Key Points:
- Machine learning (ML) methods can reliably reconstruct missing daily discharge values
- The multivariate adaptive regression spline and random forest (RF) models outperform other ML and rating curve (RC) methods
- Data pre-processing reduces the simulation time and effort

Author Contributions:
Conceptualization: Doan Van Binh, Sameh A. Kantoush, Tetsuya Sumi
Data curation: Hung Vo Thanh, Doan Van Binh
Formal analysis: Hung Vo Thanh, Doan Van Binh
Funding acquisition: Sameh A. Kantoush, Tetsuya Sumi
Investigation: Sameh A. Kantoush, Vahid Nourani, Mohamed Saber
Methodology: Hung Vo Thanh, Doan Van Binh, Sameh A. Kantoush, Vahid Nourani, Mohamed Saber
Software: Hung Vo Thanh, Vahid Nourani, Mohamed Saber
Supervision: Sameh A. Kantoush, Tetsuya Sumi
Validation: Hung Vo Thanh, Doan Van Binh
Visualization: Hung Vo Thanh, Kang-Kun Lee

Reconstructing Daily Discharge in a Megadelta Using Machine Learning Techniques

Hung Vo Thanh[1], Doan Van Binh[2,3] , Sameh A. Kantoush[2], Vahid Nourani[4,5], Mohamed Saber[2] , Kang-Kun Lee[1] , and Tetsuya Sumi[2]

[1]School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea, [2]Water Resources Research Center, Disaster Prevention Research Institute, Kyoto University, Kyoto, Japan, [3]Master Program in Water Technology, Reuse and Management, Vietnamese German University, Thu Dau Mot, Vietnam, [4]Center of Excellence in Hydroinformatics, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran, [5]Faculty of Civil and Environmental Engineering, Near East University, Nicosia, Turkey

Abstract In this study, six machine learning (ML) models, namely, random forest (RF), Gaussian process regression (GPR), support vector regression (SVR), decision tree (DT), least squares support vector machine (LSSVM), and multivariate adaptive regression spline (MARS) models, were employed to reconstruct the missing daily-averaged discharge in a mega-delta from 1980 to 2015 using upstream-downstream multi-station data. The performance and accuracy of each ML model were assessed and compared with the stage-discharge rating curves (RCs) using four statistical indicators, Taylor diagrams, violin plots, scatter plots, time-series plots, and heatmaps. Model input selection was performed using mutual information and correlation coefficient methods after three data pre-processing steps: normalization, Fourier series fitting, and first-order differencing. The results showed that the ML models are superior to their RC counterparts, and MARS and RF are the most reliable algorithms, although MARS achieves marginally better performance than RF. Compared to RC, MARS and RF reduced the root mean square error (RMSE) by 135% and 141% and the mean absolute error by 194% and 179%, respectively, using year-round data. However, the performance of MARS and RF developed for the climbing (wet season) and recession (dry season) limbs separately worsened slightly compared to that developed using the year-round data. Specifically, the RMSE of MARS and RF in the falling limb was 856 and 1,040 m^3/s, respectively, while that obtained using the year-round data was 768 and 789 m^3/s, respectively. In this study, the DT model is not recommended, while the GPR and SVR models provide acceptable results.

1. Introduction

River discharge is a crucial indicator to understand terrestrial water cycles and supplies necessary information about water resource management (Adnan et al., 2020). Direct measurement of river discharge, such as employing the acoustic Doppler current profiler, is complicated, costly, time-consuming, and labor-intensive because it requires a number of current sensors and repeated surveys performed by boats and is thus unsafe under unfavorable flow and weather conditions (Gisen & Savenije, 2015; Matte et al., 2018). Other noncontact methods, including large-scale particle image velocimetry (LSPIV) (Akbarpour et al., 2020) and remote sensing (Kebede et al., 2020), have recently begun to be used for discharge measurements. Nevertheless, the use of these methods in contiguous monitoring of river discharge is not feasible; for example, LSPIV cannot measure discharge in large rivers because of limited camera coverage, while satellite images are not always available due to cloud cover, particularly during rainy seasons. As a result, at hydrological stations situated on rivers worldwide, flow discharge is not directly measured; rather, it is indirectly estimated either from the widely used stage-discharge rating curve (RC) method or from cubature, rating-fall, tide-correction, and coaxial graphical-correction methods (Matte et al., 2018), in which the stage (water level) is recorded at specific intervals (e.g., daily, hourly, or sub-daily) depending on the goal of the measurements. Due to technical, financial, maintenance and political instability issues, long-term flow discharge datasets may have gaps, resulting in the loss of information or the misinterpretation of historical flow regime changes and hydrological processes (Tencaliec et al., 2015). Therefore, it is important to reconstruct missing discharge values to reliably provide helpful information for water resource management at the basin scale.

Several methods, including statistical methods, numerical models, and machine learning (ML) algorithms, have been employed to predict river flows. Recently, ML techniques, such as support vector regression (SVR) (Adnan et al., 2020; Luo et al., 2019), random forest (RF), Gaussian process regression (GPR) (Sun et al., 2014), M5

Writing – original draft: Hung Vo
Thanh, Doan Van Binh, Kang-Kun Lee
Writing – review & editing: Hung
Vo Thanh, Doan Van Binh, Sameh A.
Kantoush, Vahid Nourani, Mohamed
Saber, Kang-Kun Lee, Tetsuya Sumi

model tree (Nourani et al., 2019), decision tree (DT) (Choi et al.,2019), least squares support vector machine (LSSVM) (Rezaali et al., 2021), multivariate adaptive regression spline (MARS) (Jeihouni et al., 2020), and adaptive neuro-fuzzy inference system (Hadi & Tombul, 2018a) models, have been increasingly used because they are powerful, robust and efficient algorithms for streamflow prediction given their advantages compared to traditional approaches (Khan et al., 2016; Liu et al., 2020; Mispan et al., 2015).

SVR is easily adaptable for use in multiple engineering disciplines and, in many cases, outperforms other methods, such as artificial neural networks and DTs (Raghavendra & Deka, 2014). Luo et al. (2019) developed 14 ML techniques to predict the monthly discharge of the Jinsha River in Iran, revealing that a hybrid SVR method performed better than a generalized regression neural network (GRNN). In modeling the monthly discharge in the Swat River basin in Pakistan, Adman et al. (2020) found that least squares SVR was superior to other ML models and was recommended for monthly streamflow forecasting without local data. The RF nonparametric algorithm is a type of DT algorithm that includes an ensemble collection of unrelated trees for classification and regression purposes (Breiman, 2001). The advantage of using an RF over a single DT is the reduction in variance achieved by creating several trees, in which each tree is constructed based on a leverage sample of the training database (James et al., 2013). In an attempt to predict the water level in an urban reservoir in Atlanta, Georgia, Obringer and Nateghi (2018) demonstrated that an RF was the most accurate predictive model among the nonparametric ML algorithms considered, and the proposed method is highly transferable to other reservoirs. RF algorithms have been used to reliably predict the outflows of nine reservoirs in California, given reliable input parameters related to precipitation, reservoir inflows, reservoir storage, and downstream conditions (Yang et al., 2016). GPR is a Bayesian learning technique for model approximation, multivariate regression, and experimental design (Rasmussen & Williams, 2006). The power of GPR compared to other ML models is that it simplifies the integration of several ML functions, including hyperparameter evaluation, model training, and uncertainty quantification (Rasmussen & Williams, 2006; Sun et al., 2014). Thus, GPR is relatively uninfluenced by subjectivity, and the results can easily be interpreted (Sun et al., 2014). Zhu et al. (2018) used a GPR model to estimate the streamflow in the Jinsha River; they reported that GPR performed better than a GRNN but was not good at predicting extreme flows. Sun et al. (2014) established a GPR model to simulate monthly streamflow in 438 river basins in the U.S. (MOPEX database); they revealed that the GPR model outperformed regression methods in most basins.

Most recently, advanced ML techniques, including LSSVM and MARS, have received intense attention in hydrological studies. Wang et al. (2020) proposed a new method to predict the evaporation of arid areas in China by applying the MARS method. In a digital application, Jeihouni et al. (2020) employed the MARS model to map soil moisture retention parameters using only satellite data with less prediction uncertainty and high accuracy results. Additionally, Safari (2020) employed MARS and multi non-linear regression (MNLR) to improve the precision of predicting sediment accumulation in open channel flow areas.

Regarding the prospects for the application of the LSSVM model, Rezaali et al. (2021) used this advanced model for highly accurate forecasting of the urban water demand in Qom, Iran. In addition, the LSSVM model was used for water resource management by enhancing the accuracy of the prediction of mid-to long-term streamflow (Zhao et al., 2021). In methane transport modeling, Taherdangkoo et al. (2021) employed the LSSVM model to estimate methane solubility in aquatic environments for a variety of temperatures and pressures. Moreover, the LSSVM model was demonstrated to be effective for forecasting the quality of the air in the Yangtze River Delta of China (Zhou et al., 2020). Because numerous ML models are available, researchers may struggle to determine which ML model is appropriate for a particular problem. Unfortunately, no ML algorithm provides a satisfactory result for all problems involving hydrological processes, and many methods remain in the development stage. Although the SVM, RF, DT, LSSVM, MARS, and GPR models have been widely employed in various research fields (e.g., Jeihouni et al., 2020; Granata et al., 2017; Kisi & Parmar, 2016; Panahi et al., 2020; Rezaali et al., 2021), their application in estimating river discharge has been limited (e.g., Tongal & Booij, 2018; Yaghoubi et al., 2019). Therefore, this study employs these techniques to explore their power/applicability in reconstructing the daily discharge in the Mekong River.

Hydrological data are highly nonstationary (Yarar, 2014), and ML models, or artificial intelligence models in general, have demonstrated limitations in coping with nonstationary phenomena (Nourani et al., 2017). Moreover, hydrological data often contain seasonal effects driven by hydrologic cycles. It is therefore necessary to

perform data pre-processing before applying ML models, and Fourier series fitting can decompose complex original hydrological data into sub-signals with a variety of valuable features to interpret the time series structure and clarify spectral and temporal information (Nourani et al., 2019). Another challenging task in simulating hydrological processes using ML models is model input selection. Too many or too few model inputs may introduce noise, increase model complexity and increase the model run time; both instances can lead to poor model performance (Tran et al., 2015). A traditional approach in model input selection involves the use of a rank-based correlation coefficient, such as the Pearson correlation coefficient, to reflect the linear relations among variables (Zhu et al., 2018). Another more advanced metric is mutual information (MI), which can help reduce the number of model inputs (Nourani et al., 2017). Although MI, as a nonlinear measure used to explain one variable based on another random variable, is useful in reducing simulation effort, its application in the field of hydrology remains limited. This study used both MI and the Pearson correlation coefficient to derive the dominant model inputs after data pre-processing by standardization to remove trends related to the variance and mean; additionally, Fourier series fitting was performed to remove seasonal effects, and first-order differencing was used to convert a nonstationary data set to a stationary data set.

Most previous studies reconstructed/predicted monthly and annual averaged discharge series (Adnan et al., 2020; Hadi & Tombul, 2018b; Khalil et al., 2001; Liu et al., 2020; Sun et al., 2014; Yarar, 2014; Zhu et al., 2016), and studies that have reconstructed daily-averaged series are scarce. This scarcity is likely due to data availability and the complex non-stationarity and nonlinearity of daily averaged data. Notably, the complexity of hydrological data is increased by the effects of tides in the major deltas worldwide, such as the Vietnamese Mekong Delta (VMD), which is the study area considered in this paper. In tidal deltas, flow discharge is seasonally variable, with riverine and marine dominance in the flood and dry seasons, respectively. River tides are largely nonstationary and nonlinear because tides are governed by the effects of hundreds of major and minor astronomical factors (Moftakhari et al., 2013); thus, analyses of flows in tidally affected rivers are complicated by the appearance of a large number of frequencies (Hoitink & Jay, 2016). Spatial acceleration, friction, and discharge gradients also control river-tide interactions, making the direct estimation of fluvial discharge challenging (Hoitink & Jay, 2016). Being nonstationary, water levels in tidally affected rivers are continuously variable during spring-neap tidal cycles, which has led to a consensus that water levels at tidal stations are not the same under different tidal conditions (Hoitink & Jay, 2016). Moreover, tides may increase the water surface gradient and river slope (Jay et al., 2011) to transport more river water during spring tides than during neap tides. Such an increase in the water surface gradient is necessary to enhance the transport capacity of rivers against the increased river friction generated by high discharge amplitudes during spring tides (Buschman et al., 2009).

Moftakhari et al. (2013) proposed a conceptual modeling tool for tidal discharge estimation (TDE) based on sets of governing equations by combining theories of astronomical forcing, tidal constituents, and friction to hindcast the monthly averaged tidal discharges in the San Francisco Bay. Although the estimation was promising, the use of the TDE model is complicated, and many hindcast parameters and extended periods of data observations are required (Gisen & Savenije, 2015); because these are system specific, their application to other tidally affected rivers, particularly in developing countries, where river systems are largely ungauged, is difficult. Gisen and Savenije (2015) developed a semi-empirical approach to compute bankfull discharge in ungauged estuaries by combining hydraulic geometry and hydrodynamic theories. The methodology developed included five main components, namely, estuary geometry, freshwater discharge upscaling, tidal dynamics, regime relations, and estuarine flood number estimation. The derived discharges are estimated with high confidence; however, the application of this method is relatively challenging due to the introduction of several restriction criteria. Moftakhari et al. (2016) developed the multiple-gauge tidal discharge estimate (MTDE) method to estimate the discharge in tidal rivers in North America using tidally observed data at multiple stations near estuaries. The MTDE method can estimate the discharge with a temporal resolution of less than a week; this resolution is finer than that of the TDE method. However, the major shortcoming of the MTDE method is the need for at least three tide gauges, one of which must be near the ocean. This is not applicable in most of the world's tidal rivers because hydrological stations are relatively far from the river mouths (Gisen & Savenije, 2015).

