Science of the Total Environment 955 (2024) 176758

Contents lists available at ScienceDirect

Science of the Total Environment

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

Reconstructing suspended sediment concentrations in the Mekong River Basin via semi-supervised-based deep neural networks

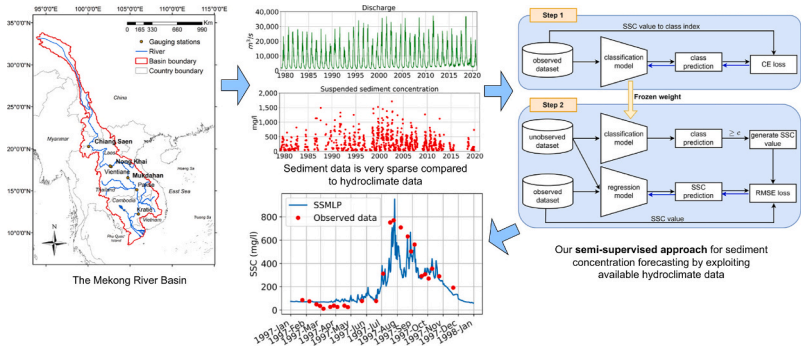
Thi Thu Ha Nguyen [a,b], Duc Quang Vu [c], Ngoc Phu Doan [d], Huynh Thi Khanh Chi [d], Peixin Li [d,e], Doan Van Binh [f], Yimeng An [d,e], Pham Tuan Dung [d], Tuan A. Hoang [d], Mai Thai Son [d,*]

[a] Laboratory of Environmental Sciences and Climate Change, Institute for Computational Science and Artificial Intelligence, Van Lang University, Ho Chi Minh City, Vietnam
[b] Faculty of Environment, School of Technology, Van Lang University, Ho Chi Minh City, Vietnam
[c] Department of Computer Science and Information System, Thai Nguyen University of Education, Vietnam
[d] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom
[e] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, China
[f] Master program in Water Technology, Reuse and Management, Faculty of Engineering, Vietnamese-German University, Ben Cat City, Binh Duong Province, Vietnam

HIGHLIGHTS

- A novel semi-supervised framework for reconstructing sediment concentration data
- The effective solution for extremely missing sediment data in the Mekong River Basin
- The proposed framework exploits other hydro-climate data as supporting sources.
- The proposed method can dramatically boost the performance of all models at the studied stations.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ashantha Goonetilleke

Keywords:
Suspended sediment concentration
Sediment reconstruction
Mekong River Basin
Semi-supervised learning

ABSTRACT

The Mekong River Basin (MRB) is crucial for the livelihoods of over 60 million people across six Southeast Asian countries. Understanding long-term sediment changes is crucial for management and contingency plans, but the sediment concentration data in the MRB are extremely sporadic, making analysis challenging. This study focuses on reconstructing long-term suspended sediment concentration (SSC) data using a novel semi-supervised machine learning (ML) model. The key idea of this approach is to exploit abundant available hydroclimate data to reduce training overfitting rather than solely relying on sediment concentration data, thus enhancing the accuracy of the employed ML models. Extensive experiments on daily hydroclimate and SSC data obtained from 1979 to 2019 at the three main stations (i.e., Chiang Saen, Nong Khai, and Mukdahan) are conducted to demonstrate the superior performance of the proposed method compared to the state-of-the-art supervised

* Corresponding author at: Computer Science Building, 16A Malone Road, Belfast, Northern Ireland BT9 5BN, United Kingdom.
E-mail addresses: ha.nguyenthithu@vlu.edu.vn (T.T.H. Nguyen), quangvd@tnue.edu.vn (D.Q. Vu), ndoan01@qub.ac.uk (N.P. Doan), thuynh01@qub.ac.uk (H.T.K. Chi), p.li@qub.ac.uk (P. Li), binh.dv@vgu.edu.vn (D.V. Binh), y.an@qub.ac.uk (Y. An), tpham01@qub.ac.uk (P.T. Dung), t.hoang@qub.ac.uk (T.A. Hoang), thaison.mai@qub.ac.uk (M.T. Son).

https://doi.org/10.1016/j.scitotenv.2024.176758
Received 30 April 2024; Received in revised form 3 October 2024; Accepted 4 October 2024
Available online 12 October 2024

techniques (i.e., Random Forest, XGBoost, CatBoost, MLP, CNN, and LSTM), and surpasses existing semi-supervised methods (i.e., CoReg, ⊓ Model, ICT, and Mean Teacher). This approach is the first semi-supervised method to reconstruct sediment data in the field and has the potential for broader application in other river systems.

1. Introduction

The Mekong River is a trans-boundary river in Southeast Asia, flowing from the Tibetan Plateau through China, Myanmar, Thailand, Laos, Cambodia, and Vietnam. It has an estimated length of approximately 5000 km with 795,000 km^2 drainage area, is home to significant biodiversity with over 20,000 plant species, 2500 animal species, and 850 fish species (MRC, 2005, 2007). However, the Mekong River Basin (MRB) faces severe threats from climate change (e.g., droughts, floods, or sea level rise) and human activities (e.g., deforestation, sand mining, ground water extraction, and dam constructions), leading to issues such as sediment starvation, salinity intrusion, and reduced water levels (Tuan and Chinvanno, 2011; Piman and Shrestha, 2017; Kondolf et al., 2014). Sediments and its attached nutrients are crucial for the MRB's ecosystem, as it supports soil nutrition, fisheries, and agriculture (Manh et al., 2014; Piman and Shrestha, 2017). However, sediment transport in the downstream MRB has drastically reduced, with estimates a decrease of 74.1 % (2012–2015 period) compared to pre-dam periods (Binh et al., 2020a), potentially reaching a 96 % reduction if all the planned dams would be successfully completed (Kondolf et al., 2014). This poses significant risks to the MRB's ecosystem and the livelihoods of millions. Historical sediment concentration data is essential for analyzing the impacts of human activities, providing significant information for further understanding about morphological changes, salinity intrusion and coastal erosion. Unlike other factors such as discharge, which has been collected daily for a long time, sediment data in the MRB are relatively sporadic, even missing data for the whole months or years in different stations (Lu and Siew, 2006). Thus, there is a need to reconstruct missing sediment concentration values to provide a long-term and reliable continuous sediment data for further studies of sediment transports in the MRB.

Most of existing works to reconstruct the long-term sediment data in the MRB rely on traditional sediment rating curves, which are a fitted relationship between suspended sediment concentration (SSC) and river water discharge (Q) (Binh et al., 2020a; Wang et al., 2011; Lu et al., 2014). However, this method is highly biased, especially during extreme events (droughts and floods) or under changing hydrological conditions (Warrick, 2015; Walling, 1977). The sediment transport models (e.g., Soil and Water Assessment Tool (SWAT), Telemac, Coupled Ocean–Atmosphere–Wave–Sediment Transport Modeling system (COAWST), Delft3D) was used to estimate sediment transport in the MRB (Sam and Khoi, 2022; Binh et al., 2022; Xue et al., 2012; Thanh et al., 2019), but require many different inputs (e.g., hydroclimate data, soil map, land cover, topographic, etc.), which are not available at all sites, extensive expert knowledge, and experiences to fine-turn them (Xu and He, 2022). Additionally, remote sensing techniques, which have been studied for other areas like Mississippi River (Umar et al., 2018), would also be adapted for reconstructing the sediment data. However, satellite data have its own limitations because the complicated climate (i.e., rainy or cloud coverage) and geography in the MRB can affect its operations and spatiotemporal data availability. The machine learning (ML) has been emerged as a tempting alternative for SSC estimation due to many attractive benefits, e.g., relying only on historical data for building models without any simulation assumption, do not need excessive amount of data, less biased by experts, or require less computational power than conventional models (Essam et al., 2022; Kaveh et al., 2021; Khan et al., 2016; Nguyen et al., 2023b). The ML techniques are very diverse, e.g., classification and regression tree (CART), Artificial Neural Networks (ANNs), multi-layer perceptron (MLP) neural network, Support Vector Machine (SVM), and long short-term memory (LSTM), which are used to predict sediment in the Haraz watershed (Iran), Mississippi River (the North American continent), Schuylkill River (USA), Ramganga River (India), and Bhagirathi river (Himalaya) (Choubin et al., 2018; Shadkani et al., 2021; Meshram et al., 2021; Kaveh et al., 2021; Singh and Khan, 2020; Khan et al., 2019a,b). The MLP, SVM, and LSTM are widely applied to predict daily or monthly sediment, and the high model performance is found in Shadkani et al. (2021); Meshram et al. (2021); Kaveh et al. (2021). Choubin et al. (2018) compared the accuracy of the CART model with the most commonly ML models (i.e., MLP and SVM) to estimate the monthly suspended sediment load, showing the best performance of the CART model. Singh and Khan (2020) and Khan et al. (2019a, 2019b) used the ANNs model to predict the SSC in India, indicating the superior of the proposed model with high values of coefficient of determination. These above-mentioned models can be used for reconstructing sediment data. However, a large amount of available sediment data is also required for effectively training them, which are currently unavailable for the MRB (for example, 93.96 % data in Mukdahan station is missed from 1979 to 2019 (c.f. Section 2)). This makes accurately predicting/reconstructing SSC in the MRB a very challenging task.

