac{\sum\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{(\sigma_x)(\sigma_y)}\right)^2 \tag{5}$$

where x_i and y_i are observed and estimated values, standard deviation of the measured and estimated data are respectively as σ_x and σ_y, and n is the number of data. It must be noted that the less value (near to zero) for *RMSE* and *MAE* and greater value (near to the unity) for R^2 imply perfect agreement between measured and calculated parameters.

Result and discussion

Using the daily discharge and suspended sediment load data series for Simga station located in the Seonath Basin, India, the stand-alone RF and PR models vs. hybrid ICO-RF and ICO-PR models have been developed and evaluated for sediment load prediction. The entire data was split into training/calibration (75%) and testing/validation (25%) sub data sets and MATLAB software was implemented for model construction.

Best input combination

To select the most important driving variable between input variables, the Pearson correlation coefficient (PCC) methods were applied (Chiang and Tsai 2011; Kisi et al. 2012; Khosravi et al. 2018). In order to investigate the correlation of different input parameters to the model output, the Pearson correlation coefficients for one-three days ahead discharge (*Q*) and suspended sediment load (*S*) with output parameter were calculated. The Pearson correlation coefficient (PCC) values

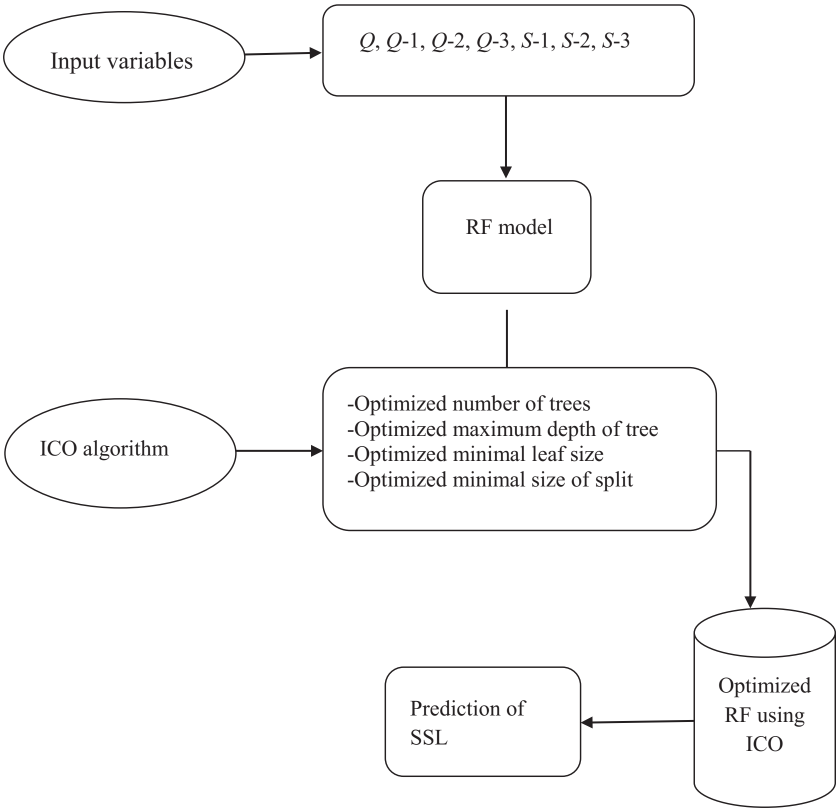
Input variables
Q, *Q*-1, *Q*-2, *Q*-3, *S*-1, *S*-2, *S*-3
RF model
ICO algorithm
-Optimized number of trees
-Optimized maximum depth of tree
-Optimized minimal leaf size
-Optimized minimal size of split
Optimized RF using ICO
Prediction of SSL

Fig. 3 Flowchart of optimized RF using ICO (ICO-RF)

in Table 2 indicate that the discharge provides the highest effect on suspended sediment load (PCC = 0.75), followed by *S*-1 (PCC = 0.63), *Q*-1 (PCC = 0.56), *Q*-2, S-2 (PCC = 0.34), and *Q*-3, S-3 (PCC = 0.23). Ten separate combinations were built and investigated on the basis of certain PCC values as shown in Table 3.

Inappropriate selection of inputs for intelligent models can decrease model's accuracy and increase modeling complexity. Likewise, a crucial stage in the process of building such models is the selection of the correct subset of applicable input variables. All the developed models in the present study (e.g., RF, PR, ICO-RF, and ICO-PR) use separate datasets for each of the different sets of input parameters. The efficiency of the models was evaluated based on the *RMSE* as shown in Table 4 using different subsets of input parameters. It is seen in Table 4 that all studied models of RF, PR, ICO-RF, and ICO-PR provide better results for scenarios No. 5 and No. 6, where input combinations are constructed in terms of one-three ahead suspended sediment load data. Although discharge has higher PCC value with suspended sediment load, it gives poor results in the modeling. Among scenarios that discharge and suspended sediment load are incorporated into the model structure, scenario No. 9 provides better results; however, its performance is not as high as scenario No. 6. The best input combination for the RF, PR, ICO-RF, and ICO-PR versions is found as scenario No. 9.