RC has been widely used to reconstruct missing data in deltas, although a special focus must be placed on various tasks, such as establishing RCs for the rising and falling limbs separately (Binh, Kantoush, et al., 2021).

Moftakhari et al. (2015) employed the RC method to reconstruct daily discharge and sediment delivered to San Francisco Bay by dividing the water level data into two subsets (i.e., <6.2 and >6.2 m) according to the effect of flooding. However, the RC method involves several limitations and uncertainties induced by dynamic changes in river geometry and roughness or the effects of backwater and tides (Matte et al., 2018). Uncertainties also arise from the difficulty in measuring the discharge during extreme floods for updating the RCs. Studies reported in the literature have shown that conventional approaches such as the RC and TDE methods have their own limitations and uncertainties under the effects of reversing tidal flow, tidal Stokes drift, spring-neap tidal cycle, lateral circulation, estuarine dynamics, and the occurrence of multiple branches in estuaries (Moftakhari et al., 2016). Therefore, the use of ML models in discharge estimation is expected to overcome the shortcomings of their conventional counterparts. Although ML models have been proven to be an efficient and promising tool, their application for daily-averaged discharge prediction in tidally affected rivers remains uncommon worldwide. In the VMD and Mekong River, no study has used ML algorithms to reconstruct daily-averaged discharge. One of the advantages of using ML models is that they are highly transferable to other river systems with little effort in acquiring a variety of datasets required compared to conceptual, statistical, and numerical models, making them time-saving and cost-effective.

This study developed a robust methodology to reconstruct missing daily-averaged discharge values in a tidally affected river in the VMD using ML techniques. First, the model inputs using multi-station data with respect to upstream-downstream relations were optimized by employing MI and Pearson correlation coefficients after three-step data pre-processing. SVR, GPR, RF, DT, LSSVM, and MARS models were then used and compared to determine the most reliable model. Finally, the best model(s) was further evaluated considering the seasonal patterns of the input data. The purpose of this analysis was to answer the following question: can ML approaches increase the daily-averaged discharge reconstruction accuracy considering seasonal patterns? In this study, we employed ML models using multi-station input data to reconstruct daily-averaged discharges in a tidally affected river. The use of only the water levels at multiple upstream stations as inputs into the ML models has two major advantages. First, the water level is directly, easily, and cheaply monitored in river systems, whereas direct discharge measurement is time-consuming, expensive, and impractical. Second, during extreme events, it is impossible to measure the discharge due to safety concerns (i.e., having to operate a boat in a flooded river), whereas measurements of the water level can be obtained anywhere, at any time, and under all conditions, although river gauges can fail or become compromised (Helaire et al., 2020). Finally, the method developed in this paper can easily be adopted for any river system even though ML models contain black box algorithms.

2. Case Study and Used Data Set

The Mekong River is the eighth largest river globally in terms of the annual discharge of 475 km^3 (Grumbine et al., 2012), and it flows through six countries from the watercourse in China to the ocean in Vietnam. The VMD (Figure 1a) has been formed and propagated over the last 6,000 years (Ta et al., 2002) by water and sediment transported by the Mekong River (Binh, Kantoush, & Sumi, 2020). The flow regime in the VMD is seasonally variable, with two distinct flood and dry seasons (Binh, Kantoush, Saber, et al., 2020). August-October (flood months) is when approximately half of the annual discharge occurs, and approximately 8% occurs in February-April (dry months). The VMD faces many hydrological problems, such as floods, droughts, and salinity intrusion (Eslami et al., 2019; Hoa et al., 2007; Kantoush et al., 2017; Loc et al., 2021; Triet et al., 2017). La Niña and El Niño have caused periodic occurrences of extreme floods (e.g., 1996, 2000, and 2011) and droughts (e.g., 1993, 1998, 2005, 2010, 2015, and 2020), resulting in tremendous damage to the delta. The flow regime in the delta is influenced by tides, with strong tidal effects in the dry season (peak in dry months) and fewer tidal effects in the flood season (Gugliotta et al., 2017). In dry seasons, tidal effects are observable at Phnom Penh, Cambodia, which is approximately 320 km from the river mouth (Gugliotta et al., 2017). The semidiurnal tide in the East Vietnam Sea (Figure 1a) causes the discharge hydrograph to have two peaks and two troughs daily.

Tan Chau and Chau Doc (in the Tien and Hau Rivers, respectively) are the first two major hydrological gauges (tidally affected) at the entrances of the VMD, and the historical data series obtained at these stations are longer
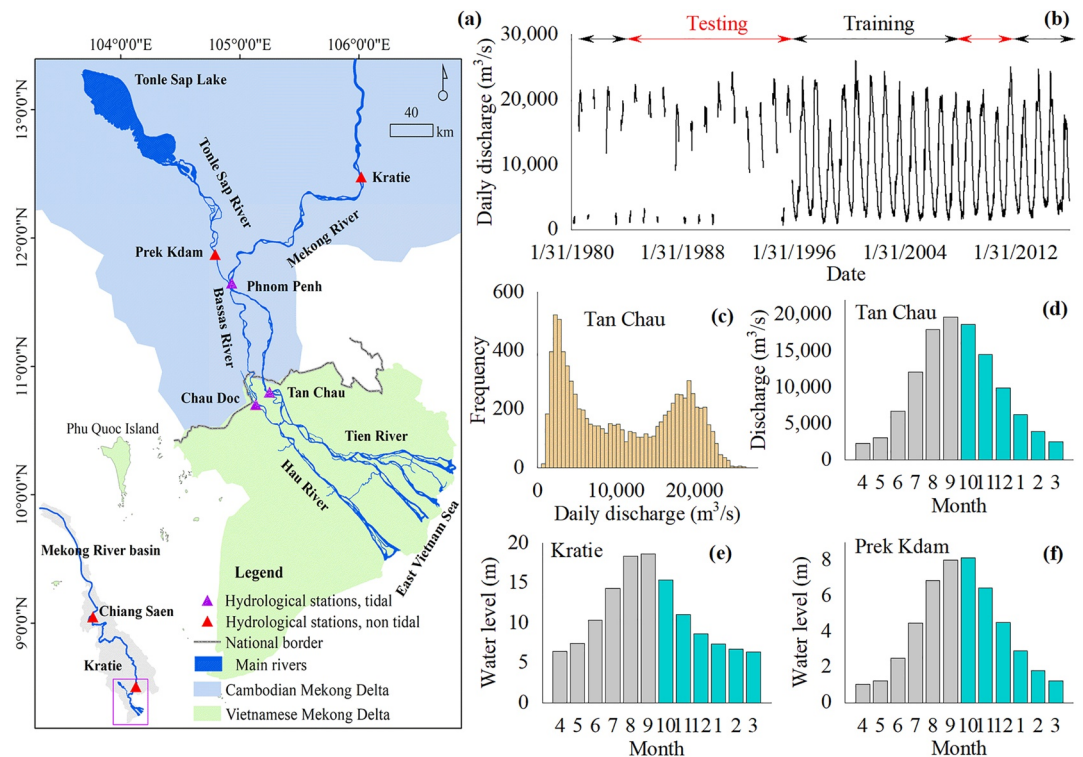
Figure 1. (a) The Mekong River basin and the Vietnamese Mekong Delta (VMD): the main rivers and hydrological stations. The tidally affected and non-tidally affected hydrological stations are distinguished by different colors. (b) Observed daily-averaged discharge values at Tan Chau (1980–2015) with the periods of training and testing in the machine learning (ML) models indicated. (c) Histogram showing the frequency of the daily-averaged discharge at Tan Chau from 1980 to 2015. Mean monthly discharge values at (d) Tan Chau and the water levels at (e) Kratie and (f) Prek Kdam from 1996 to 2015 show similar seasonality for flow regimes. The data in (d–f) were sorted to clearly illustrate the rising (gray bars) and falling limbs (green bars).

than those obtained at newer hydrological gauges. Tan Chau conveys approximately four times more water than Chau Doc (Binh, Kantoush, et al., 2021). Hourly discharge at Tan Chau was persistently monitored from 1996 to 2015, whereas daily-averaged discharge (from 1980 to 1995) was available several months per year (Figure 1b). The hourly discharge from 1996 to 2015 was averaged over a day to create the daily-averaged discharge, which is equivalent to the de-tided discharge (Binh, Kantoush, et al., 2021); however, the tidal effect does not disappear completely (Hoitink & Jay, 2016). A frequency analysis of the observed daily-averaged discharge at Tan Chau from 1980 to 2015 (Figure 1c) shows that the majority of the discharge values are less than 6,500 m³/s (39%) and vary from 16,250 to 22,250 m³/s (31%); only 2% of the daily-averaged discharge values exceed 22,500 m³/s. Given the importance of understanding long-term flow variations when assessing the corresponding causes, consequences, and appropriate actions, it is important to fill the gaps in the historical records. In this study, ML algorithms were used to reconstruct the missing daily-averaged discharge values at Tan Chau and to establish a framework for the other stations in the VMD.

The VMD receives water directly from the Mekong River, and Kratie is a gauging station at the apex of the Mekong Delta (from the Cambodian Mekong Delta) (Figure 1a). Tonle Sap Lake in Cambodia is of utmost importance in naturally regulating the flow in the VMD to the extent that the lake retards flood water and reverses the flow back to the VMD in the dry season (Park et al., 2022; Pokhrel et al., 2018). Tonle Sap Lake is connected with the Mekong River by the Tonle Sap River, and Prek Kdam is an important gauging station that records the exchanged flow regimes (Figure 1a). Figures 1d–1f show that the flow pattern at Tan Chau is physically consistent with those at Kratie and Prek Kdam, with similar rising (April–September) and falling (October–March) limbs.
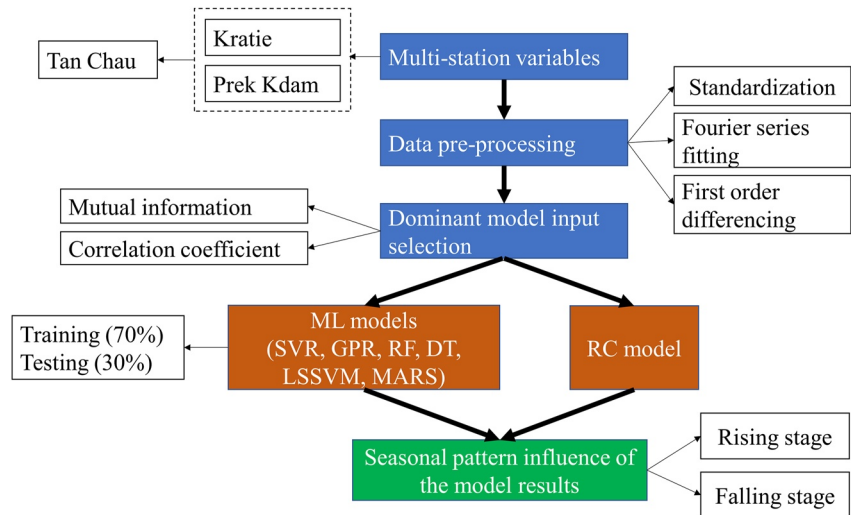
Figure 2. Flowchart of the study showing the research steps.

In reconstructing the daily-averaged discharge at Tan Chau using the ML models, data from multiple upstream stations, that is, the daily-averaged water levels at Kratie and Prek Kdam, were used. The Mekong River Commission provided water level data at Kratie and Prek Kdam from 1980 to 2015 (https://portal.mrcmekong.org/time-series). Using water levels as the input data is practically feasible because these values, rather than discharge values, are directly monitored at all hydrological stations on the Mekong River. Researchers can request such data from the Mekong River Commission. The daily-averaged discharge data at Tan Chau collected from 1980 to 2015 (38.8% of the data are missing) were obtained from the Vietnam National Centre for Hydrometeorological Forecasting. The outcomes from the ML models were compared to those from the conventional RC to assess the applicability of ML in the VMD and tidally affected river systems in general.

3. Proposed Method and Materials Used

3.1. Data Pre-Processing and Model Input Selection

Figure 2 shows the methodology proposed in this study. Six ML models were built considering the input data at multiple upstream stations (i.e., water levels at Kratie and Prek Kdam) to assess the trends at individual stations and the combined contributions to the results. In the ML models, the tidal effect is implicitly considered in the target model (output) because the input water levels are not affected by tides. The results from the ML models were compared with those obtained from linear stage-discharge RCs at the station examined (Tan Chau). These RCs were separately established based on both year-round data and data for the rising and falling limbs, following the work of Binh, Kantoush, et al. (2021). One of the purposes of this approach was to understand the advantages of nonlinear models with ML versus linear regression models based on RCs.

In this study, we applied three data pre-processing steps for the raw normalized datasets. The first step was standardization to remove trends related to the variance and mean from the datasets; the second step was the removal of seasonal effects through Fourier series fitting because the data were influenced by seasonality (Figure 1b); the third step was first-order differencing to convert a nonstationary data set to a stationary data set. Then, MI and correlation coefficient methods were applied to determine the dominant model inputs, accounting for time lags. The correlation coefficient was used in the analysis because it defines the dependence of two independent variables in time and space; therefore, it is a kind of temporal correlation for time series with different time lags.

To remove the seasonal influence from time series data, a fitting Fourier series model was used. The basic concept of this Fourier series model for time series decomposition was proposed by Delurgio (1998) as follows:

$$X_t = a + bt + \sum_{j=1}^{k} (a_j \cos jwt + b_j \sin wt) \tag{1}$$

where X_t is the fitted value at time t; a is a constant related to the series level; b is the trend estimate of the series; a_j and b_j ($j = 1, 2, 3, …, k$) are Fourier coefficients; w is the Fourier frequency; and k is the highest harmonic of w.