In this paper, a novel and effective Deep Learning (DL)-based framework to reconstruct SSC data for the MRB, which can overcome the current data scarcity problem and can provide long-term continuous and reliable SSC data to support further research, is introduced. Concretely, the contributions are summarized as the following. First, contrary to all existing works, which are supervised-based methods and need long-term continuous sediment data for training their models, e.g., Darabi et al. (2021); Essam et al. (2022), the proposed semi-supervised DL framework (SSL) is specifically designed to significantly enhance the SSC reconstruction performance, especially under data scarcity. The key idea of this algorithm is that rather than ignoring all data related to missing sediment values (e.g., daily precipitation and discharge) during the ML training process, they are exploited as an additional source to aid the learning process of the selected ML model to improve the prediction accuracy. The proposed approach is to construct two separate ML models. The first model, called the classification model or classifier, attempts to predict the concentration ranges of sediments and is trained using available sediment data (i.e., observed data). The second model, called the regression model or regressor, is initially trained with observed sediment data, then it will be updated using unobserved data under the guidance of the classifier by minimizing disagreement between them. By this way, input hydroclimate data with missing sediment values can be exploited to enrich the training data, thus avoiding overfitting problems and improving the final prediction accuracy. To the best of our knowledge, SSL is the first semi-supervised approach for reconstructing sediment data in the field. Second, a wide range of state-of-the-art supervised machine learning techniques are utilized to estimate the missing SSC data at many MRB's mainstream stations, including linear regression (LR), Support Vector Regression (SVR), Random Forest Regression (RF) (Ho, 1995), XGBoost (Chen and Guestrin, 2016), CatBoost (Prokhorenkova et al., 2018), Multi-Layer Perceptron (MLP), Convolutional neural network (CNN) (Gu et al., 2018), and Long short-term memory (LSTM) (Hochreiter, 2010). Furthermore, various state-of-the-art semi-supervised learning methods, that have not been studied before in the sediment reconstruction task, are also employed including CoReg (Zhou et al., 2005), ⊓ Model (Laine and Aila, 2016), ICT (Verma et al., 2022), and Mean Teacher (Tarvainen and Valpola, 2017). These methods are used as baselines for assessing

performance of this framework. Third, though this study focuses on the MRB, sediment data sparsity is a very common problem in many other river systems worldwide (Asselman, 1999; De Vente et al., 2007). The proposed approach can also be applied to these rivers for efficient sediment forecasting/reconstruction. The rest of the paper is organized as follows. In Section 3, the proposed semi-supervised approach is presented. Extensive experiments are conducted in Section 4 to demonstrate the performance of the proposed method, and conclusions are drawn in Section 5.

2. Study area and data

2.1. The Mekong River Basin

The MRB is characterized by a complex orography, which is covered by high mountains in the north, while lowland and floodplain are dominated in the south MRB. Originating at approximately 5000 m above the mean sea level in the Tibetan Plateau (China), the MRB spans over 795,000km^2 of China, Myanmar, Thailand, Laos, Cambodia, and Vietnam. It ends in the Vietnamese Mekong Delta with elevations mostly below 2 m above the mean sea level. The MRB is divided into the upper and lower basins. The upper part covers an average of 189,000km^2 (24 %), and the lower part has an area of 606,000km^2 (76 %) (Lu and Siew, 2006). The upper Mekong (about 2000 km) has elevations from 500 m to 4500 m with an average slope of 2 m/km (Lauri et al., 2012). The lower Mekong River (about 2900 km) has elevations between several meters above the mean sea level to 500 m, with an average slope of 0.25 m/km from Chiang Saen to Kratie and 0.03 m/km from Kratie to the East Sea (Lauri et al., 2012). The Mekong River runs through bedrock and alluvial alternately from its source to its sink. The bedrock ravines start from China territory down to 5 km upstream of Vientiane in Laos, continuing to Kratie (Rubin et al., 2014), from which the Mekong River enters the downstream alluvial portion with a spacious Mekong Delta in Cambodia and Vietnam, characterized by dense river network with mild riverbed slopes.

The MRB's climate is influenced by tropical monsoon, resulting in two distinct seasons, including the wet season (May–October) and the dry season (November–April). The entire MRB has latitudes roughly from 10°N to 35°N (c.f. Fig. 1a), which encompasses high-altitude continental and temperate in the upper basin to tropical monsoon in the lower basin. This results in remarkable variations of annual precipitation between the upstream and downstream MRB (Binh et al., 2020b). For instance, at Mukdahan station (from 1979 to 2019), the precipitation from May to October constitutes 92.63 % of the annual total rainfall, peaking in August (Fig. 1b), and the flood season spans from June to November (Fig. 1c), contributing 77.2 % of the total annual discharge.

2.2. Sediment starvation problem in the MRB

The sediment starvation of the Mekong river and its basin mentioned in Section 1 is strongly linked to human activities, in particular dam building (Binh et al., 2020a; Räsänen et al., 2017). From 1965 to 2019, more than 100 dams were built in the MRB. Many others are expected to be built in the near future, especially in Laos (WLE, 2020). Large dams in the Mekong tributaries were built in the early 1990s, and the boom for hydropower development in the MRB peaked with the operationalization of the two largest dams, Xiaowan and Nuozhadu, completed in 2010 and 2014 in China, respectively. The construction of large dams in the upstream MRB has led to the significant alteration of flow regimes and reduction of sediment in the downstream MRB (Binh et al., 2020a, 2020b; Räsänen et al., 2017). The total sediment load in the downstream MRB is estimated at around 167 Mt./yr (Binh et al., 2020a; Lu et al., 2014) before 1992. However, it was reduced by 74 %, valued at 43.1 Mt./yr from 2012 to 2015. Particularly, six large mainstream dams in the upstream (including Manwan, Dachaoshan, Xiaowan, Jinghong, Nuozhadu, and Gongguoqiao) have accounted for 40.2 % of that reduction of SSC in the downstream regions (Chua and Lu, 2022; Kondolf et al., 2014).

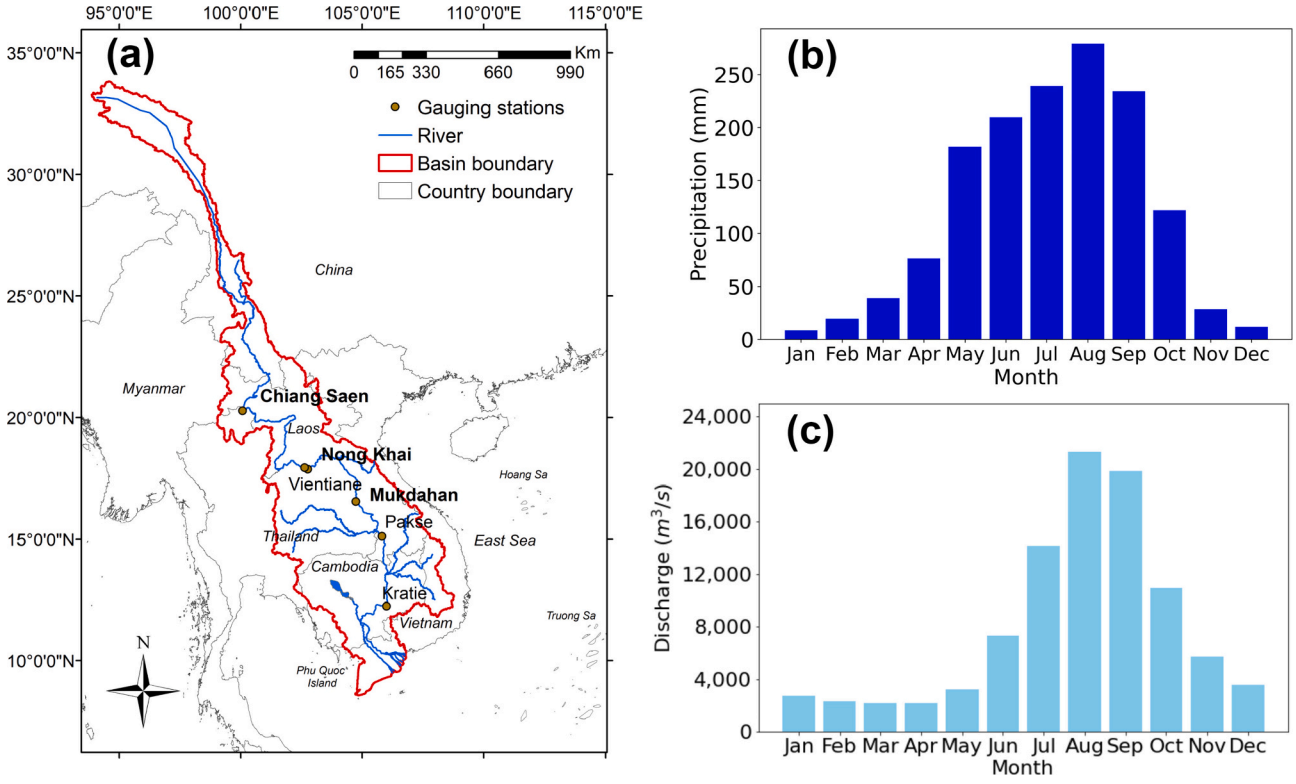


Fig. 1. (a) The Mekong River Basin and gauging stations, (b) mean monthly precipitation, and (c) mean monthly discharge at Mukdahan from 1970 to 2019.

2.3. Data

In this study, the daily discharge (Q) and suspended sediment concentration (SSC) are collected from the Mekong River Commission at main gauging stations (i.e., Chiang Saen, Nong Khai, and Mukdahan as shown in Fig. 1) from 1979 to 2019. These three stations are located in areas of dense dam constructions, and are used to estimate the sediment load in previous studies (Wang et al., 2011; Binh et al., 2020a). The discharge values are on a daily time scale. However, the SSC values are very sporadic, and have been collected from 0 to 12 times per month. The sediment data is obtained from the hydrological data set (HYMOS), which produced by the Mekong River Commission (MRC). The HYMOS data set is developed using the depth-integration method, which is measured at several vertical profiles in a cross section (Binh et al., 2020a). The daily precipitation data (P) are obtained from the Climate Prediction Center from 1979 to 2019, having a spatial resolution of $0.25^{\circ}\times0.25^{\circ}$. These data are also used for MRB's studies (Irannezhad and Liu, 2022; Nguyen et al., 2023a, 2023b; Vu et al., 2018). All gridded data are re-gridded to match with the locations of MRB's meteorological stations by using the bilinear interpolation method, which is successfully adopted in (Dang et al., 2020; Hoang et al., 2016; Nguyen et al., 2023a).