Model performance and validation

Historical discharge and suspended sediment load data are vital factors in modeling of a river suspended sediment load. The seasonality of rainfall affects discharge and influences the suspended sediment load (Yunus and Nakagoshi 2004). In this study, two stand-alone intelligent models of PR and RF were employed to calculate suspended sediment load. In order to enhance the robustness of the stand-alone models, two novel hybrid algorithms were developed by combining the stand-alone models of RR and PE with ICO optimization technique. The performance of the four developed models was compared in terms of accurate suspended sediment load prediction.

After determination of the most effective combination of input parameters, all algorithms were trained utilizing the train dataset and, then, their performances were examined on a test data set. This analysis would demonstrate how the built model fits the train data set, as the models were created using a training data set (Khosravi et al. 2016; Chen et al. 2019). To this end, model's credibility must be examined on unseen data set at testing stage.

Table 3 Different combination of input parameters

No.	Input combinations
1	*Q*
2	*Q*, *Q*-1
3	*Q*,*Q*-1, *Q*-2
4	*Q*,*Q*-1, *Q*-2, *Q*-3
5	*S*-1
6	*S*-1, *S*-2
7	*S*-1, *S*-2, *S*-3
8	*Q*, *S*-1
9	*Q*, *S*-1, *Q*-1
10	*Q*,*Q*-1, *Q*-2, *Q*-3, *S*-1, *S*-2, *S*-3

Table 5 illustrates an evaluation of the performance of the four recommended models for suspended sediment load prediction. For the sake of fair comparison of the models, the statistical parameters must be applied during both training and testing stages. However, the performance bench marks are more relevant when determining the best model during the test phase, because the performance of the models during the test phase demonstrates their ability to replicate any new data not entered in the models during the training period (Meshram et al. 2019). The R^2 values indicate that during the testing process, the ICO-RF model generates the best performance (0.81) followed by the ICO-PR (0.80), PR (0.73), and RF (0.64). R^2 is optimized for variations between mean and variance of measured and expected quantities; it is prone to outliers and must not be utilized exclusively for examination of developed models (Legates and McCabe 1999; Shiri and Kisi 2012). Therefore, alternative error measurement indices were used for model performance evaluation. The ICO-RF was superior to the other types, based on *RMSE* and *MAE*. The ICO-RF and RF models proved the greatest and least predictive capability, taking into account all the evaluation metrics together. The efficiency of ICO-RF and ICO-PR is found better during the training and testing process than the respective RF and PR versions. For example, during the testing stage, the values of the *MAE* and *RMSE* indices, i.e., 2600 and 8675 (RF model), and 3288 and 7634 (PR model), in the ICO decrease to 2252 and 6329, and 2880 and 6436, respectively. Evaluation of the performances of the developed hybrid models in contrast to the stand-alone RF and PR models shows that the hybrid models being proposed are more

Table 2 Pearson correlation coefficient for different input parameters

Input	*Q*	*Q*-1	*Q*-2	*Q*-3	*S*-1	*S*-2	*S*-3
Pearson correlation coefficient	0.75	0.56	0.34	0.23	0.63	0.34	0.23

Table 4 *RMSE* values for RF, PR, ICO-RF, and ICO-PR models performed in different scenarios

No.	Input combinations	RF	PR	ICO-RF	ICO-PR
1	*Q*	18,565	17,938	15,385	16,758
2	*Q*, *Q*-1	19,061	18,673	17,602	17,512
3	*Q*,*Q*-1, *Q*-2	19,134	18,847	16,535	17,363
4	*Q*,*Q*-1, *Q*-2, *Q*-3	19,274	19,084	16,837	17,728
5	*S*-1	10,337	10,278	10,312	10,229
6	***S*-1, *S*-2**	**10,301**	**10,105**	**10,036**	**10,038**
7	*S*-1, *S*-2, *S*-3	11,640	10,126	10,047	10,140
8	*Q*, *S*-1	16,163	16,163	15,721	15,635
9	*Q*, *S*-1, *Q*-1	15,036	14,832	12,059	12,580
10	*Q*,*Q*-1, *Q*-2, *Q*-3, *S*-1, *S*-2, *S*-3	15,245	15,146	14,699	15,497

Bold values showed minimum RMSE for RF, PR, ICO-RF and ICO-PR

reliable than stand-alone ones. In the proposed hybrid models, the less accurate results of RF and PR models are usually improved to excellent or reasonable performance by considering the R^2 predictor. Furthermore, the weak correlations of measured and computed suspended sediment load data in the RF and PR models are significantly enhanced in the ICO-RF and ICO-PR models. ICO-RF also had the highest results as compared to the other models. It is worthy to mention that among stand-alone models, PR gave better results than RF; however, between hybrid models, ICO-RF outperformed ICO-PR. It illustrates that an optimization approach greatly promotes the RF performance, where in this study the performance of stand-alone RF model has been improved by a factor of 27% in ICO-RF model with *RMSE* values of 8675 and 6329, respectively.