The first-order differencing method has been widely used as a simple procedure to convert nonstationary time series to stationary time series, as proposed by Anderson (1976). In other words, a new data set of a variable can be obtained from a measured data set by subtracting the value of that variable at time $t − 1$ ($X_{t−1}$) from its value at time t (X_t). This method can be expressed as follows:

$$Y_t = X_t + X_{t−1} \qquad (2)$$

MI is a quantitative metric based on information entropy, and it expresses the dependence or cooperation among random variables (Akca & Yozgatlıgil, 2020). Unlike traditional correlation metrics, MI does not require an assumption based on dependence, and the provided mutual information encompasses both linear and nonlinear relationships. MI stems from Shannon entropy in information theory (Shannon, 1948). The discontinuous random variable x (from x_1 to x_n) and probabilities (from P_1 to P_N) are expressed by the following equation:

$$H(x) = \sum_{i=1}^{N} P(x_i) \, Log \, [P(x_i)] \qquad (3)$$

The MI criterion is the amount of information shared among discontinuous variables X and Y (Yang et al., 2000). It is assumed that the two variables x and y correspond to probabilities m and n, and the ranges of these probabilities are indicated by i and j, respectively. Accordingly, MI is defined as MI(X, Y), where A and B share MI.

$$MI(X, Y) = \sum_{i=1}^{K} \sum_{j=1}^{L} P_{XY}(x_1, y_1) \log \left(\frac{P_{XY}(x_1, y_1)}{P_1(X = x_i) P_2(X = y_i)} \right) \qquad (4)$$

In this equation, P_i is the probability of i; $P(i, j)$ is the joint probability of i and j.

Based on the results of the MI and correlation coefficient analyses, five dominant model inputs were used: the water levels at Kratie at $t–1$ ($WK_{t−1}$) and $t–2$ ($WK_{t−2}$) and the water levels at Prek Kdam at t (WP_t), $t−1$ ($WP_{t−1}$), and $t−2$ ($WP_{t−2}$). In this case, t is the selected time, and $t–1$ and $t–2$ are the 1- and 2-day lagged times, respectively. Notably, the flow at Prek Kdam changes sooner than the flow at Kratie because Prek Kdam is closer to Tan Chau than is Kratie. The time lag concept is not considered in the RC model.

3.2. ML Models: Theoretical Background and Optimization

In the ML models, we considered three periods for the training data set (70% of the data), namely, 1980–1983, 1996–2007, and 2012–2015, and two periods in the testing data set (30% of the data), namely, 1984–1995 and 2008–2011. The training data set was selected to represent all kinds of flow events, ranging from flood years (e.g., in 2000) to drought years (e.g., in 2015). Similarly, the testing data set covered both flood years (e.g., in 2011) and drought years (e.g., in 2010). For each ML method, we established the theoretical background and adjusted hyperparameters for model optimization.

3.2.1. Decision Tree

DT is an ML method used for prediction and classification (Quinlan, 1986). This method has been employed in various studies due to its simplicity and high predictive accuracy (Choi et al., 2019).

A DT could predict responses by converting the observed values of features in ML models. These models are based on the relationship between the predictor and the response for a given data set. A DT defines each parameter and determines distinct values based on the impurities observed at roots. Therefore, the DT approach is straightforward to implement; nevertheless, its reliability is sometimes inadequate because it is prone to overfitting and linear regression loss (Ragettli et al., 2017). Therefore, to limit the likelihood of overfitting and inaccuracy, the tree size should be determined via cross-validation (Choi et al., 2020).

This study used fine tree regression and ensemble boosting regression to train the ML model. The corresponding DTs were compared with other models to find a suitable DT for reconstructing missing data at the analyzed

station. Each DT is composed of an initial point (root) and an ultimate point (leaf) in tree form (Saghafi & Arabloo, 2017). The main tuning hyperparameter in DTs is the minimum leaf size. Leaf size was used to train the DT-based ML model, while trees were used to search for the optimal result (Krzywinski & Altman, 2017).

3.2.2. Gaussian Process Regression

The GPR paradigm is a probabilistic non-parametric kernel model (Rasmussen & Williams, 2006). The GP is a potential algorithm for calculating the ideal distribution of flexible and malleable regression and classification modeling techniques that are not restricted to basic parametric forms (Weir et al., 2019). Furthermore, one of the GPR's advantages is its wide range of covariance coefficients. Notably, functions with varying degrees of smoothness or other kinds of contiguous structures may be employed. This enables the user to make an acceptable choice (Rasmussen & Nickisch, 2010). Furthermore, GPR models may determine the distributions of functions using one or more input parameters (Rasmussen & Williams, 2006). When such functions are used to calculate the average response of the regression model with Gaussian error, the related matrix computations may be modified for inference; this technique is useful for training datasets with a large number of samples (Neal, 1997; Weir et al., 2019).

In detail, GPR analyses the training database $\{(x_i, y_i); i = 1, 2, ..., n\}$, in which $x_i \in R^d$ and $y_i \in R$; both factors were selected from an undefined population. GPR models estimate values of the response parameter y_{new} based on the new input variable x_{new} and the training database. The corresponding linear equation is expressed as follows (Rasmussen & Williams, 2006):

$$y = x^T \beta + \varepsilon, \tag{5}$$

where the error term $\varepsilon \sim N(0, \sigma^2)$, the predictor observation N, the error variance σ^2, and the coefficient β are determined from the database.

Moreover, the GPR model's building blocks include a GP that uses a random variable to convert objective functions (Rasmussen & Williams, 2006). Therefore, a Gaussian second-order statistic is established, and the new form of the GP is as follows:

$$f(x) \sim GP[m(x), k(x, x')] \tag{6}$$

in which $m(x)$ and $k(x, x')$ are the mean and covariance functions, respectively.

The predicted observation values are the same as those obtained with Equation 5, but the corresponding variances depend on the noise in the observation set (Weir et al., 2019). To convert GPR to a covariance function, Equation 6 is implemented for all plausible compositions of points, and the result is rewritten in three matrices (Rasmussen & Williams, 2006):

$$K(X, X) = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2)\cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2)\cdots & k(x_2, x_n) \\ \vdots & \vdots & \vdots \\ k(x_n, x_1) & k(x_n, x_2)\cdots & k(x_n, x_n) \end{pmatrix} \tag{7}$$

To optimize the GPR training models, $k(x,x')$ is normally parameterized using a group of kernel parameters (θ), known as tuning hyperparameters. $k(x,x')$ is denoted as $k(x,x'|\theta)$ to explicitly specify the dependence on θ (Sun et al., 2014). Hence, θ and ε in Equation 5 are the major tuning hyper-parameters in this paper.

3.2.3. Support Vector Regression

SVR, first developed by Vapnik (2013), has been extensively applied for classification and prediction in many research domains. The basic equation in SVR is as follows:

$$\hat{f}(x) = \omega^T \phi(x) + b, \tag{8}$$

in which ϕ is a mapping function with a weight of ω and b is a scalar. T is the inner product/dot product parameter of the hyperplane equation. The widely used regression form of SVR, ε-SVR, was applied in this paper. Considering N_s training samples, the ordinary formula for ε-SVR is provided by Vapnik (2013):

$$\min_{\omega,b,\xi,\xi^*} \frac{1}{2}\omega^T\omega + C \sum_{i=1}^{N_s} \left(\xi_i + \xi_i^*\right) \tag{9}$$

Subject to $\omega^T \phi(x_i) + b - y_i \leq \varepsilon + \xi_i$,

$$y_i - \omega^T \phi(x_i) - b \leq \varepsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, N_s,$$

where C is used to signify the penalized variable, and ξ_i and ξ_i^* are the slack variables, which specify the upper and lower bounds, respectively, of the training errors, considering the error tolerance ε (Chen & Pawar, 2019). The optimization problem in Equation 9 can be handled with the aid of a collection of Lagrange multipliers: α_i and α_i^* (Chen & Pawar, 2019; Schölkopf & Smola, 2002). By adopting a typical quadratic programming technique, this process allows the optimization issues to be addressed quickly in dual format (Shevade et al., 2000). As a consequence, the second equation utilized to solve the SVR optimization problem is as follows (Chen & Pawar, 2019; Schölkopf & Smola, 2002; Shevade et al., 2000):

$$\min_{\alpha_i,\alpha_i^*} \frac{1}{2} \sum_{i,j=1}^{N_s} \left(\alpha_i - \alpha_i^*\right)\left(\alpha_i - \alpha_i^*\right) K(x_i, x_j) + \varepsilon \sum_{i=1}^{N_s} \left(\alpha_i + \alpha_i^*\right) + \sum_{i=1}^{N_s} y_i\left(\alpha_i - \alpha_i^*\right) \tag{10}$$

subject to $\sum_{i=1}^{N_s} \left(\alpha_i - \alpha_i^*\right) = 0$,

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, 2, \dots, N_s,$$

where $K(x_i, x_j)$, the kernel function, is the inner product of $\phi(x_i)$ and $\phi(x_j)$. The linear kernel, the polynomial kernel, the radial basis function (RBF) kernel, and the hyperbolic tangent kernel are all frequently used kernel functions. In this research, the SVR model was trained using the RBF kernel, as follows (Chen & Pawar, 2019; Schölkopf & Smola, 2002):

$$(x_i, x_j) = \exp\left(-\gamma \|x_i - x_{j_2}^2\|\right), \gamma > 0 \tag{11}$$

By using α_i and α_i^*, the predictive model is expressed as follows (Chen & Pawar, 2019):

$$\hat{f}(x) = \sum_{i=1}^{N_s} (-\alpha_i + \alpha_i^*) K(x_i, x) + b \tag{12}$$

The matrix for the nonnegative module of $(-\alpha_i + \alpha_i^*)$, where $i = 1, 2, \dots, N_s$, is referred to as the linear kernel. The SVR model can output all the support vectors if an input x is given. In this study, the three tuning hyperparameters C, γ, and ε were used to optimize the predictive models, and default values were used for all the variables.

3.2.4. Random Forest

RF is a classification and regression technique based on DT (Breiman, 2001). RF regression is also characterized as an ensemble-based ML technique that generates a set of input variables (known as training datasets and predictions) to create numerous regression trees. These trees can be merged to provide more precise and reliable results (Liaw & Wiener, 2002). Moreover, each DT regression in the ensemble is trained utilizing bootstrap samples or a sampling bag from the training data set to ensure that it performs well. Ultimately, each tree node is divided using binary splits based on a selection of randomly chosen predictors, with each split resulting in a different outcome (Liaw & Wiener, 2002). The RF method generates various independent DTs, which are described as follows (Breiman, 2001):

$$h_N(x) = \frac{1}{N_{DT}} \sum_{i=1}^{N} h_i(x) \tag{13}$$

where $h_i(x)$ is a DT and N_{DT} is defined as the total number of DTs.

The number of DTs in the forest (N_E_{st}) serves as one of the tuning hyperparameters in the optimum RF model, while the maximum depth of DTs (Max_D) and number of features are used to search for the best split (Max_F) (Kim & Shin, 2020).

3.2.5. Least Square Support Vector Machine

The LSSVM model is regarded as a more straightforward variant of the SVM regression model (Suykens & Vandewalle, 1999) and is more flexible than the original SVM method. Moreover, instead of utilizing quadratic programming to tackle regression problems, it is beneficial to determine a linear set of equations using a support vector to solve them more rapidly (Suykens et al., 2002).

A target training data set is determined as $\{x_k, y_k\}$, $k = 1, 2, ..., N$, in which $x_k \in R$ stands for the kth input data; $y_k \in R$ is the output parameter for the given input parameter; and N is the amount of data trained (Ahmadi & Ahmadi, 2016). By considering the nonlinear function $\varphi(.)$, the following regressed equation is generated (Suykens & Vandewalle, 1999):

$$y = \omega^T \varphi(x) + b \tag{14}$$

in which ω is a weight vector; $\varphi(x)$ is the nonlinear function; T depicts the transpose of the matrix; and b denotes the bias parameter. According to Equation 14, this nonlinear function plots the input data set x into the n-infinite feature space (Vong et al., 2006). When the LSSVM is used, it introduces a unique optimizing case. The approach adopted tackles the following optimization issue:

$$\min_{\omega, b, e} J(\omega, e) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2 \tag{15}$$

Equation 15 considers the following equality constraint:

$$y = \omega^T \varphi(x_k) + b + e_k \qquad k = 1, 2...., N \tag{16}$$

in which γ represents the model parameters and takes into account the model's complexity and the training error (Mehdizadeh & Movagharnejad, 2011); e_k indicates the error in the regression. The Lagrangian is constructed in the following manner to seek a resolution to the unbounded optimization problem:

$$L(\omega, b, e, \alpha) = J(\omega, e) - \sum_{k=1}^{N} \alpha_k \left\{ \omega^T \phi(x_k) + b + e_k - y_k \right\} \tag{17}$$

in which α_k denotes the Lagrange multiplier or supporting value. In obtaining a solution for Equation 17, the transformation of the equation in terms of ω, b, e_k, α_k is described as follows:

$$\frac{\partial L(\omega, b, e, \alpha)}{\partial \omega} = 0 \rightarrow \omega = \sum_{k=1}^{N} \alpha_k \varphi(x_k) \tag{18}$$

$$\frac{\partial L(\omega, b, e, \alpha)}{\partial b} = 0 \rightarrow \sum_{k=1}^{N} \alpha_k = 0 \tag{19}$$

$$\frac{\partial L(\omega, b, e, \alpha)}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, \quad k = 1, 2, ..., N \tag{20}$$

$$\frac{\partial L(\omega, b, e, \alpha)}{\partial \alpha_k} = 0 \rightarrow y_k = \varphi(x_k)\omega^T + b + e_k, \quad k = 1, 2, ..., N \tag{21}$$

When the parameters ω and e are removed, the Karush–Kuhn–Tucker system can be obtained as follows:

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & \Omega + \gamma^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{22}$$

In Equation 22, $y = [y_1.....y_N]^T$; $1_N = [1........1]^T$; $\alpha = [\alpha_1.....\alpha_N]^T$; I is the identity matrix.