Fig. 2 illustrates the daily precipitation, discharge, and SSC collected from the Mukdahan station from 1979 to 2019. As can be seen, while other data are fully available, the SSC data is very sparse with only 905 values collected during 41 years, i.e., 6.04 % available data or 93.96 % missing data. The data are entirely missing in 1983, 2008, and 2016–2017 due to many different reasons such as lacking of technology, political unrest, or financial limitation. The same situations are observed on Chiang Saen and Nong Khai stations with only 687 (4.59 %) and 879 (5.87 %) available data values, respectively. These huge amounts of missing data make daily SSC prediction/reconstruction a very challenging task for any approach.

3. Methodology

In this section, a brief overview of the problem setup is first provided. Then, the new semi-supervised approach to handle highly missing sediment data is presented.

3.1. Problem setup

Given a set of N observed samples denoted as $\mathscr{D}_o = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_N}, y_N)\}$, where $\mathbf{x_i}$ includes discharge (Q) and precipitation (P) and y_i corresponds to the SSC value. This approach aims to build an efficient model $\mathscr{F}$ to predict the SSC value y_i from the input $\mathbf{x_i}$ as shown in Eq. (1).

$$y_i \approx \hat{y}_i = \mathscr{F}(\mathbf{x_i}) \tag{1}$$

where $\hat{y}_i$ denotes the output prediction of the model $\mathscr{F}$. Far apart from the observed dataset $\mathscr{D}_o$, the unobserved (missing) dataset $\mathscr{D}_u = \{\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_M}\}$ contains M unobserved samples, where $\mathbf{z_j}$ includes Q and P values, the number of samples of $\mathscr{D}_u$ is much larger than $\mathscr{D}_o$ (i.e., $M \gg N$). There is no corresponding SSC value for each $\mathbf{z_j}$ in the dataset $\mathscr{D}_u$.

All existing works like (Tao et al., 2021; Choubin et al., 2018; Kaveh et al., 2021; Essam et al., 2022; Meshram et al., 2021) employ different supervised learning methods to predict the sediment values, i.e., they use the $\mathscr{D}_o$ dataset to train the model and totally ignore the dataset $\mathscr{D}_u$. Hence, when the numbers of observed data are small as in the cases of the MRB (i.e., small values of N), overfitting may happen and thus significantly reduce the overall performance. To overcome this drawback, it is essential to exploit more information from $\mathscr{D}_u$ to improve the model's generalization and avoid overfitting.

3.2. The proposed semi-supervised approach

To exploit more information from unobserved data, the proposed approach contains two steps: training a classification model (classifier) and training a regression model (regressor) as illustrated in Fig. 3.

3.2.1. Step 1: Training a classification model (classifier)

Instead of training a regression model on the observed dataset $\mathscr{D}_o$, the first step of this approach is to construct a classification problem to classify ranges of sediment values. To do this, the range of SSC values from the observed data is found and is divided into K disjoint ranges (i.e., classes). For example, if the SSC ranges from 0 to 3800 and K is equal to 50, that means the proposed method divides the range into 50 parts, and the magnitude (V) of each part is 76 ($V = 3800/50$). As a result, the range of class 0 is set from 0 to 76, the range of class 1 is set from 77 to 153, the range of class 2 is set from 154 to 230, etc., and the range of class 49 is set from 3723 to 3799. Assume that at a data point, the SSC value is of 1000, and the class index of the SSC value of 1000 corresponds to $1000/76 = 13$. Hence, the class index of a data point can be calculated as follows:

$$class\ index = \left\lceil \frac{SSC\ value}{V} \right\rceil \tag{2}$$

where $V = (\text{maximum value of SSC} + 1)/K$, and K is a hyper-parameter. After performing this conversion, the proposed method has transformed a regression task into a classification task with K classes. Let denote $\mathscr{D}_t$ as the new dataset generated from $\mathscr{D}_o$, i.e., $\mathscr{D}_t = \{(\mathbf{x_1}, q_1), (\mathbf{x_2}, q_2), \ldots, (\mathbf{x_N}, q_N)\}$ where $\mathbf{x_i}$ is kept the same value as in $\mathscr{D}_o$ meanwhile q_i corresponds to the class index of the y_i value. In this framework, a classification model $\mathscr{G}$ is built to address this task as follows:

$$\begin{aligned} p_i &= \mathscr{G}_\gamma(\mathbf{x_i}) \\ \hat{q}_i &= argmax(p_i) \end{aligned} \tag{3}$$

where p_i denotes the predictive distribution output from the classification model $\mathscr{G}\mathscr{G}$ and $\hat{q}_i$ are the predicted classes with the highest distribution score. To train the model $\mathscr{G}$, the difference between the prediction and the observation value is minimized by using cross-entropy (CE) (De Boer et al., 2005) as follows:

$$\mathscr{D}_{CE}(q_i, p_i) = -\frac{1}{n}\sum_{i=1}^{n} q_i log(p_i) \tag{4}$$

where n denotes the batch size and note that q_i is converted to the K-dimensional one-hot vector[1] and p_i is the predictive distribution from the network $\mathscr{G}$. The pseudo-code of training the classification model is illustrated in Algorithm 1. The classification is one of the key differences between this algorithm and existing semi-supervised regression methods. Due to very limited observations for sediment data as presented in Section 2, directly training the regression model would be very inefficient due to the continuous nature of the sediment data. So, the proposed approach is to convert the continuous sediment values into range labels and use the classifier to recognize them. First, it reduces the prediction space from indefinite to finite sets of labels, thus making it easier for the learning algorithms to recognize them. Second, the probability of class labels can be exploied from DL outputs to select those with high confident range values to enrich the training data in Step 2, which is now the regressor to produce continuous sediment prediction as described in detail below.

[1] Note that, in supervised learning, the one-hot vector is a way to present the label of the data sample. All the elements are 0 except for one element, which is set to 1. The index of the value 1 in the one-hot vector is equal to the class number of the data point. For example, in the 3-class classification problem, each sample will be assigned to one of three classes: 0, 1, and 2. The one-hot vector of class 0 is [1,0,0], class 1 is [0,1,0], and class 2 is [0,0,1].

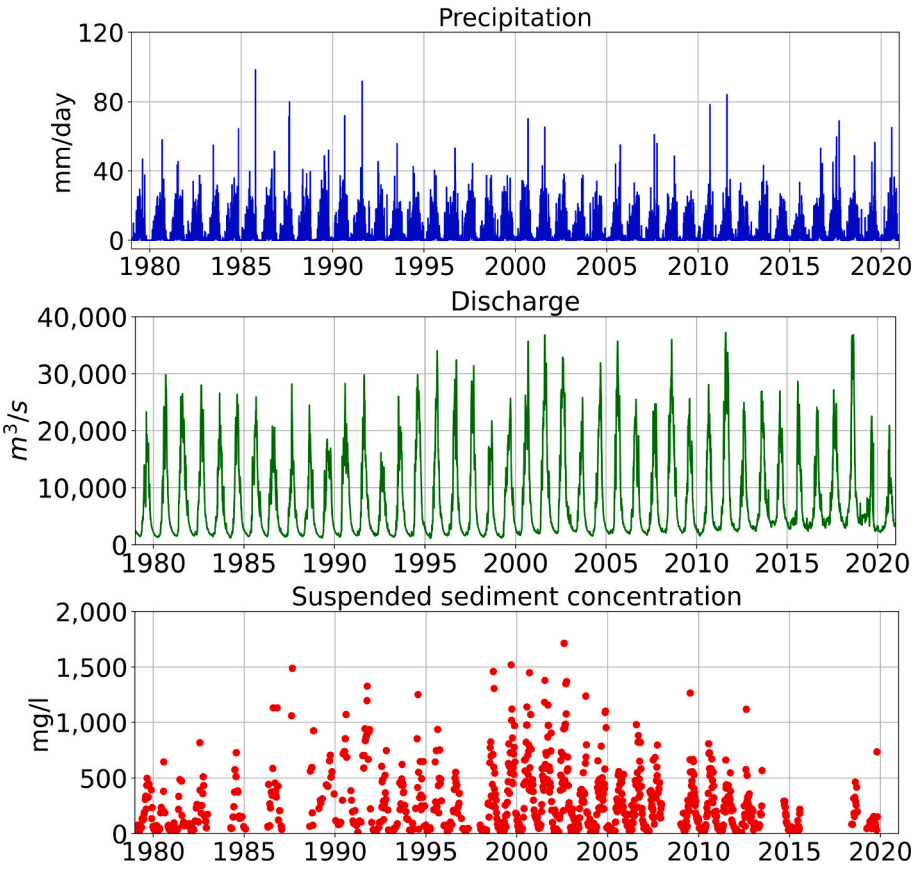


Fig. 2. Precipitation, discharge and SSC values at the Mukdahan station.

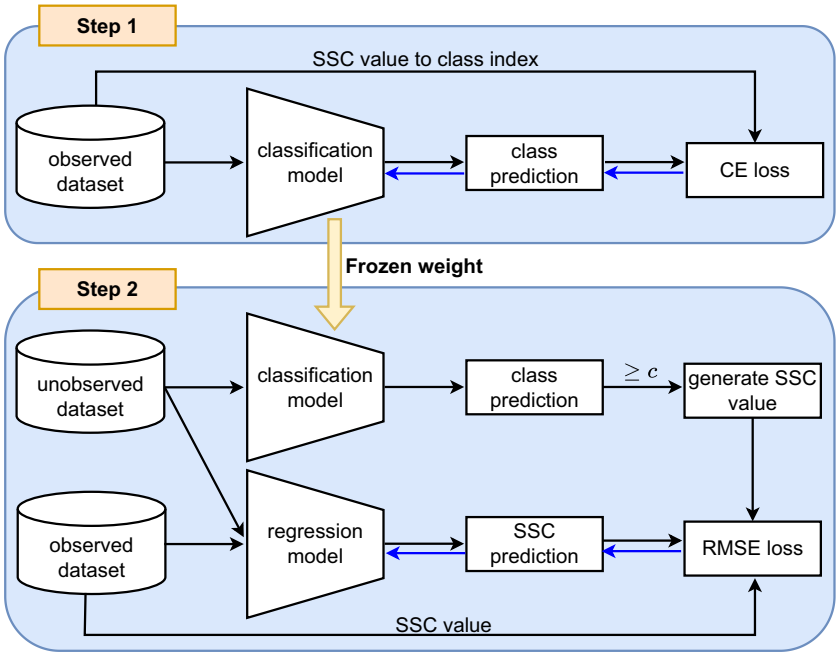


Fig. 3. Overview of the proposed semi-supervised approach (SSL). The black line is the forward path and the blue line denotes back-propagation. There are two loss functions in this approach including Cross Entropy Loss (CE Loss) and Root Mean Square Error Loss (RMSE Loss). The hyperparameter c is the confidence factor.

