Science of the Total Environment 615 (2018) 272–281

Contents lists available at ScienceDirect

Science of the Total Environment

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

River suspended sediment modelling using the CART model: A comparative study of machine learning techniques

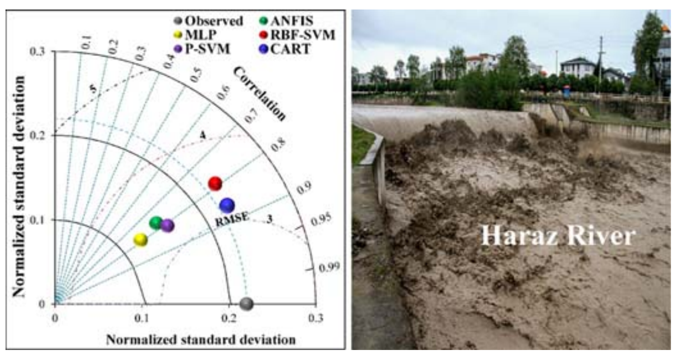
Bahram Choubin [a,*], Hamid Darabi [a], Omid Rahmati [b], Farzaneh Sajedi-Hosseini [a], Bjørn Kløve [c]

[a] Department of Watershed Management, Sari Agriculture Science and Natural Resources University, P.O. Box 737, Sari, Iran
[b] Department of Watershed Management, Faculty of Natural Resources and Agriculture, Lorestan University, Iran
[c] Water Resources and Environmental Engineering, University of Oulu, P.O. Box 4300, FIN-90014 Oulu, Finland

HIGHLIGHTS

- The CART model was employed to estimate the monthly suspended sediment load (SSL) of the Haraz River.
- The proposed model was compared with SVM, ANFIS and ANN models.
- The influence of different time lags was explored.
- The best input combination was identified using Taylor diagrams and violin plots.
- Nash-Sutcliffe efficiency coefficient for SSL prediction by the CART model was 0.77.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 July 2017
Received in revised form 26 September 2017
Accepted 27 September 2017
Available online xxxx

Editor: D. Barcelo

Keywords:
Adaptive neuro-fuzzy inference system
Classification and regression trees
Haraz watershed
Multi-layer perceptron neural network
Support vector machine
Suspended sediment load

ABSTRACT

Suspended sediment load (SSL) modelling is an important issue in integrated environmental and water resources management, as sediment affects water quality and aquatic habitats. Although classification and regression tree (CART) algorithms have been applied successfully to ecological and geomorphological modelling, their applicability to SSL estimation in rivers has not yet been investigated. In this study, we evaluated use of a CART model to estimate SSL based on hydro-meteorological data. We also compared the accuracy of the CART model with that of the four most commonly used models for time series modelling of SSL, i.e. adaptive neuro-fuzzy inference system (ANFIS), multi-layer perceptron (MLP) neural network and two kernels of support vector machines (RBF-SVM and P-SVM). The models were calibrated using river discharge, stage, rainfall and monthly SSL data for the Kareh-Sang River gauging station in the Haraz watershed in northern Iran, where sediment transport is a considerable issue. In addition, different combinations of input data with various time lags were explored to estimate SSL. The best input combination was identified through trial and error, percent bias (PBIAS), Taylor diagrams and violin plots for each model. For evaluating the capability of the models, different statistics such as Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE) and percent bias (PBIAS) were used. The results showed that the CART model performed best in predicting SSL (NSE = 0.77, KGE = 0.8, PBIAS < ±15), followed by RBF-SVM (NSE = 0.68, KGE = 0.72, PBIAS < ±15). Thus the CART model can be a helpful tool in basins where hydro-meteorological data are readily available.

* Corresponding author.
E-mail address: bahram.choubin@ut.ac.ir (B. Choubin).

https://doi.org/10.1016/j.scitotenv.2017.09.293

1. Introduction

Excess sediment in streams and reservoirs leads to degradation of surface water quality and increased drinking water treatment costs. Sediment also transports significant amounts of nutrients, metal elements, hazardous contaminants can potentially be carried by the SSL in a river (Embabi et al., 2014; Malagó et al., 2017; Moran et al., 2017; Williams et al., 2008). Sediment in excess can lead to mortality of riverine fish during the embryonic life stage, restriction of periphyton growth and limited primary productivity (Ali et al., 2012; Fox et al., 2016; Funes et al., 2017; Greig et al., 2005; Heywood and Walling, 2007; Lloyd et al., 2016; Makarynskyy et al., 2015; Malagó et al., 2017; Migiros et al., 2011; Mu et al., 2017). In addition, increased sediment loads result in loss of reservoir storage capacity, reducing the ability of reservoirs to serve various human needs, including water supply, hydropower and flood control (Fan et al., 2012; Rahmati et al., 2016; Suif et al., 2016; Tang et al., 2014). Hence, understanding the processes of sediment transport and sediment yield and their effects on sedimentation in reservoirs, shallow harbours and channels is important for policy makers and decision makers in water resources management (Palazón et al., 2016; Si et al., 2017). Reliable and accurate models for understanding the temporal variations in sediment yield and for predicting suspended sediment dynamics at watershed scale are thus needed for environmental development and management (Rodriguez-Lloveras et al., 2016; Xia et al., 2016).

Sediment dynamics in river basins are influenced by a series of complex natural processes. Sediment input to rivers due to erosion is strongly enhanced by human activities such as deforestation, land use change, overgrazing, agricultural activities and mismanagement of rivers (Ali and Abbas, 2013; Bathrellos et al., 2017; Heng and Suetsugi, 2014; Rovira et al., 2015; Silgram et al., 2010; Withers et al., 2006; Withers and Haygarth, 2007). Estimating and predicting sediment load in a river system requires robust models which can handle nonlinear relationships and missing values (Dai and Lu, 2010; Vigiak et al., 2017; Wenske et al., 2012).