$\Omega_k = \varphi(x_k)^T.\varphi(x_l) = K(x_k, x_l) \forall k, \quad l = 1, 2, ...N$; $K(x_k, x_l)$ is the kernel function that must satisfy Mercer's condition (Li et al., 2008). Three options are available for the kernel function:

$$K(x, x_k) = x_k^T x \tag{23}$$

$$K(x, x_k) = \left(\tau + x_k^T x\right)^d \tag{24}$$

$$K(x, x_k) = \exp\left(-x - x_k^2/\sigma^2\right) \tag{25}$$

Based on the above three options, the following is a description of the latest part of the LSSVM algorithm for parameter estimation:

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{26}$$

where τ denotes the slope; d represents the degree of the polynomial; σ^2 denotes the kernel sample variance; (b, α) stands for the answer to the equations' linear system illustrated in Equation 22. In this study, σ^2 and γ were considered to be the two main hyperparameters for tuning the LSSVM model. These two parameters are vital for obtaining the optimal prediction performance for daily discharge in mega deltas.

3.2.6. Multivariate Adaptive Regression Splines

MARS is defined as non-parametric regression technique proposed by Friedman (1991). It can map nonlinearities and interactions between parameters. MARS creates a predictive model ($\hat{f}$) via the following equation:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i B_i(\boldsymbol{x}) \tag{27}$$

where $\boldsymbol{x}$ represents the vector that includes all the input variables; B_i represents the basis functions; k denotes the number of basis functions defined by the regression function; and α_i denotes the i_{th} constant coefficient. Additionally, each basis function $B_i(\boldsymbol{x})$ considers one of the following constraints: (a) a single basis function has a constant value approximately equal to 1; (b) a hinge function; (c) at least two or multiple hinge functions. A hinge function is illustrated by $\max(0, x - c)$ or $\max(0, c - x)$, in which c is a constant, namely, a knot (Chen & Pawar, 2019).

The utilization of hinge functions supports MARS by splitting the response surface into different continuous areas. MARS constructs a model in two stages with forward and backward processes. MARS begins with a model that comprises the single basis function of 1. Then, MARS frequently includes paired basis functions for the available basis functions. In each iteration step, it searches for the pair of basis functions that minimizes the sum-of-squares residual error (SSRE), which is defined as follows (Friedman, 1991):

$$SSRE = \sum_{i=1}^{N_s} \left(y_i - \hat{f}(x_i)\right)^2 \tag{28}$$

in which N_s denotes the number of training points; y_i is the i_{th} output achieved from training datasets; and x_i are the i_{th} input variables of training points. The added terms repeat until the change in the residual error is less than a target value or the maximum allowable term value is obtained. The forward process often creates an overfit model. To prevent overfitting, a backwards process is implemented to shape the ML model. The model extracts one less effective term in a paired basis function at each step until the best sub-model is obtained. The term choice to be removed is based on the minimum value of the generalized cross validation (GCV). The GCV is defined as follows:

$$GCV = \frac{SSRE}{N_s * (1 - N_e/N_s)} \tag{29}$$

where N_e indicates the number of effective terms. The backwards process permits the MARS method to build a model that integrates the good fit and model parsimony criteria. According to Friedman (1991), if MARS is given an input x, it can produce all the basis functions, $B_i(x)$, and their corresponding coefficients, α_i. In this study, the maximum number of terms and max_degree was the two main tuning hyperparameters, and all the remaining parameters were set to the default values.

3.2.7. K-Fold Cross Validation and Grid Search Process for Hyperparameter Tuning

The data set was separated into two subsets (training [70%] and testing [30%]) to develop the DT, GPR, SVR, RF, LSSVM, and MARS models. These ML models can be adjusted by varying the hyperparameters that control model performance. First, the training data were subjected to k-fold cross-validation (k-FCV) to determine the optimal hyperparameters (Markatou & Hripcsak, 2005). The training samples were subdivided into k subsets: $k-1$ sets were used to train the models. The kth parameter was employed to assess the performance of the hyperparameters based on the validation data. For each candidate hyperparameter, the procedure was repeated k times. Then, the goodness of fit of the ML models was estimated based on four statistical indicators for the training and validation datasets, namely, the root mean square error (*RMSE*), correlation coefficient (*r*), Nash-Sutcliffe number (*NSE*), and mean absolute error (*MAE*).

The grid search (GS) process creates groups for all compositions of values based on the prescribed search extent of hyperparameters and assesses each group using k-FCV (Kanin et al., 2019). The lowest *RMSE* and *MAE* or the highest *r* and *NSE* help to decide the optimum hyperparameter values. The *RMSE*, *r*, *NSE*, and *MAE* were computed as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})^2} \tag{30}$$

$$r = \frac{n \sum_{i=1}^{n} (Y_{i,m} \times Y_{i,e}) - \sum_{i=1}^{n} Y_{i,m} \sum_{i=1}^{n} Y_{i,e}}{\sqrt{n \sum_{i=1}^{n} Y_{i,m}^2 - \left(\sum_{i=1}^{n} Y_{i,m}\right)^2} \sqrt{n \sum_{i=1}^{n} Y_{i,e}^2 - \left(\sum_{i=1}^{n} Y_{i,e}\right)^2}} \tag{31}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})^2}{\sum_{i=1}^{n} \left(Y_{i,m} - \overline{Y}_{i,m}\right)^2} \tag{32}$$

$$MAE = \frac{\sum_{i=1}^{n} |Y_{i,m} - Y_{i,e}|}{n} \tag{33}$$

In these equations, $Y_{i,m}$, $Y_{i,e}$, and n are the observed discharge, predicted discharge, and total number of observations, respectively. *RMSE* is the difference between the simulated and measured values. *r* expresses the agreement between the simulated and observed values. *NSE* represents how well the simulated data match the corresponding observed values. *MAE* measures the errors between the predictions and the observations.

4. Results and Discussion

4.1. Selection of the Optimization Models

The proposed ML models require many tuning hyperparameters to train the data set. Some of these variables are important for achieving satisfactory model performance; therefore, they must be defined appropriately. To evaluate the ML model performance in reconstructing the daily-averaged discharge, k-FCV and GS processes were applied to achieve the optimal hyperparameter values for the DT, GPR, SVR, RF, LSSVM, and MARS models. In the GS process, the search range was divided into 30 grid divisions, and each grid division was then assessed using k-FCV. In this work, 10-fold cross-validation was selected for ML model optimization. Then,

Table 1
The Variables Used in the ML Models

Model	Adjusted hyperparameters	Specific search range	Optimal values of hyperparameters	Input	Output
DT	Leaf Size	1–3,115	40	WK_{t-1}	Daily discharge at Tan Chau
GPR	θ	57–57,000	12,100	WK_{t-2}	
	ε	0.001–69,055	68,500	WP_t	
				WP_{t-1}	
SVR	C	0.001–1,000	930	WP_{t-2}	
	γ	0.01–2,000	960	*WK: water level at	
	ε	10–9,000	2,300	Kratie	
RF	N_Est	100–500	310	*WP: water level at	
	Max_D	10–100	60	Prek Kdam	
				*t is the time, and t-1	
				and t−2 are the 1- and	
	Max_F	0.5–1.0	0.96	2-day lagged times	
LSSVM	σ^2	0.1–10	0.5		
	γ	5–100	10		
MARS	Max-terms	500–2,000	1,000		
	Max_degree	100–1,000	300		

the GS process was used to determine the minimum MSE for all six ML models based on the specific ranges of parameters. Table 1 provides detailed descriptions of the variables employed in training the six ML models.

The DT model was first examined by the GS process. Leaf size was used as the key tuning parameter. As shown in Table 1, the optimal leaf size was 40, and the DT model achieved the best prediction performance at this leaf size. Next, the GPR model was employed to evaluate the prediction ability for the daily discharge. This study used θ and ε to achieve the best estimating outcome of the GPR scheme. The best values of θ and ε were 12,100 and 68,500, respectively. The computational time to achieve the optimal tuning parameter of the GPR model was quite long because of the complex mathematical functions. Regarding the SVR model, three tuning parameters were used to achieve the desired predictive performance. Table 1 highlights the optimal values for various hyper-parameters. As previously mentioned, SVR can produce quick predictions to obtain the optimal tuning parameters. Following SVR, the RF model was applied to predict daily discharge at the Tan Chau station. *N_Est*, *Max_D*, and *Max_F* were used for the tuning process of the RF. The optimum values for the parameters were 310, 60, and 0.96, respectively. The computational time of the GS process for RF to achieve the optimal values for the three tuning parameters was also relatively short.

Furthermore, this study adopted advanced LSSVM and MARS approaches to validate the most robust ML models and reconstruct the daily-averaged discharge in the mega delta. According to the LSSVM model, σ^2 and γ are the two key parameters for the tuning process. The computational cost of the LSSVM model was high compared to that of the prior four evaluated ML models. In addition, the MARS model was also used for the comparison. As mentioned earlier in ML theory, max-terms and max_degree were employed to determine the best prediction performance of the MARS model. By tuning these two parameters using the GS process, the optimal outcome of the MARS method to predict daily-averaged discharge at the Tan Chau station was achieved.

4.2. What Models Are Recommended?

Here, we present a comparison of the results estimated by the ML and RC models in reconstructing the missing values of the daily-averaged discharge. Generally, all ML models were superior to the RC model, as indicated by the time series comparison, statistical indicators (r, *RMSE*, *NSE*, *MAE*, and percentile values in the violin plots), and percentage differences in the peak discharge (Figures 3–5; Table 2). These results suggest that ML models are applicable for simulating the hydrology of the VMD. On the basis of the MI and correlation coefficients in the
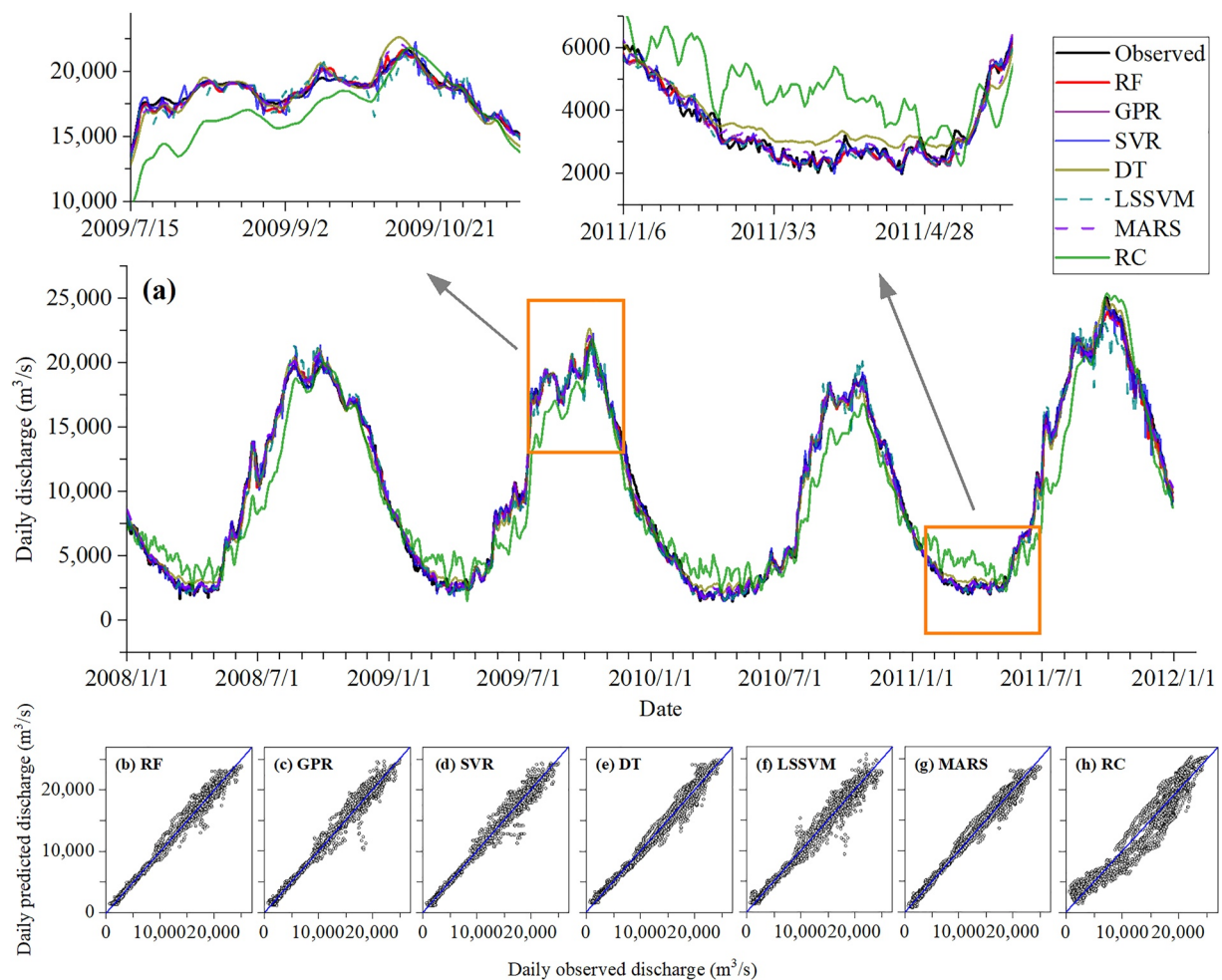
Figure 3. Results of the six machine learning (ML) and rating curve (RC) models for the testing data set. (a) Time series comparison between the predicted and measured values. (b–f) Scatter plots of the predicted versus measured values of individual ML and RC models, with the bisector line (1:1) shown.

data pre-processing, the effect of the flow at Prek Kdam on the flow at Tan Chau was found to be faster than the corresponding effect at Kratie by 1 day; in other words, the flow at Prek Kdam influenced the flow at Tan Chau sooner than the flow at Kratie. This difference is physically explained by the distance from Tan Chau to Prek Kdam (∼144 km), which is much shorter than that from Tan Chau to Kratie (∼316 km). This result confirms the importance of Tonle Sap Lake in regulating flows in the VMD, as also noted in previous research (e.g., Pokhrel et al., 2018).