In Figs. 4 and 5, the scatter and comparative plots were drawn for graphical checking of the performance of proposed ICO-RF and ICO-PR hybrid models compared to the stand-alone RF and PR models during testing and training phases. Figures 4 and 5 reflect scatter plots between regular suspended sediments observed vs. computed during the training and test phases. The ICO-RF model predicts more accurately than the other models as shown in Fig. 5. The results generally show that hybrid algorithms' predictive power depends mostly on the optimization approach (i.e., ICO) and base algorithm (i.e., PR, RF). The implementation of iterative classifier optimizer (ICO) approach improved the stand-alone model's predictive ability. A significant underestimation for suspended sediment load data has been seen for stand-alone RF and PR models, while ICO-RF and ICO-PR hybrid models generate better results with a slight underestimation. An important feature of hybrid models of ICO-RF and ICO-PR is that they have the ability to capture extreme suspended load values as shown in time series plots in Fig. 5. Stand-alone models of RF and PR are failed in predicting extreme suspended sediment load data, indicating their poor performance for molding river suspended sediment prediction.

Table 5 Comparison of the best models in terms of R^2, *MAE*, and *RMSE*

Model	Training			Testing		
	MAE	*RMSE*	R^2	*MAE*	*RMSE*	R^2
RF	**3829**	**20,512**	**0.44**	2600	8675	0.64
PR	4794	21,204	0.40	3288	7634	0.73
ICO-RF	4021	20,800	0.42	**2252**	**6329**	**0.81**
ICO-PR	4685	21,316	0.40	2880	6436	0.80

MAE, RMSE bold value showed minimum and R2 bold value showed maximum value for RF, PR, ICO-RF and ICO-PR

As a result of statistical analysis given above, it can be concluded that in the ICO-RF model the daily suspended sediment of the current day can be modeled with fewer inputs using the suspended sediment of the one day and two days ahead data. ICO-RF is found superior to its alternatives, although ICO-PR can compete with ICO-RF model in terms of accurate prediction of suspended sediment load.

As the literature review shows, suspended yield prediction by soft computing techniques was superior compared to that using traditional method (Yadav et al. 2017). The performance of the sediment rating curve (SRC) model was below expectations as it produced the least accurate results for the peak sediment values, as well as overall model performance. It is also noticed that the multiple linear regression (MLR) model predicted negative sediment yield at low values, which is completely unrealistic as suspended sediment yield cannot be negative in nature. It was also observed that suspended yield prediction by ANN was superior compared to that using MLR (Yadav et al. 2017).

It is always challenging to model sediment yield using traditional mathematical models because they are incapable of handling the complex non-linearity and non-stationarity (Yadav et al. 2020). A comparative study of different traditional models for assessment of sediment yield (Modified Universal Soil Loss Equation and Sediment delivery ratio)