Algorithm 1. The pseudo-code of the training classification model (Step 1).

Algorithm 2. The pseudo-code of the training regression model (Step 2)

3.2.2. Step 2: Training a regression model (regressor)

Different from traditional regression methods, in this proposed

Input: The observed dataset $\mathcal{D}_o$.
The network $\mathcal{G}_\theta$ where θ is the trainable parameters of $\mathcal{G}$.
The hyper-parameter K classes.

Initialize parameters θ
Convert the dataset $\mathcal{D}_o$ to $\mathcal{D}_t$ via Eq. 2
while $\mathcal{G}_\theta$ has not converged **do**
  Sample a batch $(\mathbf{x}, q)$ from the $\mathcal{D}_t$
  $p = \mathcal{G}_\theta(\mathbf{x})$
  Calculate the loss $\mathcal{L}_{CE}(q, p)$ by Eq. 4
  Update parameters θ by computing the gradient of $\mathcal{L}_{CE}$
end while
return $\mathcal{G}_\theta$

Input: The observed dataset $\mathcal{D}_o$.
The unobserved dataset $\mathcal{D}_u$.
The pre-trained classification network $\mathcal{G}_\theta$ where θ is the frozen weights of $\mathcal{G}$.
The regression network $\mathcal{F}_\gamma$ where γ is the trainable parameters of $\mathcal{F}$.
The hyper-parameters α, β and c.

Initialize parameters γ
while $\mathcal{F}_\gamma$ has not converged **do**
  Sample a batch $(\mathbf{x}, y)$ from the $\mathcal{D}_o$
  Sample a batch $(\mathbf{z})$ from the $\mathcal{D}_u$
  /*—– ***Train on observed data*** —–*/
  $\hat{y} = \mathcal{F}_\gamma(\mathbf{x})$
  Calculate the loss $\mathcal{L}_{RMSE}(y, \hat{y})$ by Eq. 5
  /*—– ***Train on unobserved data*** —–*/
  $p = \mathcal{G}_\theta(\mathbf{z})$ {Get predictive distribution from network $\mathcal{G}$}
  $\hat{q} = \arg\max(p)$ {Get the class with highest distribution score from p}
  Calculate the pseudo SSC $\tilde{y}$ via Eq. 7
  $\hat{y} = \mathcal{F}_\gamma(\mathbf{z})$
  Calculate the loss $\mathcal{L}'_{RMSE}(\tilde{y}, \hat{y}; p)$ by Eq. 8
  $\mathcal{L}_{SSDNN} = \alpha\mathcal{L}_{RMSE}(y, \hat{y}) + \beta\mathcal{L}'_{RMSE}(\tilde{y}, \hat{y}; p)$
  Update parameters γ by computing the gradient of $\mathcal{L}_{SSDNN}$
end while
return $\mathcal{F}_\gamma$

framework, both observed and unobserved data are used during the training regression network. Specifically, the proposed method trains a regression model $\mathscr{F}$ on observed data via conventional $\mathscr{L}_{RMSE}$ loss function as follows:

$$\mathscr{L}_{RMSE}(y_i\widehat{y}_i) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{5}$$

where $\widehat{y}_i = \mathscr{F}_\gamma(\mathbf{x_i})$ denotes the output prediction SSC value from the network $\mathscr{F}$ and γ is the set of the trainable parameters of the network $\mathscr{F}$. To estimate the SSC values on unobserved data, this method utilizes the classification model $\mathscr{G}$ that is pre-trained in the previous Step 1. The model $\mathscr{G}$ is first used to predict the class index and softmax scores on unobserved data in Eq. (6) as follows:

$$p_j = \mathscr{G}_\theta(\mathbf{z_j}) \tag{6}$$

In Eq. (3), the predicted class $\hat{q}_j$ is calculated by $argmax\left(p_j\right)$. The proposed method then utilizes these outputs to generate the pseudo-SSC value for unobserved data in $\mathscr{D}_u$. The pseudo-SSC value is calculated as follows:

$$\tilde{y}_j = \frac{\left(\hat{q}_j \times V\right) + \left(\left(\hat{q}_j + 1\right) \times V\right)}{2} \tag{7}$$

where $\tilde{y}_j$ illustrates the pseudo SSC value generated from pre-trained classification model $\mathscr{G}$ with input as $\mathbf{z_j}$. Note that the range of $\hat{q}_j$ values is from 0 to $K-1$. The upper bound and the lower bound of SSC values are determined by components $(\hat{q}_j \times V)$ and $(\hat{q}_j + 1) \times V$. The average value between two points (the most negligible bias probably point) is suitable for representing the SSC number for this duration. Let denote $\hat{y}_j$ as the predictive SSC value from the regression model (i.e., $\hat{y}_j = \mathscr{F}_\gamma(\mathbf{z_j})$). The pseudo SSC value will be utilized to guide the training of the network $\mathscr{F}$ by minimizing the loss function in Eq. (8)

$$\mathscr{L}'_{RMSE}\left(\tilde{y}_j, \hat{y}_j; p_j\right) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(max\left(p_j\right) \geq c\right) \times \left(\tilde{y}_j - \hat{y}_j\right)^2} \tag{8}$$

where $\mathbf{1}$ equals to 1 if the unobserved data z_j has the pseudo SSC value $\tilde{y}_j$ with $max\left(p_j\right) \geq c$. Specifically, in Eq. (8), the proposed SSL approach only retains the pseudo-SSC value $\tilde{y}_j$ with $max\left(p_j\right) \geq c$ where c is a hyper-parameter denoting the confidence factor, and the low-confidence predictions ($max\left(p_j\right) < c$) will be ignored. Combining two loss functions Eq. (5) and Eq. (8), the final objective function in this proposed approach (semi-supervised-based deep neural networks (SSDNN)) is defined as follows:

$$\mathscr{L}_{SSDNN} = \alpha \mathscr{L}_{RMSE} + \beta \mathscr{L}'_{RMSE} \tag{9}$$

The pseudo-code of training the regression model is illustrated in Algorithm 2.

4. Results and discussion

4.1. Experimental settings

In this study, 85 % and 15 % of the data are used for training and testing, respectively. The training and testing data set are selected to represent all kinds of flow events, ranging from flood events (e.g., 2001 and 2011) and dry events (e.g., 1997 and 2010), and the period of dam impacts (2012–2019) is included in the training and testing phase. Such a consideration guarantees the balance between the two data sets regarding extreme values. The construction of mega dams (e.g., Manwan (1993), Xiaowan (2010), Nuozhadu (2014)) may change the nature of sediment transport (i.e., dams trap sediment in the reservoirs) (WLE, 2020), thus affecting the learning ability of the machine learning models if not well trained. Therefore, the SSC data before and after the construction of dams need to be included in the training data set to provide plausible ranges of data values for the models to learn.

4.1.1. Pre-processing

The correlation coefficient method is applied to find independent variables (i.e., P, Q) that have significant correlations with the dependent variable (i.e., SSC). The temporal correlation for time series with different time lags is determined and suggested as dominant inputs for ML models. The model inputs for discharge were similar at Chiang Saen, Nong Khai, and Mukdahan, at time t, $t-1$, $t-2$, and $t-3$, while those for precipitation were at time $t-4$, $t-5$, $t-6$, $t-7$ at Chiang Saen and Nong Khai, and $t-1$, $t-2$ at Mukdahan. In this case, t is the selected time, and $t-1$, …, $t-7$ are 1-day, …, 7-day lagged times. Notably, the fact that the precipitation at Mukdahan has a different time lag can be explained based on the orography and geographical distance. The terrain in Mukdahan consists of a combination of plains and small hills. Due to the flat terrain, Mukdahan tends to receive more concentrated rainfall compared to Chiang Saen and Nong Khai, which have higher mountainous terrain.

4.1.2. Normalization

All input variables are rescaled into a range of [0,1] using the Min-max scaler before feeding them to the ML models to avoid range value differences (Singh and Singh, 2020), as shown in Eq. (10).

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{10}$$

where x is the feature that should be normalized, and $max(x)$ and $min(x)$ are the maximum and minimum values of the observed data, respectively.

4.1.3. Evaluation criteria

The models' performances were evaluated by the Nash–Sutcliffe efficiency coefficient (NSE) (Gupta et al., 2009) and the traditional root mean square error (RMSE) measure. NSE values range from $-\infty$ to 1, where 1 indicates a perfect match between the results from models and the observed data, and smaller NSE shows less association. In contrast, a lower RMSE value presents better models' performance. The NSE is calculated as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

where y_i is the observed data, $\hat{y}_i$ is the predicted data from the models and $\bar{y}$ is the mean of the observed data.

4.1.4. Parameter optimizations

For MLP, CNN, and LSTM, all of these models are optimized by Adam optimizer with an initial learning rate of 0.001, β_1 of 0.9, and β_2 of 0.99. The weight decay is set to 1×10^{-4}. The mini-batch of 256 samples was used and training is done in 200 epochs. The learning rate is adjusted by the cosine scheduler with a warmup of 10 epochs. SVM for regression (C = 1.0, epsilon = 0.2), Random Forest (max_depth = 2, random_state = 0), and XGBoost (n_estimators = 1000, max_depth = 7, eta = 0.1, subsample = 0.7, colsample_bytree = 0.8). Other hyperparameters are set by default values. The network architectures for these models can be found in the appendices.