The RC method produced the worst values of the statistical indicators (e.g., *RMSE* up to 2,438 m³/s in the training period) and the lowest accuracy in flood peak simulation (e.g., the flood peak was underestimated by −31% and −11.1% for the training and test datasets, respectively) (Table 2; Figures 3–5). A time series plot (Figure 3a) shows that the RC significantly underestimates the flood flow and overestimates the dry flow relative to the observed data. The underestimation of the flood peak is remarkably large in dry years, for instance, in 2010 (by −11.1%) and 1993 (by −7%) (Figure 5a). In flood months, the *NSE* values are very low (0.154 in the training and 0.523 in the testing periods) and are much lower than the respective values acquired from all the ML models (Table 2). The goodness of the predictions in the dry months from the RC is even lower than that in flood months (*NSE* = −0.77 and −2.64; *r* = 0.691 and 0.655; *RMSE* = 1,599 and 1,744 m³/s in the training and testing periods, respectively). Furthermore, Figures 4 and 5a clearly show the poor performance of the RC model compared to all six ML models.
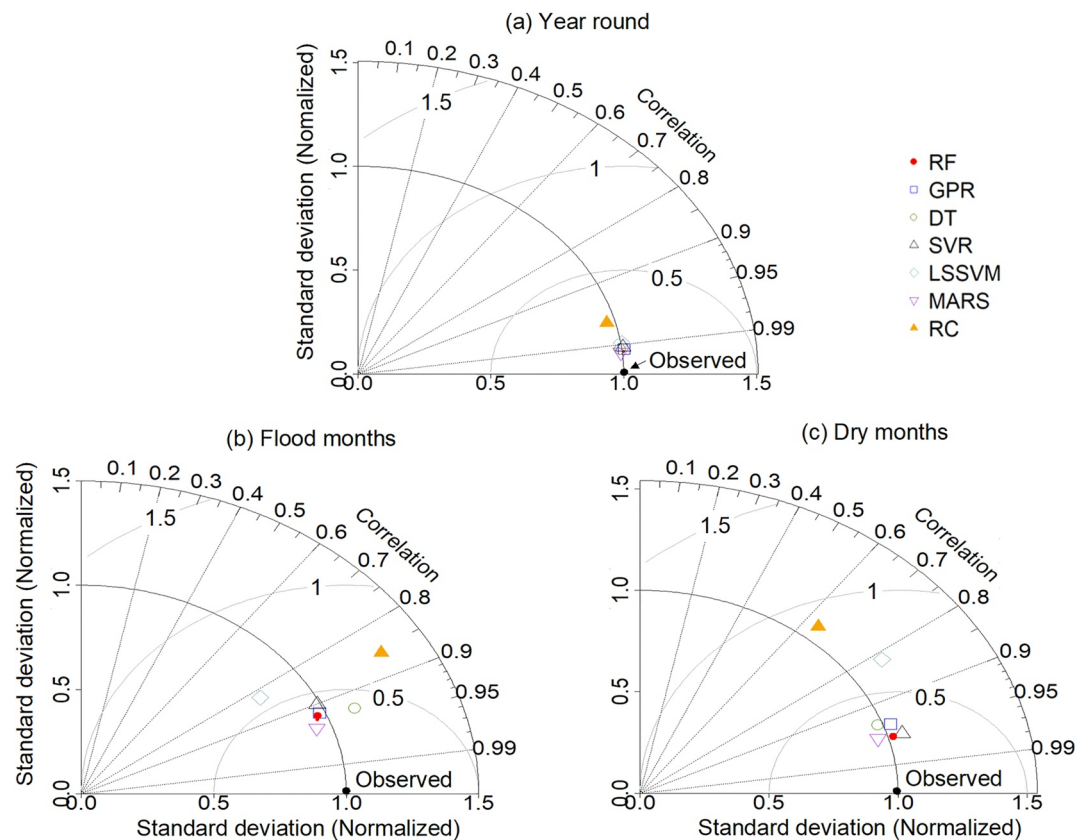
Figure 4. Taylor diagrams indicating the performance of six machine learning (ML) and rating curve (RC) models using the data from (a) year-round, (b) flood months, and (c) dry months.

These results indicate that the RC model based on year-round data alone cannot provide accurate reconstructions because of seasonal effects, which result in different hydrological behavior. To enhance the goodness of fit of the RC method, Moftakhari et al. (2015) established two RCs to predict the discharge of the Sacramento River by dividing the water level data into two subsets (i.e., <6.2 and >6.2 m) to account for seasonal effects. Binh, Kantoush, et al. (2021) revealed a clockwise hysteresis relation between the discharge and water level in the VMD, suggesting that RCs for rising and falling limbs should be developed differently. In the VMD, the flow characteristics in the rising and falling limbs are completely different: the former is controlled by riverine flood waves from upstream and the latter is strongly influenced by tides. RC is strongly influenced by the evolution of river geometry (i.e., erosion and deposition) and roughness (e.g., vegetation, infrastructure) and the effects of backwater and tides (Matte et al., 2018). Under the fluctuational tidal influence induced by oceanic and astronomical forcing (Jay et al., 2011; Moftakhari et al., 2013), a discharge magnitude does not yield a unique water level; rather, different water level values can be recorded (Hoitink & Jay, 2016). Specifically, because of the interactions among the reversing tidal flow, the tidal Stokes drift, spring-neap tidal cycle, lateral and estuarine circulations, the occurrence of multiple branches, and nonlinear frictional interactions between riverine flow and oceanic tides (Moftakhari et al., 2016), a hysteresis phenomenon is typically involved in the tidal process, in addition to many other hydrologic processes, as reported by Nourani, Parhizkar, et al. (2014). Due to this hysteresis behavior and loop, different outputs are possible for the same input; therefore, the RC method, which uses an injective function and a linear regression model, is not sufficiently robust to handle such nonlinear and complex problems. The above-mentioned phenomena are the root causes of the low prediction accuracy of the RC. To handle such issues, in the ML models established in this study, we used the data from multiple upstream gauges, as suggested by Moftakhari et al. (2016). We also used time-series data in the ML models to account for the temporal sequences involved in the data. This inclusion improved the modeling performance compared to that of RCs, which do not consider temporal data as an input. The complex hysteresis phenomenon involved in the tidal flow process can be robustly handled by the nonlinear artificial intelligence-based methods used in this study, as noted by Nourani,
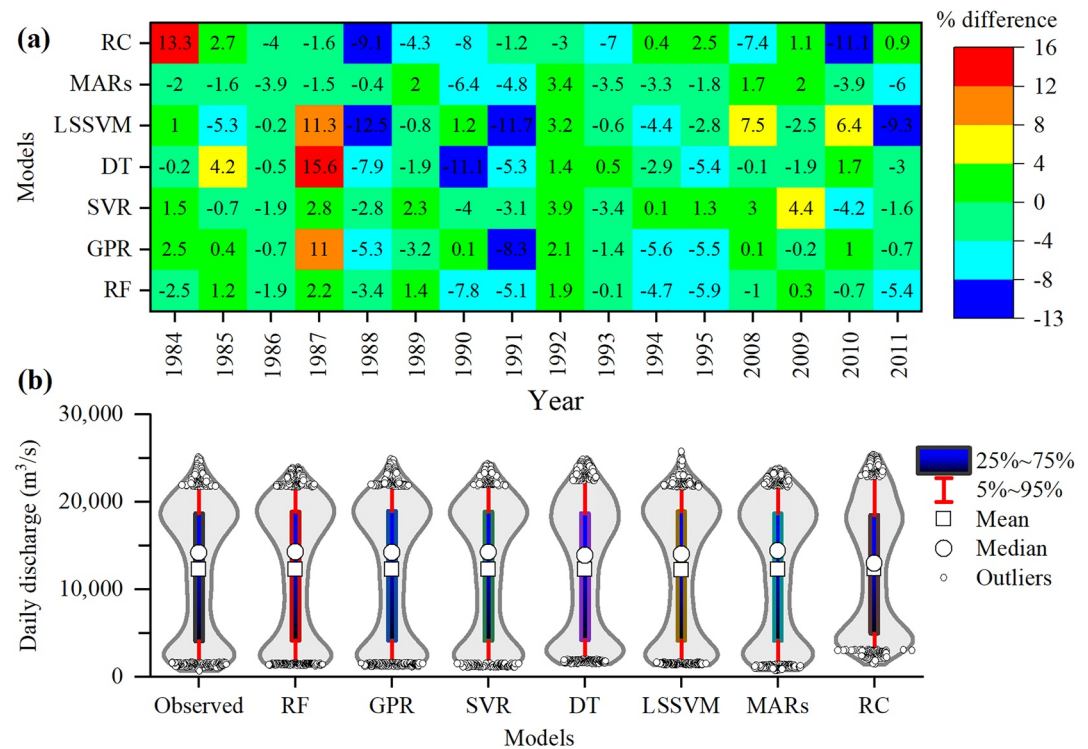
Figure 5. (a) Heatmap showing the percentage differences in the flood peaks predicted by the machine learning (ML) and rating curve (RC) models relative to the observed data. (b) Violin plots showing the goodness of fit of the predicted discharges versus the observed values. These plots are based on the data from the testing period.

Parhizkar, et al. (2014). Moreover, the use of RCs requires an extensive understanding of flow behavior and input data because both discharge and water level information from the same station are used; any misuse of the input data could directly lead to large discrepancies in the output. This influence is negligible in ML methods. Additionally, lag time must be considered in the ML models when water flows in the river from upstream to downstream, while this lag is ignored in the RC model. Moreover, ML models account for upstream-downstream relations and consider physical relations in addition to hydrological processes.

Among the ML models considered for the training data set, GPR and SVR perform better than the RF, DT, LSSVM, and MARS models in terms of statistical indicators (Table 2). For instance, the *RMSE* of GPR and SVR using year-round data is 389 and 476 m³/s, respectively, compared to 1,037 and 973 m³/s for the DT and MARS models, respectively. However, the MARS and RF models showed the best performance in the testing period, with the highest *r* and *NSE* values (e.g., r = 0.994 for both the MARS and RF models compared to 0.988 for the LSSVM model and 0.992 for the DT model using year-round data) and the lowest *RMSE* and *MAE* values (e.g., *RMSE* = 768 m³/s for the MARS model and 789 m³/s for the RF model compared to 943 m³/s for the SVR model and 1,098 for the LSSVM model considering the year-round data). Table 2 also shows the superior operation of the MARS and RF methods in both flood and dry months over the remaining four ML models with regard to all four statistical metrics.

The outstanding performance of the MARS and RF models over the other ML models can be clearly seen in the scatter plots (Figures 3b and 3g), Taylor diagrams (Figures 4b and 4c), and violin plots (Figure 5b). The scatter plots show that a majority of the scatter points using MARS and RF are concentrated around the bisector 1:1 line. Likewise, the Taylor diagrams point to the superior performance of the MARS and RF models because their results were the closest to the observed values. Based on the Taylor diagram, the LSSVM showed the worst performance, followed by the DT model because their results are the farthest from the observed points. This finding is confirmed by the *NSE* indicator shown in Table 2; for instance, the *NSE* values of the LSSVM and DT models were 0.528 and 0.676 in the dry months, respectively. Regarding the time series comparison (Figure 3a), the MARS and RF results effectively agree with the observed data. The SVR model produced unreasonable

Table 2
Statistical Indicators for the Six ML and RC Models

Model	Year round				Flood months				Dry months			
	r	*RMSE*	*NSE*	*MAE*	*r*	*RMSE*	*NSE*	*MAE*	*r*	*RMSE*	*NSE*	*MAE*
Training												
RF	0.996	602	0.992	412	0.965	695	0.931	528	0.978	251	0.956	193
GPR	0.998	389	0.997	273	0.988	399	0.977	286	0.976	263	0.952	202
DT	0.989	1,037	0.977	802	0.914	1,209	0.792	1,008	0.934	625	0.729	538
SVR	0.997	476	0.995	315	0.977	573	0.953	410	0.984	210	0.969	158
LSSVM	0.996	560	0.993	399	0.975	583	0.951	432	0.981	229	0.963	172
MARS	0.99	973	0.980	708	0.936	928	0.877	754	0.959	341	0.919	264
RC	0.954	2,087	0.91	1,629	0.841	2,438	0.154	2,026	0.691	1,599	−0.77	1,405
Testing												
RF	0.994	789	0.988	517	0.928	978	0.86	722	0.961	264	0.917	200
GPR	0.992	878	0.985	533	0.919	1,056	0.837	748	0.944	313	0.884	234
DT	0.992	886	0.985	659	0.929	1,082	0.829	815	0.939	522	0.676	454
SVR	0.991	943	0.983	588	0.901	1,160	0.804	845	0.961	274	0.911	206
LSSVM	0.988	1,098	0.977	747	0.826	1,497	0.673	1,059	0.818	631	0.528	426
MARS	0.994	768	0.989	544	0.942	872	0.889	646	0.96	257	0.921	198
RC	0.968	1,851	0.937	1,517	0.858	1,807	0.523	1,469	0.655	1,744	−2.64	1,578

Note. The models were built using year-round data; however, statistical indicators are also shown for the flood and dry months to highlight the performance of the individual models. The flood months are from August to October, and the dry months are from February to April. Time lag was not considered in the RC model. The unit of the *RMSE* and *MAE* is m³/s.

fluctuations in the discharge in the flood months; moreover, the DT model was unable to reliably reconstruct the discharge in the dry months. Another important indicator of discharge reconstruction is the flood peak, which has significant implications for flood management. All six ML models generally underestimated the flood peak, for instance, by −12.5% in 1990 using the LSSVM model; additionally, the GPR and SVR models overestimated the flood peak by up to 11% and 15.6% in 1987, respectively (Figure 5a). In flood years (e.g., 1996, 2000, and 2011), the RF and GPR models performed better than the SVR and DT models for the training data set, while the GPR and DT models performed better than the RF and SVR models for the testing data set. However, in drought years (i.e., 1993, 1998, 2005, and 2015), RF outperformed the other five ML models. For instance, the RF model underestimated the flood peak in 2010 by only −0.7%, while the DT model underestimated it by −4.2% (Figure 5a). The underestimation of the flood peaks, particularly in the extreme flood (e.g., 2000) and drought years (e.g., 2015), by the ML models was attributed to the limited data used to train the models. Underestimation of predicted flood peaks is common in hydrological modeling for both types of hydrological models (Tegegne et al., 2017; Vansteenkiste et al., 2014) and ML algorithms (Adnan et al., 2020; Tencaliec et al., 2015).

