Vietnam Journal of Earth Sciences, 46(4), 468-488, https://doi.org/10.15625/2615-9783/21067

**Vietnam Journal of Earth Sciences**
https://vjs.ac.vn/index.php/jse

# Multi-step-ahead prediction of water levels using machine learning: A comparative analysis in the Vietnamese Mekong Delta

Nguyen Duc Hanh[1], Nguyen Tien Giang[1*], Le Xuan Hoa[2], Tran Ngoc Vinh[3], Huu Duy Nguyen[4]

*[1]Faculty of Hydrology, Meteorology and Oceanography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Vietnam*
*[2]Dong Thap Provincial Hydrometeorological Station, Southern Regional Hydrometeorological Station, General Department of Hydrometeorology, Ministry of Natural Resources and Environment*
*[3]Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA*
*[4]Faculty of Geography, VNU University of Science, Vietnam National University, Hanoi, Vietnam*

Received 12 May 2024; Received in revised form 20 June 2024; Accepted 02 July 2024

ABSTRACT

This study evaluates the efficacy of five machine learning algorithms Support Vector Regression (SVR), Decision Tree (DT), Random Forest (RF), Light Gradient Boosting Machine Regressor (LGBM), and Linear Regression (LR) in predicting water levels in the Vietnamese Mekong Delta's tidal river system, a complex nonlinear hydrological phenomenon. Using daily maximum, minimum, and mean water level data from the Cao Lanh gauging station on the Tien River (2000-2020), models were developed to forecast water levels one, three, five, and seven days in advance. Performance was assessed using Nash-Sutcliffe Efficiency, coefficient of determination, Root Mean Square Error, and Mean Absolute Error. Results indicate that all models performed well, with SVR consistently outperforming others, followed by RF, DT, and LGBM. The study demonstrates the viability of machine learning in water level prediction using solely historical water level data, potentially enhancing flood warning systems, water resource management, and agricultural planning. These findings contribute to the growing knowledge of machine learning applications in hydrology and can inform sustainable water resource management strategies in delta regions.

*Keywords*: Water level, multi-step-ahead prediction, machine learning, Vietnamese Mekong delta.

## 1. Introduction

Water level prediction is a critical component in various domains, including industrial agriculture development, natural hazard management, and water resource administration (Choi et al., 2019; Wang and Wang, 2020; Nguyen et al., 2022). Fluctuations in water levels have far-reaching implications for the water cycle, sediment transport processes, water quality, and ecosystem dynamics. Consequently, the ability to forecast water level changes with high precision is paramount. Such accurate predictions can provide invaluable support to local authorities in implementing effective water resource management strategies and mitigating natural hazards (Herath et al., 2023; Kim et al., 2022; Do et al., 2022).

*Corresponding author, Email: giangnt@vnu.edu.vn

Recognizing the significance of this issue, the International Joint Commission (IJC) has emphasized the necessity of developing innovative methods and techniques to enhance the predictive capabilities and monitoring efficiency of existing systems.

Water level and discharge prediction in rivers can be accomplished through various models, generally categorized into two main types: physically-based and data-driven. A third category, hybrid models, has recently emerged, combining elements of both approaches (e.g., Sampurno et al., 2021; Ghaith et al., 2019; Li and Jun 2022; Vinh et al., 2023a). Physically based models are mathematical equations that represent conceptual models and fundamental physical principles, such as the conservation of mass and momentum (Chua, 2012). The development of these models necessitates the processing of hydrological parameters, which demands expert knowledge, high-quality data, and a comprehensive understanding of the basin's hydrological processes (Kim et al., 2015). Examples include Mike Nam (DHI, 1999), the Variable Infiltration Capacity (VIC) model (Nanda et al., 2019), TOPMODEL (Peters et al., 2003), and SWAT (Narsimlu et al., 2015). Hydrodynamic models, such as the HD module in Mike 11 (Rabuffetti and Barbero, 2005) and HEC-RAS (Hicks and Peacock, 2005), also fall into this category. While physically-based models excel at addressing 'what if' scenarios, their predictive accuracy depends on the user's ability to forecast input variables, which often involves additional forecasting tasks. Moreover, these models require complex computational processes and parameter adjustments, potentially leading to uncertainties in results and increased time for water level estimation or prediction (Vinh et al., 2020; Vinh and Jongho, 2022; Vinh and Jongho, 2019). The complexity and non-linearity of water level changes, influenced by factors such as meteorological conditions, tidal effects, and inter-watershed flow exchanges, pose significant challenges for traditional hydraulic models in achieving high-precision predictions, especially for nonlinear problems (Pan et al., 2020; Park et al., 2022).