4.2. Performance evaluation

4.2.1. Performance comparisons

Table 1 shows performances of SSC-Q rating curves, traditional ML techniques, state-of-the-art semi-supervised methods, and the proposed semi-supervised model (with different ML models as the classifier and regressor) based on the test data for all three stations Chiang Saen, Nong Khai, and Mukdahan in the MRB.

The SSC-Q rating curve has been successfully used in the MRB (Wang et al., 2011; Lu et al., 2014; Darby et al., 2016; Binh et al., 2020a) to estimate the SSC. Despite its simplicity, SSC-Q rating curves acquire better performance than LR, SVR, and RF in major cases. For example, it outperforms LR and SVR by both RMSE and NSE at the Chiang Saen, Nong Khai, and Mukdahan stations. It also performs better than RF at the Nong Khai and Mukdahan stations. Barberena et al. (2023) estimated the SSC from turbidity, and indicated that more variables should be included to estimate the SSC. According to the authors, this conclusion is specifically applicable to small and medium basins which should be different from a large basin like the MRB. The SSC in the Vietnamese

Table 1
Performances of different approaches for sediment reconstruction at all stations. SSMLP, SSCNN and SSLSTM are the results of the proposed SSL frameworks using MLP, CNN and LSTM as the main classifier/regressor, respectively. STD denotes for standard deviation. The best results are shown in **bold.**

Models/method	Chiang Saen		Nong Khai		Mukdahan	
	RMSE (STD)	NSE (STD)	RMSE (STD)	NSE (STD)	RMSE (STD)	NSE (STD)
SSC-Q rating curves	592.16 (0)	0.3 (0)	265.78 (0)	0.36 (0)	213 (0)	0.46 (0)
Linear Regression (LR)	611.97 (0)	0.12 (0)	214.97 (0)	0.19 (0)	246.63 (0)	0.41 (0)
SVR (SVM)	782.38 (0)	−0.44 (0)	203.53 (0)	0.27 (0)	309.07 (0)	0.08 (0)
Random Forest (RF)	443.66 (2.1)	0.54 (0.005)	252.84 (1.62)	−0.12 (0.013)	258.84 (0.74)	0.36 (0.005)
XGBoost	420.80 (16.21)	0.58 (0.033)	256.81 (7.71)	−0.16 (0.071)	236.04 (9.39)	0.46 (0.044)
CatBoost	396.73 (4.66)	0.63 (0.01)	257.85 (2.43)	−0.17 (0.023)	225.12 (1.97)	0.51 (0.01)
MLP	400.91 (10.23)	0.60 (0.019)	181.50 (6.62)	0.51 (0.037)	201.66 (13.44)	0.61 (0.054)
CNN	406.36 (6.62)	0.59 (0.014)	181.98 (9.21)	0.51 (0.049)	201.13 (6.94)	0.61 (0.027)
LSTM	490.62 (2.98)	0.40 (0.008)	174.27 (9.05)	0.54 (0.045)	207.91 (2.69)	0.57 (0.001)
CoReg	475.43 (13.61)	0.47 (0.028)	261.05 (5.69)	−0.20 (0.052)	250.04 (2.28)	0.40 (0.013)
⊓ Model	399.32 (5.19)	0.62 (0.008)	232.32 (2.2)	0.05 (0.017)	227.36 (2.2)	0.50 (0.008)
ICT	399.81 (3.27)	0.62 (0.008)	232.15 (2.25)	0.05 (0.019)	227.45 (1.69)	0.50 (0.008)
Mean Teacher	398.82 (3.85)	0.62 (0.008)	231.96 (2.01)	0.05 (0.018)	227.76 (1.65)	0.50 (0.007)
SSMLP	**378.96** (12.23)	**0.64** (0.022)	**181.08** (13.33)	**0.52** (0.073)	**201.28** (13.74)	**0.61** (0.057)
SSCNN	406.15 (21.55)	0.59 (0.04)	184.19 (17)	0.49 (0.098)	196.56 (2.43)	0.62 (0.01)
SSLSTM	423.29 (12.37)	0.55 (0.027)	193.55 (9.48)	0.44 (0.056)	210.78 (4.07)	0.57 (0.016)

Mekong Delta can be estimated using the turbidity-SSC rating curve with very high coefficients of determination of up to 0.9945 (Binh, 2019), thus this method could be applied to predict the SSC in the MRB. The extreme data sparsity problem in these stations might be a reason. With too little data to train the models like in this case, ML models tend to be overfitted to train data and thus reduce their overall forecasting performance. XGBoost, with its ability to handle data sparsity, acquires significantly better performance than SSC-Q rating curves at Chiang Saen station but still fails to overcome SSC-Q rating curves at the other two stations. On the other hand, CatBoost dominates XGBoost and RF, and it also performs better than LR, SVR, and SSC-Q rating curves at two over three stations including Chiang Saen and Mukdahan.

Deep learning-based methods including MLP, CNN, and LSTM outperform SSC-Q rating curves at all stations. However, CatBoost still has better performances than MLP, CNN, and LSTM at Chiang Saen. Though MLP and CNN has better accuracy than LSTM in major cases, those are only slightly better than CatBoost with comparable performances at Chiang Saen and Mukdahan while being better at Nong Khai.

Though there is no existing semi-supervised approach for sediment reconstruction before, some state-of-the-art semi-supervised regression techniques are also employed as baselines for assessing the performance of the proposed method including CoReg, ⊓ Model, ICT, and Mean Teacher. As shown in Table 1, though these methods with the exception of CoReg have better results than SSC-Q, they are not better than CatBoost, MLP, CNN, or LSTM overall.

With its ability to exploit unobserved data using combined classification and regression models, the proposed semi-supervised approach helps to boost the performances of all its employed ML methods in most of cases. For instance, the NSE increases from 0.4 to 0.55 at Chiang Saen station for LSTM, while RMSE decreases from 201.13 mg/l to 196.56

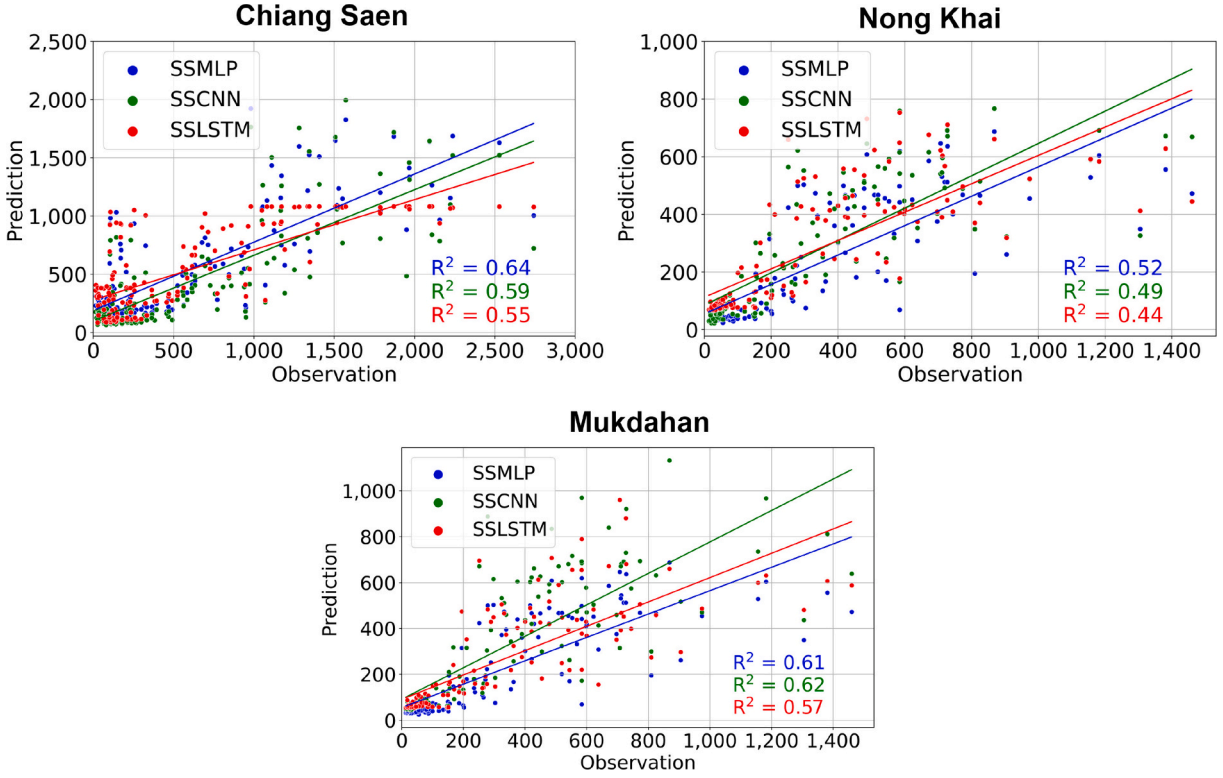


Fig. 4. Performance of the proposed semi-supervised approach with respect to different core ML algorithms at all stations.

mg/l at Mukdahan station for CNN. When using MLP as its core component, that semi-supervised approach, denoted as SSMLP, outperforms all other methods (including MLP itself) at all stations. Particularly, its RMSE are 378.96 mg/l, 181.08 mg/l, and 201.28 mg/l at Chiang Saen, Nong Khai, and Mukdahan, respectively, while its NSE values ranged from 0.52 to 0.64 at three main stations.

Fig. 4 illustrates the performance of the proposed SSL framework with respect to different core ML algorithms including MLP, CNN, and LSTM at all three stations using scatter plots between observed and predicted values. The results are consistent with Table 1, SSMLP has the best performance in all stations.

For the rest of this paper, the performances of the proposed semi-supervised approach using MLP as a core ML model, denoted as SSMLP, will be further studied unless otherwise stated.