A comparison of the predicted and observed values using the time series plots, scatter plots, Taylor diagram, violin plots, heatmap, and statistical indicators reveals that the MARS and RF models most reliably reconstructed the daily-averaged discharge at Tan Chau (Figures 3–5; Table 2), although the performance of MARS was slightly better than that of RF. Thus, the MARS and RF models are recommended for daily-averaged discharge reconstruction in tidally affected river systems such as that in the VMD. This finding agrees with those of previous research that used ML models to assess hydrological processes (e.g., Hussain & Khan, 2020; Jeihouni et al., 2020; Li et al., 2016; Obringer & Nateghi, 2018). For instance, Obringer and Nateghi (2018) verified that the RF model was the best model among nonparametric ML algorithms in predicting riverine water levels. The GPR model is also trustworthy for reconstructing the discharge in tidally affected rivers. Notably, the DT model is not recommended in this study based on the abnormal and unreasonable fluctuations in the time series results, especially in dry months, compared to the measured data. In this study, we used individual ML models to reconstruct the missing discharge values. However, using a hybrid model by combining an ML model with an analytical, empirical or numerical model may improve the performance, and this research direction will be examined in future work.
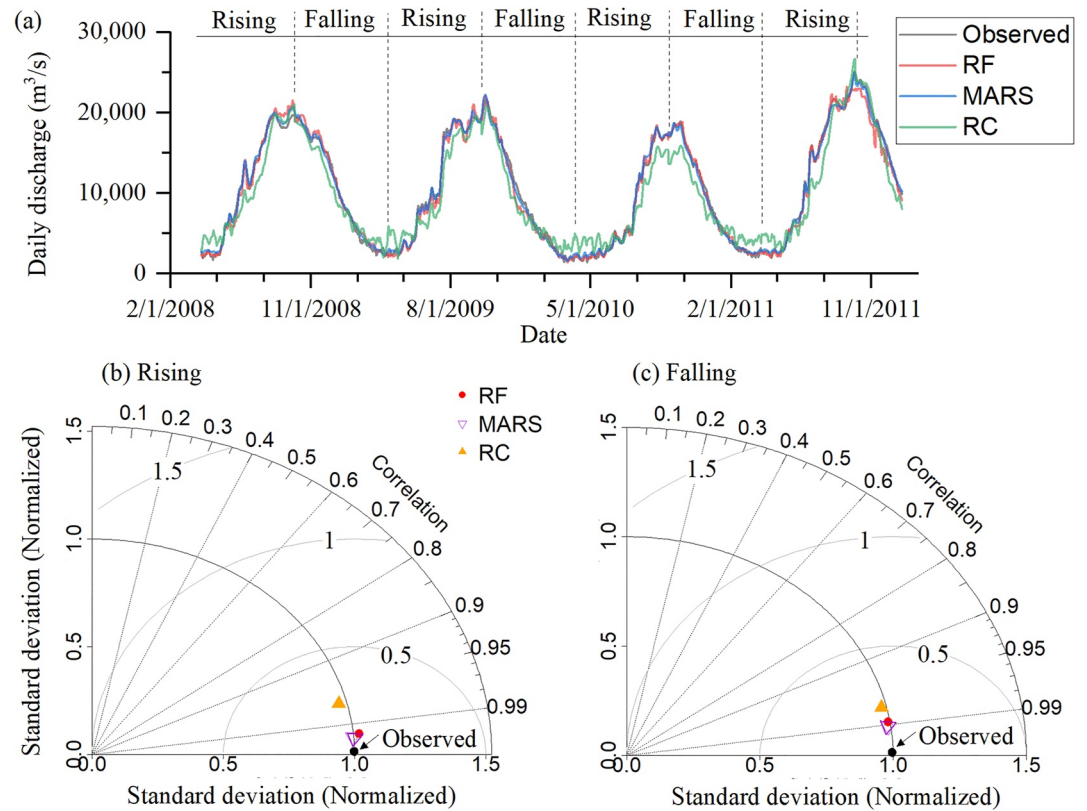
Figure 6. (a) Time series plot comparing the predicted discharge from multivariate adaptive regression spline (MARS), random forest (RF), and rating curve (RC) models with the observed data. (b, c) Taylor diagrams showing the performance of the three models in the rising and falling phases.

For instance, Safari (2020) found that the hybridization of the MARS and RF models with the empirical MNLR provides better predictions of incipient sediment deposition compared to the individual MARS and RF models.

4.3. Is Prediction Accuracy Increased by Considering Seasonal Patterns?

Binh, Kantoush, et al. (2021) found that separately constructing stage-discharge RCs for the rising and falling limbs can improve the reconstruction of the missing daily-averaged discharge in the VMD. Therefore, in this section, we attempt to apply the MARS and RF models in assessments of these two limbs to determine whether the prediction accuracy is enhanced. This approach also reduces the computational time and the effort required to collect data if the research focus is in flood or drought seasons.

Comparison of Figures 6 and 7 with Figures 3–5 along with statistical indicators (Table 2) reveals that the MARS and RF models established for the rising and falling limbs separately do not improve the prediction accuracy compared to the results obtained using the year-round data; in fact, the accuracy decreases slightly. Specifically, the statistical indicators of the RF model from the falling phase in the testing period ($r = 0.987$, $RMSE = 1,040$ m^3/s, $NSE = 0.973$, $MAE = 742$ m^3/s) are lower than those obtained using the year-round data ($r = 0.994$, $RMSE = 789$ m^3/s, $NSE = 0.988$, $MAE = 517$ m^3/s). Similarly, the corresponding values of the MARS model are $r = 0.991$, $RMSE = 856$ m^3/s, $NSE = 0.982$, $MAE = 485$ m^3/s, while those from the year-round data are $r = 0.994$, $RMSE = 768$ m^3/s, $NSE = 0.989$, $MAE = 544$ m^3/s. In flood peak prediction, however, the accuracy of the MARS and RF models considering seasonal patterns improves slightly compared to that in the original case. For instance, relative to the measured data, the mean predicted flood peaks in the test period of MARS and RF are underestimated by −1.4% and −1.1%, respectively, when seasonal patterns are considered and by −1.9% and −2%, respectively, when year-round data are used. The unexpected lack of improvements in the MARS and RF models for the rising and falling limbs is likely because of the reduction in the number of data points used (∼50% reduction). The above results, together with the Taylor diagrams (Figures 6b and 6c) and the scatter and violin
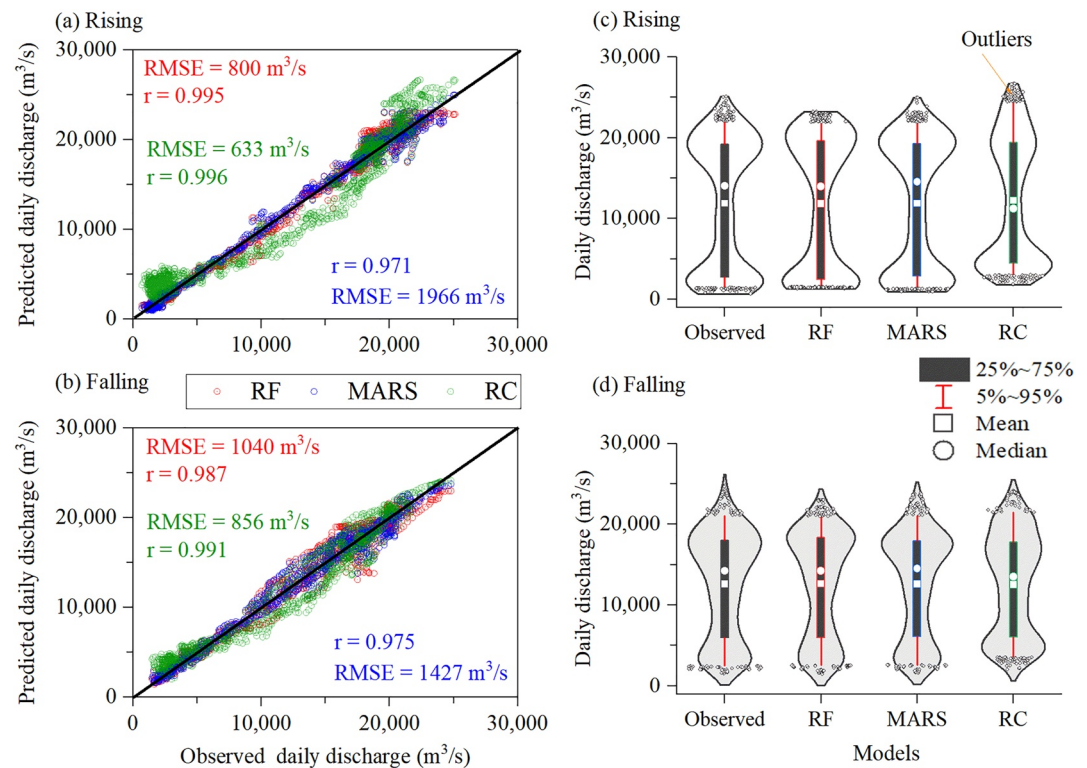
Figure 7. Predicted versus observed discharge values using the multivariate adaptive regression spline (MARS), random forest (RF), and rating curve (RC) models for the rising and falling phases. (a, b) Scatter plots with statistical indicators. (c, d) Violin plots comparing the results of the MARS, RF, and RC models with their observed counterparts for the rising and falling phases. All data plotted are from the test period. The MARS and RF models outperform the RC model.

plots (Figure 7), demonstrate that MARS is marginally better than RF. We also observed that the accuracy of MARS and RF in predicting the discharge in the rising phase was greater than that in the falling phase.

In contrast, the RCs that were separately developed for the rising and falling limbs enhanced the prediction accuracy substantially compared to those based on the year-round data. For example, in the test period, the performance of the RC model considering seasonal patterns (e.g., $r = 0.975$, $RMSE = 1,427$ m^3/s in the falling limb) was improved relative to that of the model using the year-round data (e.g., $r = 0.968$, $RMSE = 1,851$ m^3/s). Although improved, the results of the RC model are still inferior to those of the MARS and RF models, particularly for the rising limb (Figures 6 and 7).

In short, there are two main implications drawn from the above results. First, although using separate RCs for the rising and falling limbs can improve the prediction accuracy compared to those based on year-round data, the MARS and RF models still outperformed the RC model. However, if only the water level at a station at which the discharge needs to be reconstructed is available, RCs considering seasonal patterns can produce acceptable results. Second, the MARS and RF models that were separately established for the rising and falling limbs should be used with care, especially for the falling limb, if the research interest is the flood or drought period alone.

4.4. Prospects of Using ML in Hydrological Assessment in the Mekong River Basin

ML has been used by hydrologists to assess hydrological processes in many rivers worldwide, and its applications include rainfall (Alizadeh et al., 2017; Tikhamarine et al., 2020), streamflow (Luo et al., 2019; Zia et al., 2015), salinity (Tran et al., 2021), water quality (Elkiran et al., 2019; Imani et al., 2021), and sediment (Huang et al., 2019; Zounemat-Kermani et al., 2020) assessments. In tidally affected river systems, such as the Mekong River basin, the use of ML to explore hydrology and hydrological processes is limited. Notably, the use of ML to study hydrology is especially limited in the VMD, where data availability is a constraint for scientific research and resource management. Therefore, the use of ML in the VMD is recommended for (a) data reconstruction,

such as for tide, sediment, and salinity concentration data, and (b) hydrological and water quality prediction. We acknowledge that reconstructing salinity and sediment data is even more challenging than reconstructing river flows; thus, the latter is vital for the former and will be the focus of our future research. Such predictions have been made using statistical models (Apel et al., 2020). However, most predictions are medium-term estimates (up to 9 months), and appropriate and proactive water resource planning and management tasks require long-term predictions, especially considering variations in upstream inflows because of changes in dam management and climate change, downstream rising sea level and saline water instruction, and the increasing water demand within the delta (e.g., Binh, Kantoush, Saber, et al., 2020; Park et al., 2021, 2022). Therefore, ML methods are promising tools for gaining insight into hydrological processes and improving water resource management.

Flooding is an annual event in the VMD; however, extreme floods, such as the historical floods in 1996, 2000, and 2011, have appeared periodically and may cause disastrous damage to society (Triet et al., 2018). Therefore, flood prediction has an important role in flood preparedness. In the study of floods, it is crucial to predict flood water levels because it is the water level, not the discharge, that causes flood problems. In the tidally affected rivers in the VMD, the water level fluctuates remarkably under tidal effects within a day. Given the rapid evolution of flood water, it is necessary to predict hourly water levels instead of daily-averaged values, as was done in this study for the discharge reconstruction. Collectively, predicting the hourly flood water levels is even more challenging than predicting the daily-averaged discharge. As such, ML and RC models may yield undesirable predictions. Although challenging, we intend to develop a robust methodology using deep learning models to predict hourly flood water levels in the VMD in future studies, with the goal of helping the delta to prepare to cope with future flood disasters.

5. Concluding Remarks

Six ML-based methods were employed in the present research to reconstruct the missing daily-averaged discharge in a tidal river system, the VMD. The missing historical data have limited studies of long-term flow regime variations under the effects of climate change and intensifying anthropogenic intervention, such as the construction of hydropower dams and irrigation expansion. We used multi-station data considering upstream-downstream relations, and the water level at two upstream stations in different geographical settings was used to reconstruct the discharge at a downstream station. While many studies have ignored data pre-processing when completing similar tasks, we performed it in three steps: first, the raw data were normalized to remove trends in variance and mean; second, the Fourier series were fitted to remove seasonal effects; and third, first-order differencing was conducted. Additionally, MI and correlation coefficient methods were applied to optimize the model inputs and avoid the use of too many simulation parameters; moreover, lagged data were considered, thereby reducing the simulation time and effort. The results of our study are important for long-term water resource management in the delta, and the methodology developed can easily be adopted for other river systems.