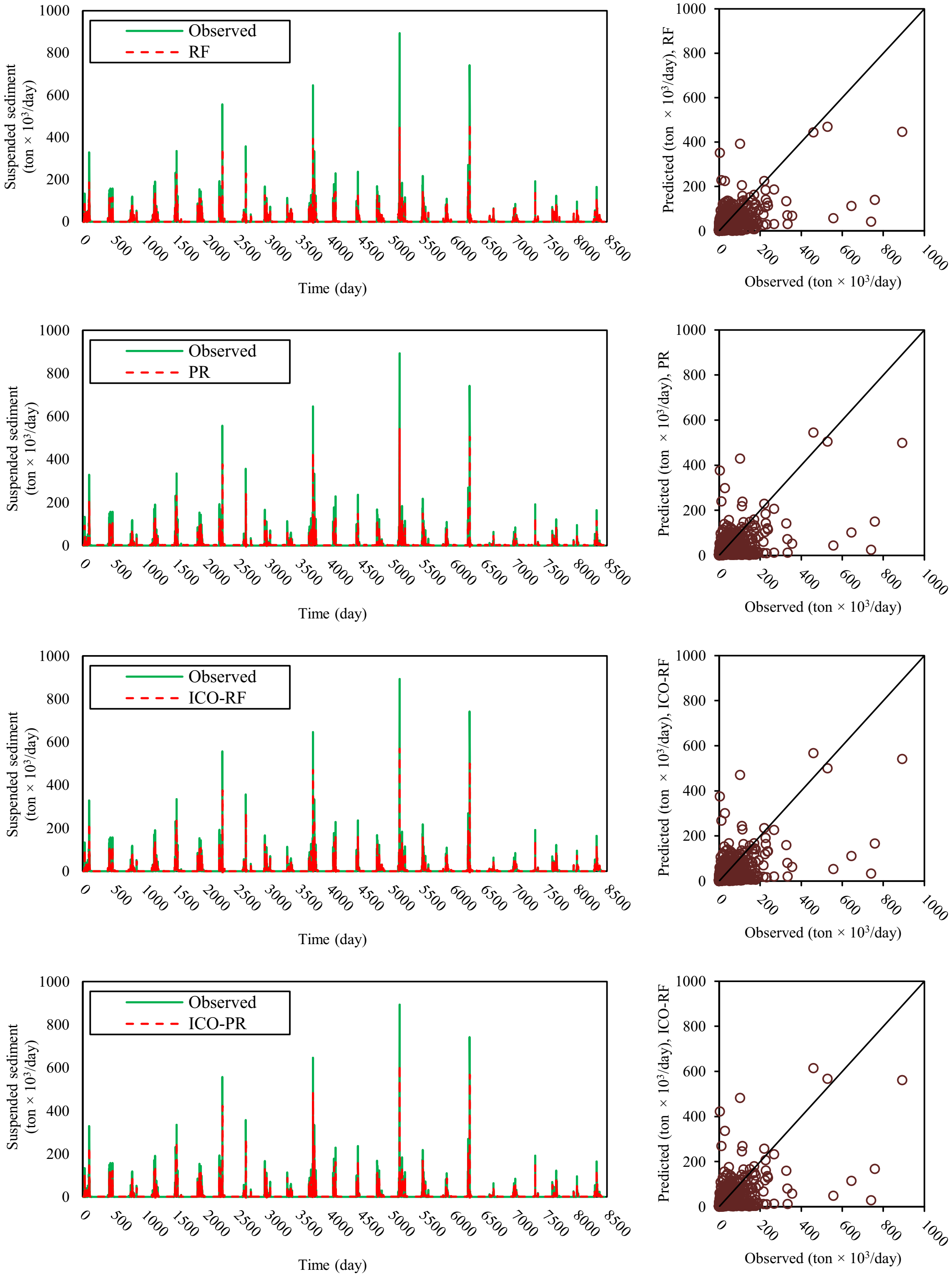

Fig. 4 Observed and predicted suspended sediment load for RF, PR, ICO-RF, and ICO-PR during training phase

was carried out in Pairi Watershed, Chhattisgarh, India. It is found that MUSLE model for sediment yield has been found to be most reliable as compared to RUSLE (Kumar et al. 2019). The soil loss estimated by the RUSLE method was quite close to the direct field measurement (Nigam et al. 2017).

Despite the AI-based models, promising implementation in the many fields of scientific research has been implemented and demonstrated, but there are still some notable challenges attributed to AI-based models. The main drawback of the ANN model is weak generalization potential, lack of strict design programs