4.2.2. Performance of SSMLP in dry and flood seasons

The performances of SSMLP, the best-performing method, in the dry and wet seasons at three main stations are illustrated in Table 2. The SSMLP obtains better results in the dry seasons than those in the flood seasons. For example, the NSE in the dry seasons are 0.7, 0.71, and 0.67 at Chiang Saen, Nong Khai, and Mukdahan, respectively, while those range from 0.35 to 0.48 at all main stations in the flood seasons. It can be explained that the heavy rains during the flood seasons significantly increase the sediment loads, thus making it harder to predict due to high fluctuations of values. The model performed better in the dry months because most of the SSC values in the time series used in the model were low and medium, which occurred mostly in the transitional and dry months. For instance, the SSC values of 0–1000 mg/l (14.24 % - 83.68 %) dominate other ranges as shown in Fig. 7. Thus, the proposed model has more data of low and medium values to learn and produces more reliable results. Although not being examined, if we use the data only in the dry months to estimate the SSC, the results should be worse because the number of data for the ML models to learn is very limited. In that case, the model may not get enough information to learn to produce reliable results. Thanh et al. (2022) found that by establishing the ML models for the dry and flood seasons separately, the estimation of the discharge at a hydrological station in the Mekong basin became worse compared to using all the year-round data. They explained that such a result happened because the number of data for the ML models to learn became half for which the ML models could not leverage sufficient information to learn.

4.2.3. Performance of SSMLP in dry and flood years

Fig. 5 illustrates the reconstructed daily SSC values produced by SSMLP in comparison with observed data in 1997 (left) and 2001 (right) at Nong Khai station. Years 1997 and 2001 were the extreme drought and flood years and coincided with the occurrences of strong El Niño and La Niña that resulted in severe economic losses and reductions in agricultural productions (Cosslett and Cosslett, 2018). The reconstructed SSC in the dry season (c.f., Fig. 5a) shows a good fit between observed data and prediction one than that in the flood season (Fig. 5b). Particularly, the results from SSMLP almost cover the peak values in the flood months and low values in the dry months in 1997, while the predicted results are underestimated in both dry and wet seasons in 2001. The results indicate that SSMLP produces better results to predict SSC in the dry months than those in the flood months. The same reason can be explained by the above study in the wet and flood seasons.

Table 2
The performances of SSMLP at mainstream stations in the dry and flood seasons.

Stations	Dry season		Flood season	
	RMSE	NSE	RMSE	NSE
Chiang Saen	179.55	0.70	484.37	0.48
Nong Khai	80.17	0.71	207.83	0.36
Mukdahan	78.42	0.67	254.87	0.35

4.3. Ablation studies

4.3.1. Effects of the confidence factor c

Fig. 6a shows the performances of SSMLP compared to different values of the confidence factor c (c.f. Section 3). When c increases from 0.1 to 0.9, the performance of SSMLP increases until it reaches a peak at 0.5 and then starts to decrease. Too low confidence values can lead to many uncertain data being placed in the training data, thus lowering the overall performance. On the other hand, too high confidence values would reduce the number of new data to be put into the training set. Hence, this does not help to reduce the overfitting problem to improve the result. A default value of $c = 0.5$ is suggested.

4.3.2. Effects of the number of classes K

Fig. 6b shows the effects of the parameter K on the performance of this proposed method. When K is too small, the classification range is too big, which leads to the loss of important information and lowers the performance of the proposed algorithm. On the other hand, when K is too large, it reduces the performance of the classification model (Model 1) due to many class labels, thus reducing the performance of the whole algorithm consequently. The model performances with different K values are not significant different. For all three stations, the best values are acquired when K is between 30 and 70 for RMSE and NSE. For all three stations, the best values are acquired when $K = 50$ for RMSE and NSE.

4.3.3. Effects of different training data on SSMLP

For all ML models, having good training data without noisy/abnormal samples would help to improve the learning process and the overall prediction performance. Hence, in this part, the performance of SSMLP with respect to different training data is evaluated by adding/removing some years with extreme events to/from the original training data. Overall, four cases are studied as follow:

- **Case 1**: excluding extreme flood years (2000 and 2011) (Cosslett and Cosslett, 2018)
- **Case 2**: excluding extreme drought years (1997 and 2006) (Cosslett and Cosslett, 2018)
- **Case 3**: excluding years of dam impacts (2010–2019) (Binh et al., 2020a; Lu and Chua, 2021)
- **Case 4**: including extreme flood years, extreme drought years, and years of high-dam impacts.

The results, shown in Fig. 6c, indicate that the performances of SSMLP are improved significantly with the contribution of all kinds of flow events, ranging from flood years to drought years and the period of high-dam impacts in the training phase (Case 4). For Cases 1 and 2, without extreme flood and drought years in the training data, the performance of SSMLP is not significant different at Chiang Saen and Mukdahan, but it performs worse at Nong Khai. For example, the NSE values is 0.45 and 0.44 at Nong Khai for Case 1 and Case 2, respectively. For Case 3, the influences of dams are clearly observed after the completion of mega dams (e.g., Manwan (1993), Xiaowan (2010), Nuozhadu (2014)). Hence, the period of high dam impacts was not taken into account in the training data set, from which the performance of SSMLP is slightly better compared to Case 1 and Case 2. For Case 4, the training data set includes extreme flood and drought events and years of dam impacts, resulting in the best results. Particularly, NSE values are 0.64, 0.52, and 0.61 at Chiang Saen, Nong Khai, and Mukdahan, respectively. Overall, the SSMLP produces the best model performance when the model is trained with all extreme cases (including drought, flood, and dam impact years).

4.3.4. Should each station to be trained separately?

Fig. 6d shows the performances of SSMLP in two different training scenarios. In the first scenario (Case A), SSMLP is trained separately for

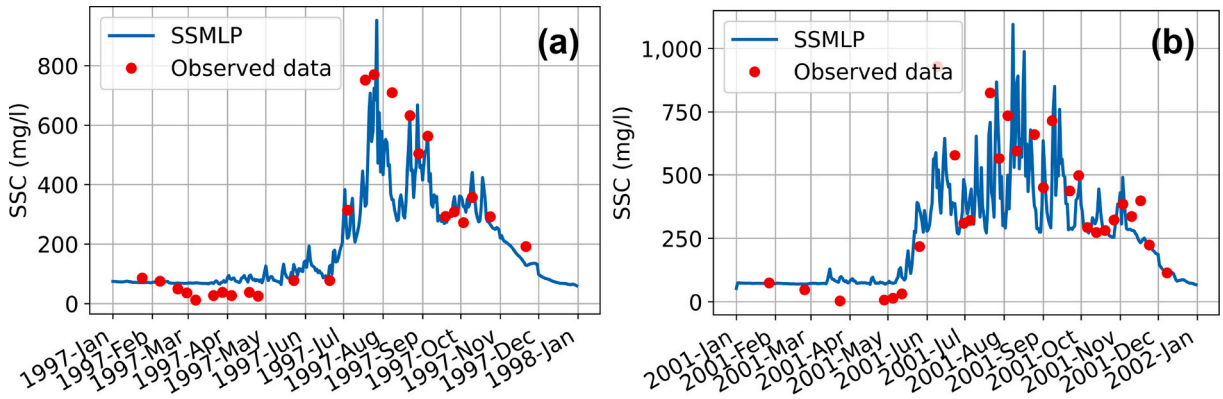


Fig. 5. Reconstructed daily SSC values by SSMLP with respect to observed data at Nong Khai station in (a) dry year (1997) and (b) flood year (2001).

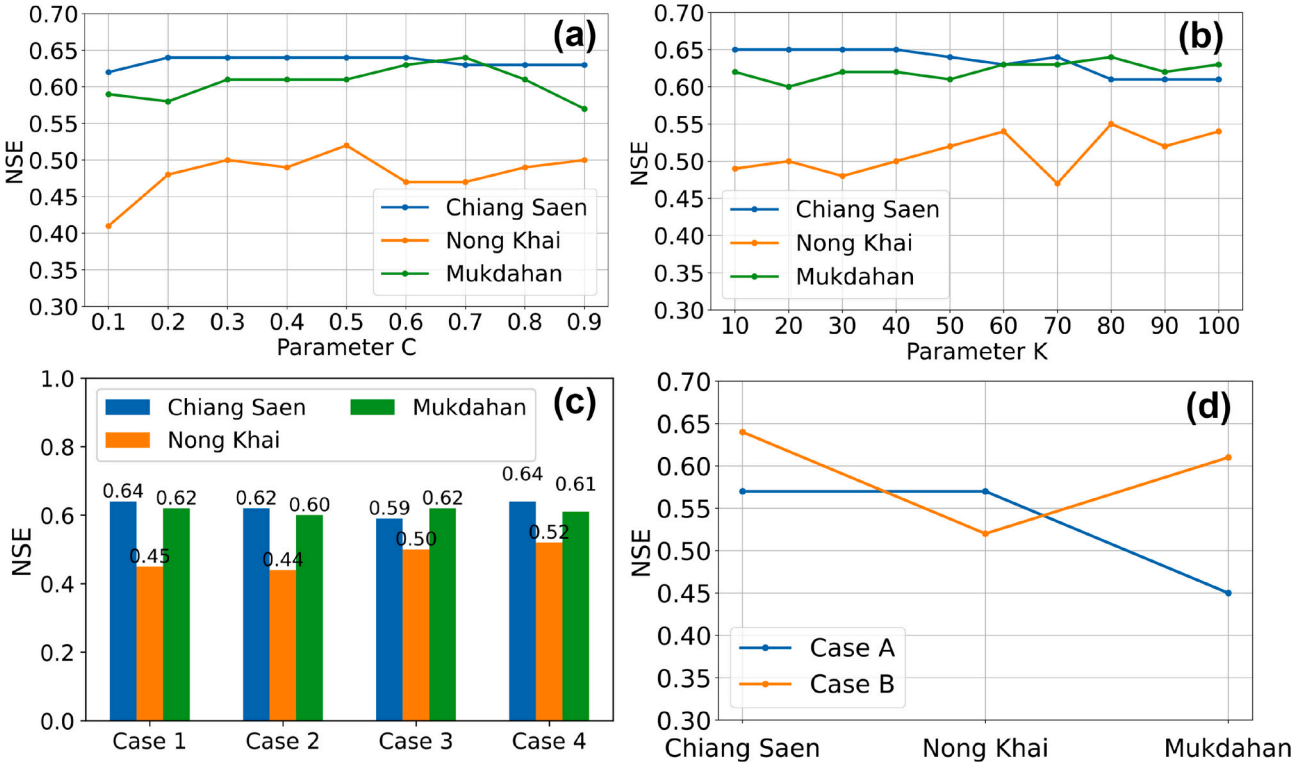


Fig. 6. Ablation studies: (A) Effects of the confidence factor *c*; (B) Effects of the number of classes *K*; (C) Effects of training data on the performances of SSMLP; and (D) The performances of SSMLP when being trained for each station separately.

three stations. NSE values are 0.57, 0.57, and 0.45 at Chiang Saen, Nong Khai, and Mukdahan. In the second scenario (Case B), SSMLP is trained using data from all stations. As can be seen, the acquired results are better than those in the first scenario. Particularly, the NSE values range between 0.52 and 0.64 at Chiang Saen, Nong Khai, and Mukdahan, respectively. Overall, by aggregating data from all stations to train SSMLP, the proposed approach can enrich the training data and reduce overfitting, thus enhancing its performance.