In recent years, the advancement of computational capabilities has led to increased application of data-driven methods for water level prediction across various global regions. These models excel in analyzing and predicting nonlinear relationships between variables, such as rainfall and streamflow, significantly enhancing hydrological model performance. Data-driven modeling approaches have emerged as valuable tools in hydrology, offering the ability to analyze complex terrains and situations with limited data without requiring extensive knowledge of underlying physical processes (Wunsch et al., 2018). These methods often allow for rapid model development with minimal input requirements (Mosavi et al., 2018; Vinh et al., 2024; Vinh et al., 2023b; Nguyen et al., 2023). While comparing data-driven and physically-based models presents challenges due to their distinct characteristics and data needs, several studies have indicated that machine-learning techniques may offer superior water-level prediction capabilities in specific contexts (Baek et al., 2020; Zhao et al., 2020; Zhu et al., 2020, Özdoğan-Sarıkoç and Dadaser-Celik, 2024). Various machine learning techniques, including Gaussian process regression, multilayer perceptron, random forest, and multiple linear regression, have been successfully applied to water level prediction in diverse geographical contexts, such as Lake Erie in North America (Wang and Wang, 2020), the Red River in Vietnam (Phan and Nguyen, 2020), and the Durian Tunggal river in Malaysia (Ahmed et al., 2022). However, the literature reveals that model performance varies across different regions, emphasizing the absence of a universal water-level prediction model. Consequently, selecting and testing appropriate models for accurate water level prediction remains crucial to support decision-makers and local authorities in effective water resource management, economic development, and natural hazard mitigation, particularly flooding.

In this study, we comprehensively evaluated various machine learning (ML) models for water level prediction in the Vietnamese Mekong Delta to identify the most suitable model for forecasting in this region. Specifically, we developed and compared five ML models: support vector regression (SVR), Decision Tree (DT), Random Forest (RF), Light Gradient Boosting Machine Regressor (LGBM), and linear regression (LR). These models were applied to predict water levels at the Cao Lanh station in Dong Thap province, where water resource management poses significant challenges to agricultural development. The predictive performance of these five models was comparatively analyzed using multiple evaluation metrics. It is important to note that most previous related studies in the Vietnamese Mekong Delta primarily used hydrodynamic models to simulate flow. The applicability of ML in this context remains an open question. This study represents the first application of these five models for multi-step-ahead water level prediction at the Cao Lanh station. The findings from this research have the potential to assist decision-makers and farmers in developing effective strategies for water management in the study area.

## 2. Study area and data used

Dong Thap Province is one of the 13 provinces in the Mekong Delta, located at 10°07'-10°58' North latitude and 105°12'-105°56' East longitude. Dong Thap Province has a natural area of approximately 3384 km$^2$ and is home to more than 1.6 million people. The topography of the study area is relatively flat, with an average height of 1-2 m. The altitude decreases from north to south and from west to east. The Tien River system divides Dong Thap province into two areas: the north of the Tien River, with an area of about 250,731 ha, and the area south of the Tien River, with approximately 73,074 ha.

The study area has a dense river system with two major rivers: the Tien and Hau Rivers. Dong Thap province is located in the tropical monsoon climate zone, the climate of which is divided into two distinct seasons. The rainy season starts from May to November, and the dry season starts from December to April of the following year. Precipitation in the study area varies from 1392 to 2388 mm/year, of which about 90% of the precipitation is concentrated in August, September, and October. The uneven distribution of precipitation throughout the year leads to difficulties in allocating water resources for agricultural development. The tide in the study area belongs to the semi-diurnal tide regime, comprising 2 high tides and 2 low tides with an average height ranging from 3 to 4 m.