Unlike the traditional stage-discharge RC method using linear regression, ML models can provide reliable reconstructions of the missing daily-averaged discharge. The water levels, with lagged time considerations, at Kratie and Prek Kdam are appropriate to input into the ML models. The present study provides a basis for hydrologists and researchers who plan to employ ML models for future water resource management in the VMD in the context of global warming. Our next step is to use ML to predict the discharge in the Mekong River from the short to the long term for optimal water resource allocation, particularly to enhance flood and drought preparedness.

The MARS and RF models are the two most suitable algorithms for reconstructing the missing daily-averaged discharge, although the MARS model performs slightly better than the RF model. These approaches can reliably predict flood peak, flood flow, and drought flow discharges and are therefore suitable for flood, drought and salinity intrusion studies. Compared to the RC method, the RF model reduces the *RMSE* and *MAE* by 135% and 194% (year-round data), 85% and 104% (flood-month data), and 561% and 691% (dry-month data), respectively. The respective values for the MARS and RC models are 141% and 179% (year-round data), 107% and 127% (flood-month data), and 578% and 696% (dry-month data). Establishing MARS and RF models for the rising and falling limbs separately did not improve the prediction accuracy; however, acceptable results could be obtained for a specific flood or drought period. MARS and RF reconstruct the daily-averaged discharge in the rising phase better than in the falling phase. The GPR and SVR models are also appropriate for daily-averaged discharge

reconstruction in the delta. The DT model, however, is not recommended because it produces abnormal, unreasonable fluctuations in the predicted discharge.

Since this paper is our first attempt to use ML models to estimate hydrological parameters in the VMD, we applied simple techniques to pre-process the input data. In future works, more advanced de-noising methods, such as the wavelet de-noising approach (Nourani, Baghanam, et al., 2014) or ensemble empirical mode decomposition (EEMD) (Gaci, 2016), are recommended for obtaining better performance in the application of ML models. Moreover, more advanced artificial intelligence models, such as deep learning neural networks (Ha et al., 2021), should be considered in the next attempt to forecast the discharge, sediment, and salinity in the VMD to support strategic decision making. The hybridization of ML models with other empirical, analytical or numerical models is also a good strategy to improve the prediction power of ML. Finally, our future work will focus on estimating the tidal water level and discharge on an hourly interval to quantify the net riverine and tidal flux exchange between rivers and seas.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Daily-averaged water levels at Kratie and Prek Kdam are available at https://portal.mrcmekong.org/time-series (the station names and type of data requested must be typed into the search boxes), and daily-averaged discharges at Tan Chau can be downloaded from Binh, Thanh, et al. (2021).

References

Adman, R. M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O., & Li, B. (2020). Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. *Journal of Hydrology*, *586*, 124371. https://doi.org/10.1016/j.jhydrol.2019.124371

Ahmadi, M. A., & Ahmadi, A. (2016). Applying a sophisticated approach to predict CO_2 solubility in brines: Application to CO_2 sequestration. *International Journal of Low Carbon Technologies*, *11*(3), 325–332. https://doi.org/10.1093/ijlct/ctu034

Akbarpour, F., Fathi-Moghadam, M., & Schneider, J. (2020). Application of LSPIV to measure supercritical flow in steep channels with low relative submergence. *Flow Measurement and Instrumentation*, *72*, 101718. https://doi.org/10.1016/j.flowmeasinst.2020.101718

Akca, E., & Yozgatlıgil, C. (2020). Mutual information model selection algorithm for time series. *Journal of Applied Statistics*, *47*(12), 2192–2207. https://doi.org/10.1080/02664763.2019.1707516

Alizadeh, M. J., Kavianpour, M. R., Kisi, O., & Nourani, V. (2017). A new approach for simulating and forecasting the rainfall-runoff process within the next two months. *Journal of Hydrology*, *548*, 588–597. https://doi.org/10.1016/j.jhydrol.2017.03.032

Anderson, O. (1976). *Time series analysis and forecasting: The box-Jenkins approach*. Butterworth.

Apel, H., Khiem, M., Quan, N. H., & Toan, T. Q. (2020). Brief communication: Seasonal prediction of salinity intrusion in the Mekong Delta. *Natural Hazards and Earth System Sciences*, *20*(6), 1609–1616. https://doi.org/10.5194/nhess-20-1609-2020

Binh, D. V., Kantoush, S., & Sumi, T. (2020). Changes to long-term discharge and sediment loads in the Vietnamese Mekong Delta caused by upstream dams. *Geomorphology*, *353*, 107011. https://doi.org/10.1016/j.geomorph.2019.107011

Binh, D. V., Kantoush, S. A., Saber, M., Mai, N. P., Maskey, S., Phong, D. T., & Sumi, T. (2020). Long-term alterations of flow regimes of the Mekong River and adaption strategies for the Vietnamese Mekong Delta. *Journal of Hydrology: Regional Studies*, *32*, 100742. https://doi.org/10.1016/j.ejrh.2020.100742

Binh, D. V., Kantoush, S. A., Sumi, T., Mai, N. P., Ngoc, T. A., Trung, L. V., & An, T. D. (2021). Effects of riverbed incision on the hydrology of the Vietnamese Mekong Delta. *Hydrological Processes*, *35*(2), e14030. https://doi.org/10.1002/hyp.14030

Binh, D. V., Thanh, H. V., Kantoush, S. A., Nourani, V., Saber, M., Lee, K. K., & Sumi, T. (2021). Long-term daily-averaged discharge time series in a megadelta. *Mendeley Data*. V1. https://doi.org/10.17632/xnzvfcyy68.1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Buschman, F. A., Hoitink, A. J. F., van der Vegt, M., & Hoekstra, P. (2009). Subtidal water level variation controlled by river flow and tides. *Water Resources Research*, *45*(10), W10420. https://doi.org/10.1029/2009WR008167

Chen, B., & Pawar, R. J. (2019). Characterization of CO_2 storage and enhanced oil recovery in residual oil zones. *Energy*, *183*, 291–304. https://doi.org/10.1016/j.energy.2019.06.142

Choi, C., Kim, J., Han, H., Han, D., & Kim, H. S. (2020). Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea. *Water (Switzerland)*, *12*(1), 93. https://doi.org/10.3390/w12010093

Choi, C., Kim, J., Kim, J., & Kim, H. S. (2019). Development of combined heavy rain damage prediction models with machine learning. *Water (Switzerland)*, *11*(12), 2516. https://doi.org/10.3390/w11122516

Delurgio, S. A. (1998). *Forecasting principles and applications*. Mc Graw-Hill.

Elkiran, G., Nourani, V., & Abba, S. I. (2019). Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *Journal of Hydrology*, *577*, 123962. https://doi.org/10.1016/j.jhydrol.2019.123962

Eslami, S., Hoekstra, P., Trung, N. N., Kantoush, S. A., Binh, D. V., Dung, D. D., et al. (2019). Tidal amplification and salt intrusion in the Mekong Delta driven by anthropogenic sediment starvation. *Scientific Reports*, *9*(18746). https://doi.org/10.1038/s41598-019-55018-9

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, *19*(1), 1–67. https://doi.org/10.1214/aos/1176347963

Gaci, S. (2016). A new ensemble empirical mode decomposition (EEMD) denoising method for seismic signals. *Energy Procedia*, *97*, 84–91. https://doi.org/10.1016/j.egypro.2016.10.026

Gisen, J. I. A., & Savenije, H. H. G. (2015). Estimating bankfull discharge and depth in ungauged estuaries. *Water Resources Research*, *51*(4), 2298–2316. https://doi.org/10.1002/2014WR016227

Granata, F., Papirio, S., Esposito, G., Gargano, R., & de Marinis, G. (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. *Water (Switzerland)*, *9*(2), 105. https://doi.org/10.3390/w9020105

Grumbine, R. E., Dore, J., & Xu, J. (2012). Mekong hydropower: Drivers of change and governance challenges. *Frontiers in Ecology and the Environment*, *10*(2), 91–98. https://doi.org/10.1890/110146

Gugliotta, M., Saito, Y., Nguyen, V. L., Ta, T. K. O., Nakashima, R., Tamura, T., et al. (2017). Process regime, salinity, morphological and sedimentary trends along the fluvial marine transition zone of the mixed-energy Mekong River Delta, Vietnam. *Continental Shelf Research*, *147*, 7–26. https://doi.org/10.1016/j.csr.2017.03.001

Ha, S., Liu, D., & Mu, L. (2021). Prediction of Yangtze River streamflow based on deep learning neural network with El Niño-Southern Oscillation. *Scientific Reports*, *11*(1), 11738. https://doi.org/10.1038/s41598-021-90964-3

Hadi, S. J., & Tombul, M. (2018a). Forecasting daily streamflow for basins with different physical characteristics through data-driven methods. *Water Resources Management*, *32*(10), 3405–3422. https://doi.org/10.1007/s11269-018-1998-1

Hadi, S. J., & Tombul, M. (2018b). Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination. *Journal of Hydrology*, *561*, 674–687. https://doi.org/10.1016/j.jhydrol.2018.04.036

Helaire, L. T., Talke, S. A., Jay, D. A., & Chang, H. (2020). Present and future flood hazard in the lower Columbia river estuary: Changing flood hazards in the Portland-Vancouver metropolitan area. *Journal of Geophysical Research: Oceans*, *125*(7), e2019JC015928. https://doi.org/10.1029/2019JC015928

Hoa, L. T. V., Nhan, N. H., Wolanski, E., Cong, T. T., & Shigeko, H. (2007). The combined impact on the flooding in Vietnam's Mekong River delta of local man-made structures, sea level rise, and dams upstream in the river catchment. *Estuarine, Coastal and Shelf Science*, *71*(1–2), 110–116. https://doi.org/10.1016/j.ecss.2006.08.021

Hoitink, A. J. F., & Jay, D. A. (2016). Tidal river dynamics: Implications for deltas. Reviews of. *Geophysics*, *54*(1), 240–272. https://doi.org/10.1002/2015RG000507

Huang, C. C., Fang, H. T., Ho, H. C., & Jhong, B. C. (2019). Interdisciplinary application of numerical and machine-learning-based models to predict half-hourly suspended sediment concentrations during typhoons. *Journal of Hydrology*, *573*, 661–675. https://doi.org/10.1016/j.jhydrol.2019.04.001

Hussain, D., & Khan, A. A. (2020). Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan. *Earth Science Informatics*, *13*(3), 939–949. https://doi.org/10.1007/s12145-020-00450-z

Imani, M., Hasan, M. M., Bittencourt, L. F., McClymont, K., & Kapelan, Z. (2021). A novel machine learning application: Water quality resilience prediction model. *The Science of the Total Environment*, *768*, 144459. https://doi.org/10.1016/j.scitotenv.2020.144459

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R* (pp. 303–335). Springer Texts in Statistics Book Series.

Jay, D. A., Leffler, K., & Degens, S. (2011). Long-term evolution of Columbia river tides. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *137*(4), 182–191. https://doi.org/10.1061/(ASCE)WW.1943-5460.0000082

Jeihouni, M., Alavipanah, S. K., Toomanian, A., & Jafarzadeh, A. A. (2020). Digital mapping of soil moisture retention properties using solely satellite-based data and data mining techniques. *Journal of Hydrology*, *585*, 124786. https://doi.org/10.1016/j.jhydrol.2020.124786

Kanin, E. A., Osiptsov, A. A., Vainshtein, A. L., & Burnaev, E. V. (2019). A predictive model for steady-state multiphase pipe flow : Machine learning on lab data. *Journal of Petroleum Science and Engineering*, *180*, 727–746. https://doi.org/10.1016/j.petrol.2019.05.055

Kantoush, S., Binh, D. V., Sumi, T., & Trung, L. V. (2017). Impact of upstream hydropower dams and climate change on hydrodynamics of Vietnamese Mekong Delta. *Journal of Japan Society of Civil Engineers, Series B1 (Hydraulic Engineering)*, *73*(4), I_109–I_114. https://doi.org/10.2208/jscejhe.73.i_109

Kebede, M. G., Wang, L., Li, X., & Hu, Z. (2020). Remote sensing-based river discharge estimation for a small river flowing over the high mountain regions of the Tibetan Plateau. *International Journal of Remote Sensing*, *41*(9), 3322–3345. https://doi.org/10.1080/01431161.2019.1701213

Khalil, M., Panu, U. S., & Lennox, W. C. (2001). Groups and neural networks-based streamflow data infilling procedures. *Journal of Hydrology*, *241*(3–4), 153–176. https://doi.org/10.1016/S0022-1694(00)00332-2

Khan, M. Y. A., 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. *Hydrological Sciences Journal*, *61*(11), 2084–2095. https://doi.org/10.1080/02626667.2015.1083650

Kim, M., & Shin, H. (2020). Machine learning-based prediction of the shale barrier size and spatial location using key features of SAGD production curves. *Journal of Petroleum Science and Engineering*, *191*, 107205. https://doi.org/10.1016/j.petrol.2020.107205

Kisi, O., & Parmar, K. S. (2016). Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *Journal of Hydrology*, *534*, 104–112. https://doi.org/10.1016/j.jhydrol.2015.12.014

Krzywinski, M., & Altman, N. (2017). Classification and regression trees. *Nature Methods*, *14*(8), 757–759. https://doi.org/10.1038/nmeth.4370

Li, B., Yang, G., Wan, R., Dai, X., & Zhang, Y. (2016). Comparison of random forests and other statistical methods for the prediction of lake water level: A case study of the Poyang Lake in China. *Hydrology Research*, *47*(S1), 69–83. https://doi.org/10.2166/nh.2016.264

Li, C. H., Zhu, X. J., Cao, G. Y., Sui, S., & Hu, M. R. (2008). Identification of the Hammerstein model of a PEMFC stack based on least squares support vector machines. *Journal of Power Sources*, *175*(1), 303–316. https://doi.org/10.1016/j.jpowsour.2007.09.049

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*, 18–22.