5. Conclusion

Reconstructing missing SSC data plays an important role in the MRB to have long-term reliable SSC data for further research, such as assessing the impacts of climate change and human activities on sediment load in global river basins. However, it is a non-trivial task due to severe sediment data sparsity. The proposed semi-supervised DL framework provides a unique way to cope with this challenge by exploiting existing climate data to enrich the training process, thus enhancing the overall prediction accuracy. In addition, the performances of many different supervised ML methods and especially existing semi-supervised learning techniques, which have not been employed for sediment reconstruction before, are thoroughly studied. This will provide a comprehensive view on the performances of different approaches and play an important role for researchers to select suitable models for their works in other locations. Extensive experiments conducted on data collected from 1979 to 2019 at three main stations in the Mekong River including Chiang Saen, Nong Khai and Mukdahan show that CatBoost and MLP acquire better accuracies than the SSC-Q rating curves and other state-of-the-art ML models like Random Forest, XGBoost, CNN, and LSTM. More importantly, the proposed semi-supervised framework can dramatically boost the performance of all employed ML models at all stations. Compared to state-of-the-art semi-supervised methods like CoReg, ⊓ Model, ICT, or Mean Teacher, the proposed SSL approach also acquires much better reconstruction accuracy. Deep studies show that

the proposed SSL framework can predict SSC during the dry periods better than during the flood ones. Moreover, the more diverse data are included in the training process (i.e., dry years, flood years, and dam impact years), the better the acquired performance.

CRediT authorship contribution statement

Thi Thu Ha Nguyen: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Duc Quang Vu:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Ngoc Phu Doan:** Writing – review & editing, Software. **Huynh Thi Khanh Chi:** Writing – review & editing, Software. **Peixin Li:** Writing – review & editing, Software. **Doan Van Binh:** Writing – review & editing. **Yimeng An:** Writing – review & editing. **Pham Tuan Dung:** Writing – review & editing, Software. **Tuan A. Hoang:** Writing – review & editing. **Mai Thai Son:** Writing – review & editing, Writing – original draft, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We special thank anonymous reviewers for helpful comments to improve this paper. This project is partly supported by QUB Agility+ program, DfE Project (R3648ECS) and EU Horizon\UKRI funded project RELAX (R1142ECS). Doan Van Binh received financial support from the collaborative research project (2023IG-01) of the Disaster Prevention Research Institute of Kyoto University and the Asia-Pacific Network for Global Change Research under project reference number CRRP2023-04MY-Doan Van.

Appendix A

Networks Architectures. The MLP network architecture is shown in Table 3. In which, FC, BN, and ReLU denote Fully Connected, Batch Normalization, and ReLU activation layers, respectively. The unit means the number of units in each FC layer. The CNN network architecture is shown in Table 4 where Conv denotes the convolution layer with f as the number of filters and k as kernel size. Table 5 presents the architecture of the LSTM network, in which, unit in each LSTM layer denotes the number of LSTM cells.

Table 3
The MLP network architecture.

Layer	Specification	Output size
Input		10
FC	$unit = 8192$	8192
BN, ReLU		8192
Dropout	$droprate = 0.5$	8192
FC	$unit = 4096$	4096
BN, ReLU		4096
Dropout	$droprate = 0.5$	4096
FC	$unit = 2048$	2048
BN, ReLU		2048
Dropout	$droprate = 0.5$	2048
ReLU		2048
FC	$unit = 1$	1

Table 4
The CNN network architecture.

Layer	Specification	Output size
Input		10
Conv	$f = 256, k = 3$	256
BN, ReLU		256
Conv	$f = 512, k = 3$	512
BN, ReLU		512
Conv	$f = 1024, k = 3$	1024
BN, ReLU		1024
Flatten		1024
Dropout	$droprate = 0.5$	1024
FC	$unit = 1$	1

Table 5
The LSTM network architecture.

Layer	Specification	Output size
Input		10
LSTM	$unit = 512, layers = 3, bidirectional = True$	512
LSTM	$unit = 512, layers = 3, bidirectional = True$	512

(*continued on next page*)

Table 5 (*continued*)

Layer	Specification	Output size
LSTM	$unit = 1024, layers = 3, bidirectional = True$	1024
Flatten		1024
Dropout	$droprate = 0.5$	1024
FC	$unit = 1$	1

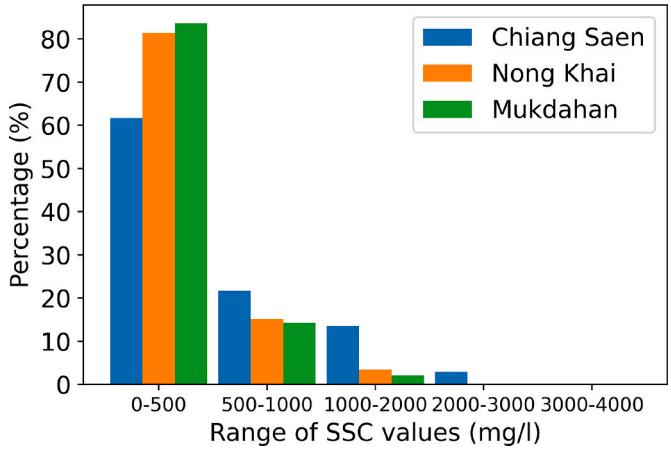


Fig. 7. Data distribution of SSC values across different gauging stations.

Extensive experiments are conducted to evaluate the model's performance with different test sizes. The test years include all extreme flow events (e.g., flood and drought years), and years of dams' impact. In Tables 6 and 7, the results of MLP and SSMLP are compared with test size ranging from 5 % to 25 %. Results are highlighted in bold indicate SSMLP outperforms MLP, and vice versa. As can be seen, SSMLP helps to boost the performance of its barebone MLP in most cases (24/30 cases). That proves the effectiveness of the proposed semi-supervised learning framework.

Table 6
Model performance of MLP with different test sizes.

Test sizes	Chiang Saen		Nong Khai		Mukdahan	
	RMSE	NSE	RMSE	NSE	RMSE	NSE
5 %	168.59	−4	97.72	−5.64	**187.02**	**0.31**
10 %	392.24	0.64	**179.44**	**0.52**	211.11	0.57
15 %	400.91	0.60	181.50	0.51	201.66	0.61
20 %	399.14	0.57	197.65	0.32	**210.42**	**0.49**
25 %	423.26	0.44	197.22	0.18	226.98	0.35

Results are highlighted in bold indicate MLP outperforms SSMLP.

Table 7
Model performance of SSMLP with different test sizes.

Test sizes	Chiang Saen		Nong Khai		Mukdahan	
	RMSE	NSE	RMSE	NSE	RMSE	NSE
5 %	**151.17**	**−3.07**	**77.65**	**−3.00**	212.38	0.13
10 %	**387.45**	**0.65**	189.13	0.47	**200.33**	**0.61**
15 %	**378.96**	**0.64**	**181.08**	**0.52**	**201.28**	0.61
20 %	**375.16**	**0.62**	**192.06**	**0.36**	218.65	0.45
25 %	**421.78**	**0.45**	**194.69**	**0.20**	**225.65**	0.35

Results are highlighted in bold indicate SSMLP outperforms MLP.

Data availability

Daily precipitation and temperature were retrieved from CPC datasets, available at https://www.cpc.ncep.noaa.gov. The daily discharge and suspended sediment concentration at mainstream stations were obtained from the Mekong River Commission (https://portal.mrcmekong.org/home). All these data are publicly available and were accessed in December 2022.

References

Asselman, N.E., 1999. Suspended sediment dynamics in a large drainage basin: the river Rhine. Hydrol. Process. 13, 1437–1450.
Barberena, I., Luquin, E., Campo-Bescós, M.Á., Eslava, J., Giménez, R., Casal, J., 2023. Challenges and progresses in the detailed estimation of sediment export in agricultural watersheds in Navarra (Spain) after two decades of experience. Environ. Res. 234, 116581.
Binh, D.V., 2019. Impacts of Upstream Dam Development on Flow, Sediment and Morphological Changes in Vietnamese Mekong Delta (Ph.D. thesis).
Binh, D.V., Kantoush, S., Sumi, T., 2020a. Changes to long-term discharge and sediment loads in the vietnamese mekong delta caused by upstream dams. Geomorphology 353, 107011.