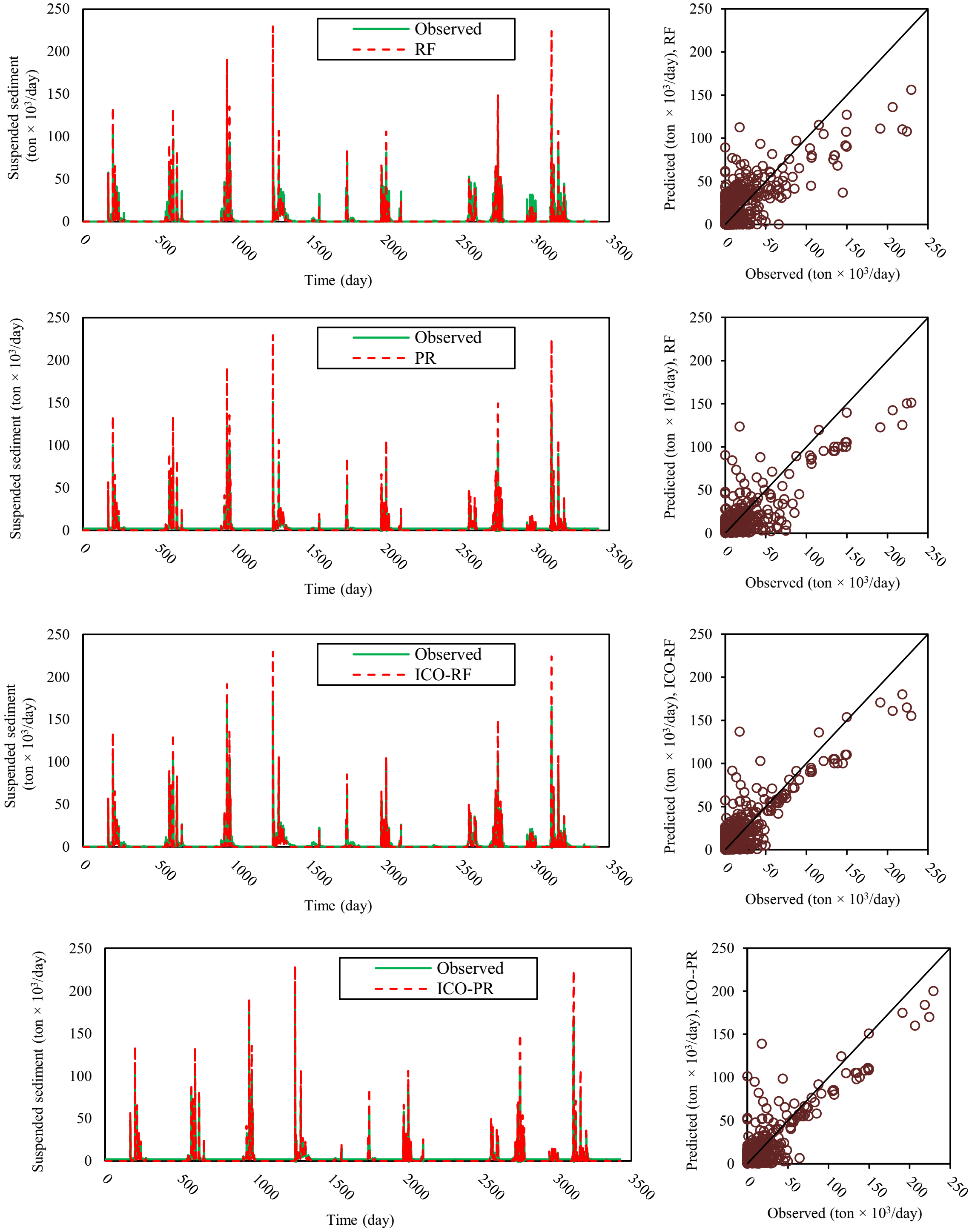

Fig. 5 Observed and predicted suspended sediment load for RF, PR, ICO-RF, and ICO-PR during testing phase

with theoretical basis, and difficult to manage the training process, and slow convergence and inefficiency-related issues.

Since the random forest (RF) can be used to identify and regression, it is common because it can be applied to a wide range of predictive problems, it has a few parameters to change, it is easy to use, and it has been applied successfully to many practical problems and can handle small sample sizes, high-dimensional feature spaces, and complex data structures (Tyralis and Papacharalampous 2017).

The pace regression may be an improvement over the other regression, as it includes measuring the impact of each variable and using a clustering approach to strengthen the statistical basis (Wei 2016). As indicated in Wang and Witten (1999), the pace regression is outperforming, because it in a general sense contradicts the least squares theory. According to Naing and Htike (2015), RF algorithm performs well in short time series one-step ahead of prediction.

Conclusions

In this study, the efficiency of four artificial intelligent techniques of the stand-alone models of RF, PR, and hybrid models of ICO-RF and ICO-PR, was assessed for estimation of the suspended sediment load over a station in the Seonath river basin located in India. Daily discharge and suspended sediment load data of one-three ahead historical records are used for the modeling. Different input combinations were examined on all studied models to select the best scenario for further analysis. Comparison of the developed models based on the variety of statistical error measurement indices showed that the hybrid ICO-RF and ICO-PR techniques provide better performance for estimating the suspended sediment load, and have been performed as the best-ranked 1st and 2nd models, respectively. The results obtained in this study show a satisfactory basis for integrating the ICO as an optimizer technique to promote RF and PR model performance in prediction problems. Results show that optimization of RF with ICO approach enhances the model performance by a factor of 27%. The stand-alone models of RF and PR significantly underestimate suspended sediment load. Hybrid models of ICO-RF and ICO-PR can accurately capture the extreme suspended sediment load values, demonstrating their robustness for application in hydrological problems. Considering the results, the potential alternative optimizer techniques such as fire fly algorithm, multi-verse optimization can be used to boost the single RF and PR model for suspended sediment load prediction and applied to alternative hydrological problems which may be considered future research directions.

Acknowledgments The authors extend their sincere thanks to the editors of this journal and the anonymous reviewers for their valuable comments and suggestions that have significantly improved the manuscript.

Author contributions Conceptualization: Sarita Gajbhiye Meshram, Mir Jafar Sadegh Safari; Data curation: Sarita Gajbhiye Meshram; Formal analysis: Mir Jafar Sadegh Safari, Khabat Khosravi; Investigation: Sarita Gajbhiye Meshram, Mir Jafar Sadegh Safari; Methodology: Mir Jafar Sadegh Safari, Khabat Khosravi; Resources: Sarita Gajbhiye Meshram; Software: Khabat Khosravi; Supervision: Sarita Gajbhiye Meshram, Chandrashekhar Meshram; Validation/ Visualization: Sarita Gajbhiye Meshram, Mir Jafar Sadegh Safari; Writing - original draft: Sarita Gajbhiye Meshram, Chandrashekhar Meshram; Writing - review & editing: Mir Jafar Sadegh Safari, Chandrashekhar Meshram.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Compliance with ethical standards

Conflicts of interests The authors declare that they have no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

Adnan H, Habib A, Ashraf J, Mussadiq S, Ali Raza A (2019) Deep neural network based m-learning model for predicting mobile learners' performance. Turk J Electr Eng Comput Sci 28:1422–1441. https://doi.org/10.3906/elk-1907-8