Several approaches have been used to model suspended sediment load (SSL), including (i) hydraulic/numerical (i.e., sediment transport models); (ii) physical/mathematical (i.e., distributed physically-based and lumped conceptual models); (iii) statistics-based (e.g., copula function-based models); and (iv) empirical models (e.g., data-driven models) (Bezak et al., 2014; Jha and Bombardelli, 2011; Merkhali et al., 2015). Hydraulic/numerical models are complex and time-consuming, because they solve the differential equations of two-phase stream discharge and sediment modelling (Jha and Bombardelli, 2011). Sophisticated physical/mathematical models require data on the true spatial distribution of most relevant variables (e.g., precipitation, evaporation etc.) and on spatial variation in watershed properties, and generally involve lengthy modelling processes (Hamel et al., 2017; Kothyari et al., 1997; Refsgaard, 1997). Since the data required are often not available in developing countries, physical/mathematical models are generally impractical or lead to uncertainty in the results (Goyal et al., 2014; Guldal and Muftuoglu, 2001).

Data-driven models such as artificial intelligence and soft computing techniques have also recently attracted attention with respect to prediction of SSL (Afan et al., 2015; Ali et al., 2013; Alp and Cigizoglu, 2007; Kisi, 2008; Kisi et al., 2012; Makarynskyy et al., 2015; Mustafa et al., 2011; Partal and Cigizoglu, 2008). For example, Partal and Cigizoglu (2008) applied multi-layer perceptrons (MLPs), the most frequently used artificial neural network model in water resources engineering, for forecasting daily suspended sediment and found that MLPs can successfully model the complex non-linear pattern of sediment data series significantly better than conventional models. Chiang and Tsai (2011) predicted SSL in the Kaoping river basin, Taiwan, using support vector machine (SVM) and artificial neural network (ANN) models and demonstrated that SVM outperforms ANN. Similarly, Çimen (2016) investigated the performance of SVM to predict the SSL of two rivers in the USA and found that a SVM model can predict SSL without producing negative sediment discharge values. Melesse et al. (2011) developed an ANN-based modelling approach for estimating SSL in three major rivers in the USA and demonstrated that the ANN predictions were better than results from integrated moving average (ARIMA), multiple linear regression (MLR) and multiple non-linear regression (MNLR) models. Azamathulla et al. (2012) introduced an approach based on adaptive neuro-fuzzy inference system (ANFIS), which is a combination of neural network and fuzzy logic, for predicting sediment transport in Malaysia. Their results indicated that the ANFIS model gives robust and satisfactory results that can be useful for policy makers. Kisi et al. (2012) compared the genetic programming (GP), ANFIS, ANN and SVM algorithms for predicting daily SSL at two stations in the Cumberland River in the USA and demonstrated that GP is superior to the other models. Recently, Zounemat-Kermani et al. (2016) evaluated the ability of different data-driven models, including four different support vector regression (SVR) model kernels (linear (L-SVM), polynomial (P-SVM), radial basis function (RBF-SVM) and sigmoid (S-SVM)) and three ANN model algorithms (conjugate gradient, gradient descent and Broyden-Fletcher-Goldfarb-Shannon (BFGS)), to estimate daily SSL in rivers in the USA. They found that the BFGS-ANN and RBF-SVR models outperformed the other models in simulating SSL. However, to our knowledge, classification and regression trees (CART), a popular machine learning algorithm, has not been used previously for SSL prediction in river engineering and hydrological studies.

The sediment yield from catchments in northern Iran is mainly produced during high-intensity precipitation events. Due to the complexity of erosion and sediment transport processes in the region, the current strategy in environmental research is to use data-based approaches and computing methods for modelling e.g., suspended sediment in the Haraz river watershed (one of the most important coastal watersheds in northern Iran; IWRMC, 2015). Therefore, current research developed a modelling approach that can be applied to reduce the frequency of costly operations for SSL estimation in watersheds in developing countries where hydrological datasets are incomplete. The main objectives of this study were to: i) develop and apply a CART model to estimate SSL, since its accuracy in predicting SSL has not previously been determined, but may exceed that of previous predictive models; ii) evaluate the performance of the CART model in SSL estimation using the most commonly employed criteria and observed sediment data; iii) compare results obtained using the CART model to results obtained using the MLP, ANFIS, RBF-SVM and P-SVM models; and iv) explore the influence of different time lags and various combinations of input datasets on the accuracy of the estimations.

2. Materials and methods

2.1. Study area

The Haraz watershed (4014 km^2) is one of the largest coastal watersheds in northern Iran, extending between 51°43′31″ to 52°36′12″ East and 35°45′44″ to 36°22′25″ North (Fig. 1). The Kareh-Sang hydrometric station, located at the outlet of the Haraz watershed, had a mean measured annual discharge of 27.72 $m^3 s^{-1}$ during 1981–2010. The triangles in Fig. 1 indicate the location of meteorological stations. The climate type in the region is dry sub-humid (aridity index: 0.64 based on the Sahin, 2012), with 832 mm annual precipitation and 1300 mm potential evapotranspiration (1981–2010). The soil type of case study is divided into 4 classes that include Inceptisol (lithic or paralithic contact and soil with no accumulation of clays and organic matter), Mollisol (an argillic or natric horizon or soil with high organic matter and typically under a grassland cover), Entisol (aquic conditions within 50 cm soil surface and soil with no any profile development other than an A horizon), and Alfisol (aquic conditions for some time in normal years and soil with high native fertility and typically under a hardwood forest cover and they are widely used both in agriculture and deciduous