Binh, D.V., Kantoush, S.A., Saber, M., Mai, N.P., Maskey, S., Phong, D.T., Sumi, T., 2020b. Long-term alterations of flow regimes of the mekong river and adaptation strategies for the vietnamese mekong delta. *Journal of hydrology*. Reg. Stud. 32, 100742.
Binh, D.V., Kantoush, S.A., Ata, R., Tassi, P., Nguyen, T.V., Lepesqueur, J., Abderrezzak, K.E.K., Bourban, S.E., Nguyen, Q.H., Phuong, D.N.L., et al., 2022. Hydrodynamics, sediment transport, and morphodynamics in the vietnamese mekong delta: field study and numerical modelling. Geomorphology 413, 108368.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
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.
Chua, S.D.X., Lu, X.X., 2022. Sediment load crisis in the mekong river basin: severe reductions over the decades. Geomorphology 419, 108484.
Cosslett, T.L., Cosslett, P.D., 2018. Sustainable Development of Rice and Water Resources in Mainland Southeast Asia and Mekong River Basin. Springer.
Dang, T.D., Chowdhury, A., Galelli, S., 2020. On the representation of water reservoir storage and operations in large-scale hydrological models: implications on model parameterization and climate change impact assessments. Hydrol. Earth Syst. Sci. 24, 397–416.
Darabi, H., Mohamadi, S., Karimidastenaei, Z., Kisi, O., Ehteram, M., ELShafie, A., Torabi Haghighi, A., 2021. Prediction of daily suspended sediment load (ssl) using new optimization algorithms and soft computing models. Soft. Comput. 25, 7609–7626.
Darby, S.E., Hackney, C.R., Leyland, J., Kummu, M., Lauri, H., Parsons, D.R., Best, J.L., Nicholas, A.P., Aalto, R., 2016. Fluvial sediment supply to a mega-delta reduced by shifting tropical-cyclone activity. Nature 539, 276–279.
De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134, 19–67.
De Vente, J., Poesen, J., Arabkhedri, M., Verstraeten, G., 2007. The sediment delivery problem revisited. Prog. Phys. Geogr. 31, 155–178.
Essam, Y., Huang, Y.F., Birima, A.H., Ahmed, A.N., El-Shafie, A., 2022. Predicting suspended sediment load in peninsular Malaysia using support vector machine and deep learning algorithms. Sci. Rep. 12, 1–29.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al., 2018. Recent advances in convolutional neural networks. Pattern Recogn. 77, 354–377.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.
Ho, T.K., 1995. Random decision forests. In: ICDAR, volume 1. IEEE, pp. 278–282.
Hoang, L.P., Lauri, H., Kummu, M., Koponen, J., Van Vliet, M.T., Supit, I., Leemans, R., Kabat, P., Ludwig, F., 2016. Mekong river flow and hydrological extremes under climate change. Hydrol. Earth Syst. Sci. 20, 3027–3041.
Hochreiter, S., 2010. Long short-term memory. Neural Comput. 9, 1735–1780.
Irannezhad, M., Liu, J., 2022. Evaluation of six gauge-based gridded climate products for analyzing long-term historical precipitation patterns across the lancang-mekong river basin. Geogr. Sustainability 3, 85–103.
Kaveh, K., Kaveh, H., Bui, M.D., Rutschmann, P., 2021. Long short-term memory for predicting daily suspended sediment concentration. Eng. Comput. 37, 2013–2027.
Khan, M., Hasan, F., Panwar, S., Chakrapani, G.J., 2016. Neural network model for discharge and water-level prediction for ramganga river catchment of ganga basin, India. Hydrol. Sci. J. 61, 2084–2095.
Khan, M.Y.A., Hasan, F., Tian, F., 2019a. Estimation of suspended sediment load using three neural network algorithms in ramganga river catchment of ganga basin, India. Sustainable Water Resour. Manage. 5, 1115–1131.
Khan, M.Y.A., Tian, F., Hasan, F., Chakrapani, G.J., 2019b. Artificial neural network simulation for prediction of suspended sediment concentration in the river ramganga, ganges basin, India. Int. J. Sediment Res. 34, 95–107.
Kondolf, G.M., Rubin, Z.K., Minear, J., 2014. Dams on the mekong: cumulative sediment starvation. Water Resour. Res. 50, 5158–5169.
Laine, S., Aila, T., 2016. Temporal Ensembling for Semi-Supervised Learning. *arXiv Preprint arXiv:1610.02242*.
Lauri, H., de Moel, H., Ward, P.J., Räsänen, T.A., Keskinen, M., Kummu, M., 2012. Future changes in mekong river hydrology: impact of climate change and reservoir operation on discharge. Hydrol. Earth Syst. Sci. 16, 4603–4619.
Lu, X.X., Chua, S.D.X., 2021. River discharge and water level changes in the mekong river: droughts in an era of mega-dams. Hydrol. Process. 35, e14265.
Lu, X.X., Siew, R., 2006. Water discharge and sediment flux changes over the past decades in the lower mekong river: possible impacts of the chinese dams. Hydrol. Earth Syst. Sci. 10, 181–195.
Lu, X., Kummu, M., Oeurng, C., 2014. Reappraisal of sediment dynamics in the lower mekong river, Cambodia. Earth Surf. Process. Landf. 39, 1855–1865.
Manh, N.V., Dung, N.V., Hung, N.N., Merz, B., Apel, H., 2014. Large-scale suspended sediment transport and sediment deposition in the mekong delta. Hydrol. Earth Syst. Sci. 18, 3033–3053.
Meshram, S.G., Pourghasemi, H.R., Abba, S.I., Alvandi, E., Meshram, C., Khedher, K.M., 2021. A comparative study between dynamic and soft computing models for sediment forecasting. Soft. Comput. 25, 11005–11017.
MRC, 2005. Overview of the Hydrology of the Mekong Basin. Mekong River Commission, Vientiane, p. 82.
MRC, 2007. Consumption and the Yield of Fish and Other Aquatic Animals from the Lower Mekong Basin. Mekong River Commission, Vientiane, p. 82.
Nguyen, T.-T.-H., Li, M.-H., Vu, T.M., Chen, P.-Y., 2023a. Multiple drought indices and their teleconnections with enso in various spatiotemporal scales over the mekong river basin. Sci. Total Environ. 854, 158589.
Nguyen, T.-T.-H., Vu, D.-Q., Mai, S.T., Dang, T.D., 2023b. Streamflow Prediction in the Mekong River Basin Using Deep Neural Networks. IEEE Access.
Piman, T., Shrestha, M., 2017. Case Study on Sediment in the Mekong River Basin: Current State and Future Trends. Stockholm, Sweden, Stockholm Environment Institute.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: unbiased boosting with categorical features. Adv. Neural Inf. Proces. Syst. 31.
Räsänen, T.A., Someth, P., Lauri, H., Koponen, J., Sarkkula, J., Kummu, M., 2017. Observed river discharge changes due to hydropower operations in the upper mekong basin. J. Hydrol. 545, 28–41.
Rubin, Z.K., Kondolf, G.M., Carling, P.A., 2014. Anticipated geomorphic impacts from mekong basin dam construction. International Journal of River Basin Management 13, 105–121.
Sam, T.T., Khoi, D.N., 2022. The responses of river discharge and sediment load to historical land-use/land-cover change in the mekong river basin. Environ. Monit. Assess. 194, 700.
Shadkani, S., Abbaspour, A., Samadianfard, S., Hashemi, S., Mosavi, A., Band, S.S., 2021. Comparative study of multilayer perceptron-stochastic gradient descent and gradient boosted trees for predicting daily suspended sediment load: the case study of the Mississippi river, us. Int. J. Sediment Res. 36, 512–523.
Singh, N., Khan, M.Y.A., 2020. Ann modeling of the complex discharge-sediment concentration relationship in bhagirathi river basin of the himalaya. Sustainable Water Resour. Manage. 6, 1–8.
Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 97, 105524.
Tao, H., Al-Khafaji, Z.S., Qi, C., Zounemat-Kermani, M., Kisi, O., Tiyasha, T., Chau, K.-W., Nourani, V., Melesse, A.M., Elhakeem, M., et al., 2021. Artificial intelligence models for suspended river sediment prediction: state-of-the art, modeling framework appraisal, and proposed future research directions. Eng. Appl. Comput. Fluid Mech. 15, 1585–1612.
Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Proces. Syst. 30.
Thanh, V.Q., Reyns, J., Van, S.P., Anh, D.T., Dang, T.D., Roelvink, D., et al., 2019. Sediment transport and morphodynamical modeling on the estuaries and coastal zone of the vietnamese mekong delta. Cont. Shelf Res. 186, 64–76.
Thanh, H.V., Binh, D.V., Kantoush, S.A., Nourani, V., Saber, M., Lee, K.-K., Sumi, T., 2022. Reconstructing daily discharge in a megadelta using machine learning techniques. Water Resour. Res. 58 e2021WR031048.
Tuan, L.A., Chinvanno, S., 2011. Climate change in the Mekong river delta and key concerns on future climate threats. In: Environmental Change and Agricultural Sustainability in the Mekong Delta, pp. 207–217.
Umar, M., Rhoads, B.L., Greenberg, J.A., 2018. Use of multispectral satellite remote sensing to assess mixing of suspended sediment downstream of large river confluences. J. Hydrol. 556, 325–338.
Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D., 2022. Interpolation consistency training for semi-supervised learning. Neural Netw. 145, 90–106.
Vu, T.M., Raghavan, S.V., Liong, S.-Y., Mishra, A.K., 2018. Uncertainties of gridded precipitation observations in characterizing spatio-temporal drought and wetness over Vietnam. Int. J. Climatol. 38, 2067–2081.
Walling, D., 1977. Limitations of the rating curve technique for estimating suspended sediment loads, with particular reference to british rivers. In: Erosion and Solid Matter Transport in Inland Waters, 122, pp. 34–78.
Wang, J.-J., Lu, X., Kummu, M., 2011. Sediment load estimates and variations in the lower mekong river. River Res. Appl. 27, 33–46.
Warrick, J.A., 2015. Trend analyses with river sediment rating curves. Hydrol. Process. 29, 936–949.
WLE, 2020. Mekong Dam database. In: WLE Greater Mekong, CGIAR Research Program on Water, Land and Ecosystems (WLE), Vientiane, Lao PDR.
Xu, B., He, X., 2022. A physics-informed bayesian storyline approach to assess sediment transport in the mekong. Water Resour. Res. 58, e2022WR032681.
Xue, Z., He, R., Liu, J.P., Warner, J.C., 2012. Modeling transport and deposition of the mekong river sediment. Cont. Shelf Res. 37, 66–78.
Zhou, Z.-H., Li, M., et al., 2005. Semi-supervised regression with co-training. In: IJCAI, volume 5, pp. 908–913.