Alp M, Cigizoglu HK (2007) Suspended sediment load simulation by two artificial neural network methods using hydro-meteorological data. Environ Model Softw 22(1):2–13

Altunkaynak A (2009) Sediment load prediction by genetic algorithms. Adv Eng Softw 40(9):928–934

Asadi S, Shahrabi J, Abbaszadeh P, Tabanmehr S (2013) A new hybrid artificial neural networks for rainfall–runoff process modeling. Neurocomputing 121:470–480

Besaw LE, Rizzo DM, Bierman PR, Hackett WR (2010) Advances in ungauged stream flow prediction using artificial neural networks. J Hydrol 386:27–37

Biau G, Scornet E (2016) A random forest guided tour. https://doi.org/10.1007/s11749-016-0481-7

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Cutler A (2004) Random forests. Department of Statistics, University of California, Berkeley.

Bui XN, Nguyen H, Choi Y, Nguyen-Thoi T, Zhou J, Dou J (2020) Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. Sci Rep 10:9939

Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831

Chen XY, Chau KWA (2016) Hybrid double feedforward neural network for suspended sediment load estimation. Water Recourses Management 30:2179–2194

Chen X, Wang M, Zhang H (2011) The use of classification trees for bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 1:55–63

Chen W, Hong H, Li S, Shahabi H, Wang Y, Wang X, Ahmad BB (2019) Flood susceptibility modeling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. J Hydrol 575:864–873

Chiang JL, Tsai YS (2011) Suspended sediment load estimate using support vector machines in Kaoping river basin, in: Consumer Electronics, Communications and Networks (CECNet), International Conference On. pp. 1750–1753.

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. Sci Total Environ 615:272–281

Cigizoglu HK (2004) Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons. Adv Water Resour 27(2):185–195

Coulibaly P, Anctil F, Bobée B (2000) Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J Hydrol 230(3-4):244–257

Dudek G (2014) Short-term load forecasting using random forests. DOI: https://doi.org/10.1007/978-3-319-11310-4_71

Ehteram M, Ghotbi S, Kisi O, Najah Ahmed A, Hayder G, Ming Fai C, Krishnan M, Abdulmohsin A, Fan H, El-Shafie A (2019) Investigation on the potential to integrate different artificial intelligence models with metaheuristic algorithms for improving river suspended sediment predictions. Appl Sci 9(19):41–49

Fathian F, Mehdizadeh S, Sales AK, Safari MJS (2019) Hybrid models to improve the monthly river flow prediction: integrating artificial intelligence and non-linear time series models. J Hydrol 575:1200–1213

Firat M, Gungor M (2009) Generalized regression neural networks and feed forward neural networks for prediction of scour depth around bridge piers. Adv Eng Softw 40(8):731–737

Gajbhiye S, Mishra SK, Pandey A (2015) Simplified sediment yield index model incorporating parameter CN. Arab J Geosci 8(4): 1993–2004

Goeschel K (2016) Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis. Con, Southeast

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, Second Edition (Springer Series in Statistics). Edition: 2nd ed. 20 Publisher: Springer, ISBN: 0387848576.

Herrera M, Torgo L, Izquierdo J, Pérez-García R (2010) Predictive models for forecasting hourly urban water demand. J Hydrol 387(1-2):141–150

Huo Z, Feng S, Kang S, Dai X (2012) Artificial neural network models for reference evapotranspiration in an arid area of northwest China. J Arid Environ 82:81–90

Iraji H, Mohammadi M, Shakouri B, Meshram SG (2020) Predicting reservoirs volume reduction using artificial neural network. Arab J Geosci. https://doi.org/10.1007/s12517-020-05772-2

Jayech K, Mahjoub MA (2011) Clustering and Bayesian network for image of faces classification, International Journal of Advanced Computer Science and Applications, Special Issue on Image Processing and Analysis.

Juan C, Genxu W, Tianxu M, Xiangyang S (2017) ANN model-based simulation of the runoff variation in response to climate change on the Qinghai-Tibet Plateau, China. Adv Meteorol 2017(9451802):1–13. https://doi.org/10.1155/2017/9451802

Kargar K, Safari MJS, Mohammadi M, Samadianfard S (2019) Sediment transport modeling in open channels using neuro-fuzzy and gene expression programming techniques. Water Sci Technol 79(12): 2318–2327

Khalil B, Ouarda Taha BMJ, St-Hilaire (2011) A Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J Hydrol 405(3):277–287

Khosravi K, Nohani E, Maroufinia E, Pourghasemi HR (2016) A GIS -based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights of evidence bivariate statistical models with multi -criteria method. Nat Hazards 83(2):1–41

Khosravi K, Mao L, Kisi O, Yaseen ZM, Shahid S (2018) Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile. J Hydrol 567: 165–179

Khosravi K, Daggupati P, Alami MT, Awadh SM, Ghareb MI, Panahi M (2019) Meteorological data mining and hybrid data -intelligence models for reference evaporation simulation: a case study in Iraq. Comput Electron Agric 167:105041

Khosravi H, Sadiq S, Gasevic D (2020) Development and adoption of an adaptive learning system. In: Proceedings of the 51st ACM technical symposium on computer science education.