Liu, D., Jiang, W., Mu, L., & Wang, S. (2020). Streamflow prediction using deep learning neural network: Case study of Yangtze River. *IEEE Access*, *8*, 90069–90086. https://doi.org/10.1109/ACCESS.2020.2993874

Loc, H. H., Binh, D. V., Park, E., Shrestha, S., Dung, T. D., Son, V. H., et al. (2021). Intensifying saline water intrusion and drought in the Mekong Delta: From physical evidence to policy outlooks. *The Science of the Total Environment*, *757*, 143919. https://doi.org/10.1016/j.scitotenv.2020.143919

Luo, X., Yuan, X., Zhu, S., Xu, Z., Meng, L., & Peng, Z. (2019). A hybrid support vector regression framework for streamflow forecast. *Journal of Hydrology*, *568*, 184–193. https://doi.org/10.1016/j.jhydrol.2018.10.064

Markatou, M., & Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, *6*, 1127–1168.

Matte, P., Secretan, Y., & Morin, J. (2018). Reconstruction of tidal discharges in the St. Lawrence fluvial estuary: The method of cubature revisited. *Geophysical Research: Oceans*, *123*(8), 5500–5524. https://doi.org/10.1029/2018JC013834

Mehdizadeh, B., & Movagharnejad, K. A. (2011). Comparative study between LS-SVM method and semi empirical equations for modeling the solubility of different solutes in supercritical carbon dioxide. *Chemical Engineering Research and Design*, *89*(11), 2420–2427. https://doi.org/10.1016/j.cherd.2011.03.017

Mispan, M. R., Rahman, N. F. A., Ali, M. F., KHalid, K., Bakar, M. H. A., & Haron, S. H. (2015). Missing river discharge data imputation approach using artificial neural network. *ARPN Journal of Engineering and Applied Sciences*, *10*(22), 10480–10485.

Moftakhari, H. R., Jay, D. A., & Talke, S. A. (2016). Estimating river discharge using multiple-tide gauges distributed along a channel. *Journal of Geophysical Research: Oceans*, *121*(4), 2078–2097. https://doi.org/10.1002/2015JC010983

Moftakhari, H. R., Jay, D. A., Talke, S. A., Kukulka, T., & Bromirski, P. D. (2013). A noval approach to flow estimation in tidal rivers. *Water Resources Research*, *49*(8), 4817–4832. https://doi.org/10.1002/wrcr.20363

Moftakhari, H. R., Jay, D. A., Talke, S. A., & Schoellhamer, D. H. (2015). Estimation of historic flows and sediment loads to San Grancisco Bay, 1849–2011. *Journal of Hydrology*, *529*, 1247–1261. https://doi.org/10.1016/j.jhydrol.2015.08.043

Neal, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *Preparation Physics* (Technical Report No. 9702, pp. 1–24). Department of Statistics, University of Toronto.

Nourani, V., Andalib, G., & Dabrowska, D. (2017). Conjunction of wavelet transform and SOM-mutual information data pre-processing approach for AI-based multi-station nitrate modeling of watersheds. *Journal of Hydrology*, *548*, 170–183. https://doi.org/10.1016/j.jhydrol.2017.03.002

Nourani, V., Baghanam, A. H., Rahimi, A. Y., & Nejad, F. H. (2014). Evaluation of Wavelet-based de-noising approach in hydrological models linked to artificial neural networks. In I. Islam, P. K. Srivastava, M. Gupta, X. Zhu, & S. Mukherjee(Eds.), *Computational intelligence techniques in Earth and environmental sciences*. Springer.

Nourani, V., Parhizkar, M., Vousoughi, F. D., & Amini, B. (2014). Capability of artificial neural network for detecting hysteresis phenomenon involved in hydrological processes. *Journal of Hydrologic Engineering*, *19*(5), 896–906. https://doi.org/10.1061/(asce)he.1943-5584.0000870

Nourani, V., Tajbakhsh, A. D., Moladjou, A., & Gokcekus, H. (2019). Hybrid wavelet-M5 model tree for rainfall-runoff modeling. *Journal of Hydrologic Engineering*, *24*(5), 04019012. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001777

Obringer, R., & Nateghi, R. (2018). Predicting urban reservoir levels using statistical learning techniques. *Scientific Reports*, *8*(1), 5164. https://doi.org/10.1038/s41598-018-23509-w

Panahi, M., Sadhasivam, N., Pourghasemi, H. R., Rezaie, F., & Lee, S. (2020). Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). *Journal of Hydrology*, *588*, 125033. https://doi.org/10.1016/j.jhydrol.2020.125033

Park, E., Ho, H. L., Binh, D. V., Kantoush, S., Poh, D., Alcantara, E., et al. (2022). Impacts of agricultural expansion on floodplain water and sediment budgets in the Mekong River. *Journal of Hydrology*, *605*, 127296. https://doi.org/10.1016/j.jhydrol.2021.127296

Park, E., Loc, H. H., Binh, D. V., & Kantoush, S. (2021). The worst 2020 saline water intrusion disaster of the past century in the Mekong Delta: Impacts, causes, and management implications. *Ambio*, *51*(3), 691–699. https://doi.org/10.1007/s13280-021-01577-z

Pokhrel, Y., Burbano, M., Roush, J., Kang, H., Sridhar, V., & Hyndman, D. W. (2018). A review of the integrated effects of changing climate, land use, and dams on Mekong River hydrology. *Water*, *10*(3), 266. https://doi.org/10.3390/w10030266

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/bf00116251

Ragettli, S., Zhou, J., Wang, H., Liu, C., & Guo, L. (2017). Modeling flash floods in ungauged mountain catchments of China: A decision tree learning approach for parameter regionalization. *Journal of Hydrology*, *555*, 330–346. https://doi.org/10.1016/j.jhydrol.2017.10.031

Raghavendra, S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, *19*, 372–386. https://doi.org/10.1016/j.asoc.2014.02.002

Rasmussen, C. E., & Nickisch, H. (2010). Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, *11*, 3011–3015.

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes in machine learning. In *Adaptive computation and machine learning* (p. 48). MIT Press.

Rezaali, M., Quilty, J., & Karimi, A. (2021). Probabilistic urban water demand forecasting using wavelet-based machine learning models. *Journal of Hydrology*, *600*, 126358. https://doi.org/10.1016/j.jhydrol.2021.126358

Safari, M. J. S. (2020). Hybridization of multivariate adaptive regression splines and random forest models with an empirical equation for sediment deposition prediction in open channel flow. *Journal of Hydrology*, *590*, 125392. https://doi.org/10.1016/j.jhydrol.2020.125392

Saghafi, H., & Arabloo, M. (2017). Modeling of CO_2 solubility in MEA, DEA, TEA, and MDEA aqueous solutions using AdaBoost-decision tree and artificial neural network. *International Journal of Greenhouse Gas Control*, *58*, 256–265. https://doi.org/10.1016/j.ijggc.2016.12.014

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, *11*(5), 1188–1193. https://doi.org/10.1109/72.870050

Sun, A. Y., Wang, D., & Xu, X. (2014). Monthly streamflow forecasting using Gaussian process regression. *Journal of Hydrology*, *511*, 72–81. https://doi.org/10.1016/j.jhydrol.2014.01.023

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letter*, *9*(3), 293–300. https://doi.org/10.1023/A:1018628609742

Suykens, J. A., Van Gestel, T., De Brabanter, J., & Van Moor, B. (2002). *Least squares support vector machines*. World Scientific. https://doi.org/10.1142/5089

Ta, T. K. O., Nguyen, V. L., Tateishi, M., Kobayashi, I., Tanabe, S., & Saito, Y. (2002). Holocene delta evolution and sediment discharge of the Mekong River, southern Vietnam. *Quaternary Science Review*, *21*(16–17), 1807–1819. https://doi.org/10.1016/S0277-3791(02)00007-0

Taherdangkoo, R., Liu, Q., Xing, Y., Yang, H., Cao, V., Sauter, M., & Butscher, C. (2021). Predicting methane solubility in water and seawater by machine learning algorithms: Application to methane transport modeling. *Journal of Contaminant Hydrology*, *242*, 103844. https://doi.org/10.1016/j.jconhyd.2021.103844

Tegegne, G., Park, D. K., & Kim, Y. O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies*, *14*, 49–66. https://doi.org/10.1016/j.ejrh.2017.10.002

Tencaliec, P., Favre, A. C., Prieur, C., & Mathevet, T. (2015). Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research*, *51*(12), 9447–9463. https://doi.org/10.1002/2015WR017399

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A. N., Sammen, S. S., Kisi, O., Huang, Y. F., & El-Shafie, A. (2020). Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. *Journal of Hydrology*, *589*, 125133. https://doi.org/10.1016/j.jhydrol.2020.125133

Tongal, H., & Booij, M. J. (2018). Simulation and forecasting of stream flows using machine learning models coupled with base fl ow separation. *Journal of Hydrology*, *564*, 266–282. https://doi.org/10.1016/j.jhydrol.2018.07.004

Tran, D. A., Tsujimura, M., Ha, N. T., Nguyen, V. T., Binh, D. V., Dang, T. D., et al. (2021). Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. *Ecological Indicators*, *127*, 107790. https://doi.org/10.1016/j.ecolind.2021.107790

Tran, H. D., Muttil, N., & Perera, B. J. C. (2015). Selection of significant input variables for time series forecasting. *Environmental Modelling & Software*, *64*, 156–163. https://doi.org/10.1016/j.envsoft.2014.11.018

Triet, N. V. K., Dung, N. V., Fujii, H., Kummu, M., Merz, B., & Apel, H. (2017). Has dyke development in the Vietnamese Mekong Delta shifted flood hazard downstream? *Hydrology and Earth System Sciences*, *21*(8), 3991–4010. https://doi.org/10.5194/hess-21-3991-2017

Triet, N. V. K., Dung, N. V., Merz, B., & Apel, H. (2018). Towards risk-based flood management in highly productive paddy rice cultivation— Concept development and application to the Mekong delta. *Natural Hazards and Earth System Sciences*, *18*(11), 2859–2876. https://doi.org/10.5194/nhess-18-2859-2018

Vansteenkiste, T., Tavakoli, M., Steenbergen, N. V., Smedt, F. D., Batelaan, O., Pereira, F., & Willems, P. (2014). Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. *Journal of Hydrology*, *511*, 335–349. https://doi.org/10.1016/j.jhydrol.2014.01.050

Vapnik, V. N. (2013). *The nature of statistical learning theory*. Springer.

Vong, C. M., Wong, P. K., & Li, Y. P. (2006). Prediction of antomotive engine power and torque using least squares support vector machines and Bayesian inference. *Engineering Application of Artificial Intelligence*, *19*(3), 277–287. https://doi.org/10.1016/j.engappai.2005.09.001

Wang, H., Yan, H., Zeng, W., Lei, G., Ao, C., & Zha, Y. (2020). A novel nonlinear Arps decline model with salp swarm algorithm for predicting pan evaporation in the arid and semi-arid regions of China. *Journal of Hydrology*, *582*, 124545. https://doi.org/10.1016/j.jhydrol.2020.124545

Weir, K., Salmasi, F., Arvanaghi, H., Karbasi, M., & Farsadizadeh, D. (2019). Application of Gaussian process regression model to predict discharge coefficient of gated Piano. *Water Resources Management*, *33*(11), 3929–3947. https://doi.org/10.1007/s11269-019-02343-3

Yaghoubi, B., Hosseini, S. A., & Nazif, S. (2019). Monthly prediction of streamflow using data-driven models. *Journal of Earth System Science*, *128*(6), 141. https://doi.org/10.1007/s12040-019-1170-1

Yang, H. H., Van Vuuren, S., Sharma, S., & Hermansky, H. (2000). Relevance of time–frequency features for phonetic and speaker-channel classification. *Speech Communication*, *31*(1), 35–50. https://doi.org/10.1016/s0167-6393(00)00007-8

Yang, T., Gao, X., Sorooshian, S., & Li, X. (2016). Simulating California reservoir operation using the classification and regression-treen algorithm combined with a shuffled cross-validation scheme. *Water Resources Research*, *52*(3), 1626–1651. https://doi.org/10.1002/2015WR017394

Yarar, A. (2014). A hybrid wavelet and neuro-fuzzy model for forecasting the monthly streamflow data. *Water Resources Management*, *28*(2), 553–565. https://doi.org/10.1007/s11269-013-0502-1

Zhao, X., Lv, H., Lv, S., Sang, Y., Wei, Y., & Zhu, X. (2021). Enhancing robustness of monthly streamflow forecasting model using gated recurrent unit based on improved grey wolf optimizer. *Journal of Hydrology*, *601*, 126607. https://doi.org/10.1016/j.jhydrol.2021.126607

Zhou, W., Wu, X., Ding, S., & Cheng, Y. (2020). Predictive analysis of the air quality indicators in the Yangtze River Delta in China : An application of a novel seasonal grey model. *The Science of the Total Environment*, *748*, 141428. https://doi.org/10.1016/j.scitotenv.2020.141428

Zhu, S., Luo, X., Xu, Z., & Ye, L. (2018). Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection. *Hydrology Research*, *50*(1), 200–214. https://doi.org/10.2166/nh.2018.023

Zhu, S., Zhou, J., Ye, L., & Meng, C. (2016). Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. *Environmental Earth Sciences*, *75*(6), 531. https://doi.org/10.1007/s12665-016-5337-7

Zia, H., Harris, N., Merrett, G., & Rivers, M. (2015). Predicting discharge using a low complexity machine learning model. *Computers and Electronics in Agriculture*, *118*, 350–360. https://doi.org/10.1016/j.compag.2015.09.012

Zounemat-Kermani, M., Mahdavi-Meymand, A., Alizamir, M., Adarsh, S., & Yaseen, Z. M. (2020). On the complexities of sediment load modeling using integrative machine learning: Application of the great river of Loiza in Puerto Rico. *Journal of Hydrology*, *585*, 124759. https://doi.org/10.1016/j.jhydrol.2020.124759