Kisi O (2010) River suspended sediment concentration modeling using a neural differential evolution approach. J Hydrol 389(1-2):227–235

Kisi O, Yaseen ZM (2019) The potential of hybrid evolutionary fuzzy intelligence model for suspended sediment concentration prediction. Catena 174:11–23

Kisi O, Haktanir T, Ardiclioglu M, Ozturk O, Yalcin E, Uludag S (2009) Adaptive neuro-fuzzy computing technique for suspended sediment estimation. Adv Eng Softw 40(6):438–444

Kisi O, Dailr AH, Cimen M, Shiri J (2012) Suspended sediment modeling using genetic programming and soft computing techniques. J Hydrol 450–451:48–58. https://doi.org/10.1016/j.jhydrol.2012.05.031

Kumar D, Pandey A, Sharma N, Flügel WA (2016) Daily suspended sediment simulation using machine learning approach. Catena 138:77–90

Kumar T, Jhariya DC, Pandey HK (2019) Comparative study of different soil erosion and sediment yield models for Pairi Watershed, Chhattisgarh, India. Geocarto Int 35:1245–1266. https://doi.org/10.1080/10106049.2019.1576779

Legates DR, McCabe GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35(1):233–241

Mehr AD, Kahya E, Ahin AS, Nazemosadat MJ (2015) Successive-station monthly stream flow prediction using different artificial neural network algorithms. Int J Environ Sci Technol 12:2191–2200

Meshram SG, Ghorbani MA, Shamshirband S, Karimi V, Meshram C (2019b) River flow prediction using hybrid PSOGSA algorithm based on feedforward neural network. Soft Comput 23(20):10429–10438. https://doi.org/10.1007/s00500-018-3598-7

Meshram SG, Powar PL, Singh VP, Meshram C (2018) Application of cubic spline in soil erosion modelling from Narmada Watersheds, India. Arab J Geosci 11:362. https://doi.org/10.1007/s12517-018-3699-8

Meshram SG, Singh VP, Kisi O, Karimi V, Meshram C (2020) Application of artificial neural networks, support vector machine and multiple model-ANN to sediment yield prediction. Water Resource Management. https://doi.org/10.1007/s11269-020-02672-8

Meshram SG, Ghorbani MA, Deo RC, Kashani MH, Meshram C, Karimi V (2019a) New approach for sediment yield forecasting with a two-phase feedforward neuron network-particle swarm optimization model integrated with the gravitational search algorithm. Water Resour Manag 33(7):2335–2356

Mirabbasi R, Kisi O, Sanikhani H, Meshram SG (2019) Monthly long-term rainfall estimation in Central India using M5Tree, MARS, LSSVR, ANN and GEP models. Neural Comput Applic 31:6843–6862

Morris GL, Annandale G, Hotchkiss R (2008) Reservoir sedimentation, in Garcia M.H. ed. Sedimentation Engineering, Published by American Society of Civil Engineering.

Mustafa MR, Rezaur RB, Saiedi S, Isa MH (2012) River suspended sediment prediction using various multilayer perceptron neural network training algorithms—a case study in Malaysia. Water Resour Manag 26(7):1879–1897

Naing WYN, Htike ZZ (2015) Forecasting of monthly temperature variations using random forests. APRN J Eng Appl Sci 10:10109–10112

Narayana VV, Ram Babu D (1983) Estimation of soil erosion in India. J. Irrig. and Drainage Eng. ASCE 109(4):419–433

Nigam GK, Sahu RK, Sinha MK, Deng X, Singh RB, Kumar P (2017) Field assessment of surface runoff, sediment yield and soil erosion in opencast mines in Chirimiri area, Chhattisgarh, India. Phys Chem Earth 101:137–148. https://doi.org/10.1016/j.pce.2017.07.001

Nourani V, Alizadeh F, Roushangar K (2016) Evaluation of a two-stage SVM and spatial statistics methods for modeling monthly river suspended sediment load. Water Resour Manag 30(1):393–407

Nourani V, Molajou A, Tajbakhsh AD, Najafi H (2019) A wavelet based data mining technique for suspended sediment load modeling. Water Resour Manag 33(5):1769–1784

Omondi RA, Rajapakse CJ (2010) "FPGA Implementations of Neural Networks", 1st edition, Springer publishing company.

Rajaee T, Mirbagheri SA, Zounemat-Kermani M, Nourani V (2009) Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models. Sci Total Environ 407(17):4916–4927

Ramezani F, Nikoo M, Nikoo M (2015) Artificial neural network weights optimization based on social-based algorithm to realize sediment over the river. Soft Comput 19(2):375–387

Robbins H (1964) The empirical Bayes approach to statistical decision problems. Ann Math Stat 35(1):1–20

Rudžianskaitė-Kvaraciejienė R, Apanavičienė R, Gelžinis A (2015) Modelling the effectiveness of PPP road infrastructure projects by applying random forests. J Civ Eng Manag 21:290–299

Safari MJS (2020) Hybridization of multivariate adaptive regression splines and random forest models with an empirical equation for sediment deposition prediction in open channel flow. J Hydrol 590:125392

Safari MJS, Ebtehaj I, Bonakdari H, Es-haghi MS (2019) Sediment transport modeling in rigid boundary open channels using generalize structure of group method of data handling. J Hydrol 577:123951

Safari MJS, Arashloo SR, Danandeh Mehr A (2020) Rainfall-runoff modeling through regression in the reproducing kernel Hilbert space algorithm. J Hydrol 587:125014

Sattari MT, Yurekli K, Pal M (2012) Performance evaluation of artificial neural network approaches in forecasting reservoir inflow. Appl Math Model 36(6):2649–2657

Sharghi E, Nourani V, Najafi H, Soleimani S (2019) Wavelet-exponential smoothing: a new hybrid method for suspended sediment load modeling. Environ Process 6(1):191–218

Shiri J, Kisi O (2012) Estimation of daily suspended sediment load by using wavelet conjunction models. J Hydrol Eng 17(9):986–1000

Shirzad A, Safari MJS (2019) Pipe failure rate prediction in water distribution networks using multivariate adaptive regression splines and random forest techniques. Urban Water J 16(9):653–661

Tao H, Sulaiman SO, Yaseen ZM, Asadi H, Meshram SG, Ghorbani MA (2018) What is the potential of integrating phase space reconstruction with SVMFFA data-intelligence model? application of rainfall forecasting over regional scale. Water Resour Manage 32:3935. https://doi.org/10.1007/s11269-018-2028-z

Tayebiyan A, Ahmad Mohammed T, Ghazali AH, Abdul Malek M, Mashohor S (2016) Potential impacts of climate change on precipitation and temperature at Jor Dam Lake. Pertanika J Sci Technol 24(1):213–224

Tayfur G (2002) Artificial neural networks for sheet sediment transport. Hydrol Sci J 47(6):879–892

Tyralis H, Papacharalampous G (2017) Variable selection in time series forecasting using random forests. Algorithms 10:114. https://doi.org/10.3390/a10040114

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. Pattern Recogn 44:330–349

Wang Y (2000) A new approach to fitting linear models in high dimensional spaces. PhD thesis. Department of Computer Science, University of Waikato, New Zealand.

Wang Y, Witten IH (1999) Pace regression (working paper 99/12). University of Waikato, Department of Computer Science, Haminton, New Zealand.

Wei CC (2016) Comparing single- and two-segment statistical models with a conceptual rainfall-runoff model for river streamflow prediction during typhoons. Environ Model Softw 85:112–128

White R (2001) Evacuation of sediments from reservoirs. Thomas Telford Press, London

Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G (2016) Achieving human parity in conversational speech recognition, Microsoft Res. Tech. Rep. MSR-TR-2016-71. http://arxiv.org/abs/1610.05256. Accessed 17 Feb 2017

Yadav A, Chatterjee S, Equeenuddin SM (2017) Prediction of suspended sediment yield by artificial neural network and traditional mathematical model in Mahanadi river basin, India. Sustain Water Resour Manag 4:745–759. https://doi.org/10.1007/s40899-017-0160-1

Yadav A, Chatterjee S, Equeenuddin SKMD (2018) Prediction of suspended sediment yield by artificial neural network and traditional mathematical model in Mahanadi river basin, India. Sustain Water Resourc Manag 4:745–759

Yadav A, Chatterjee S, Equeenuddin SM (2020) Suspended sediment yield modeling in Mahanadi River, India by multi-objective optimization hybridizing artificial intelligence algorithms. Int J Sediment Res. https://doi.org/10.1016/j.ijsrc.2020.03.018

Yaseen ZM, Ebtehaj I, Kim S, Sanikhani H, Asadi H, Ghareb MI, Bonakdari H, Mohtar WHMW, Al-Ansari N, Shahid S (2019a) Novel hybrid data-intelligence model for forecasting monthly rainfall with uncertainty analysis. Water 11:502

Yaseen ZM, Ehteram M, Hossain S, Chow MF, Koting S et al (2019b) A novel hybrid evolutionary data-intelligence algorithm for irrigation and power production management: application to multi-purpose reservoir systems. Sustainability 11:1953

Yilmaz B, Aras E, Nacar S, Kankal M (2018) Estimating suspended sediment load with multivariate adaptive regression spline, teaching-learning based optimization, and artificial bee colony models. Sci Total Environ 639:826–840

Yunus AJM, Nakagoshi N (2004) Effects of seasonality on streamflow and water quality of the Pinang River in Penang Island, Malaysia. Chin Geogr Sci 14(2):153–161

Zounemat-Kermani M (2016) Assessment of several nonlinear methods in forecasting suspended sediment concentration in streams. Hydrol Res 48(5):1240–1252

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.