Dong Thap province, in particular, and the Mekong Delta, in general, are considered areas severely affected by climate change. According to the Ministry of Natural Resources and Environment's climate change scenario, by the end of the 21st century, sea levels will rise by about 46 cm, making drought and saltwater intrusion more severe. Therefore, the development of modern methods to forecast water levels with high accuracy plays a vital role in helping people develop appropriate adaptation measures (Fig. 1).

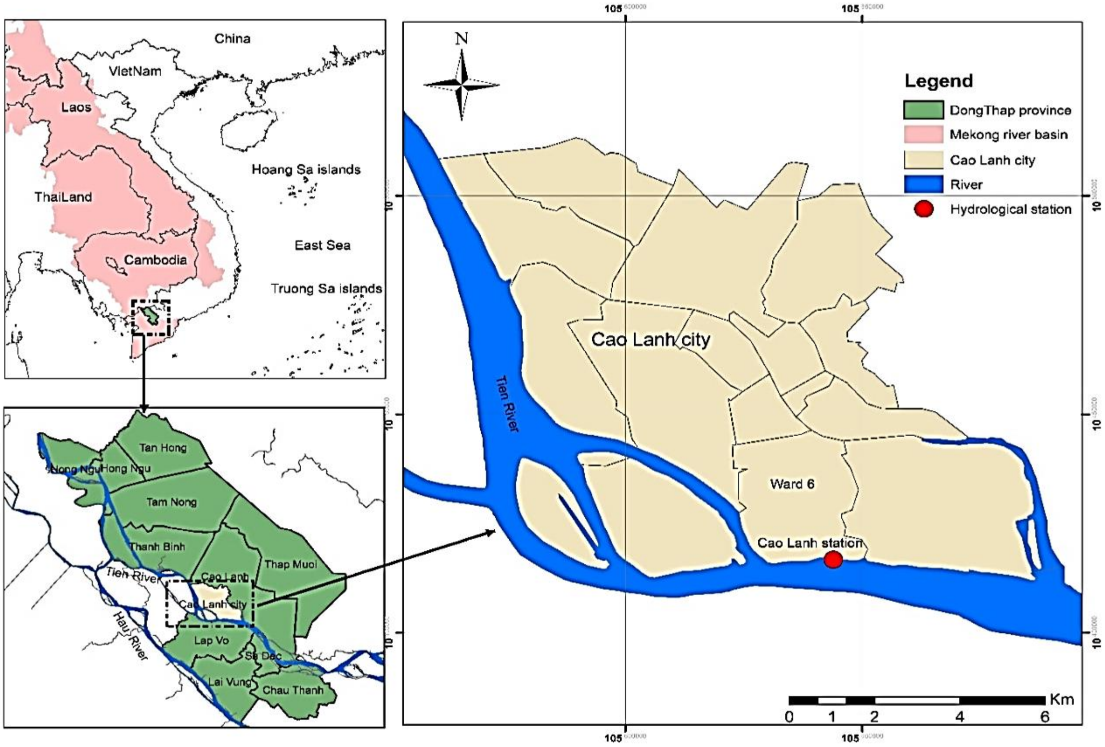

*Figure 1.* Location of study area

Several countries upstream of the Mekong River have built dams and reservoirs to meet the growing demand for water resources. Due to the relatively staggered information-sharing policy between countries, downstream farmers cannot get the information in time. This influences agricultural development policies downstream. In this study, the daily maximum, minimum, and average water level time series extracted from observed hourly data at Cao Lanh station on the Tien River from 2000 to 2020 (sources: Southern Regional Hydrometeorological Center - Vietnam Meteorological and Hydrological Administration) was used to build the water level prediction models. These three predictors were chosen following the requirements dictated in legal documents relating to hydrological forecasting in Vietnam (e.g., MONRE, 2023). These data were divided into two parts: 80% of the data for data training and 20% of the data for validation. More specifically, daily data from January 1, 2000, to October 19, 2016, were used for model construction/training; from October 20, 2016, to December 31, 2019, for model validation; and from January 1, 2020, to December 31, 2020, for prediction testing. Figure 2 shows the observed daily maximum water level time series as an example.

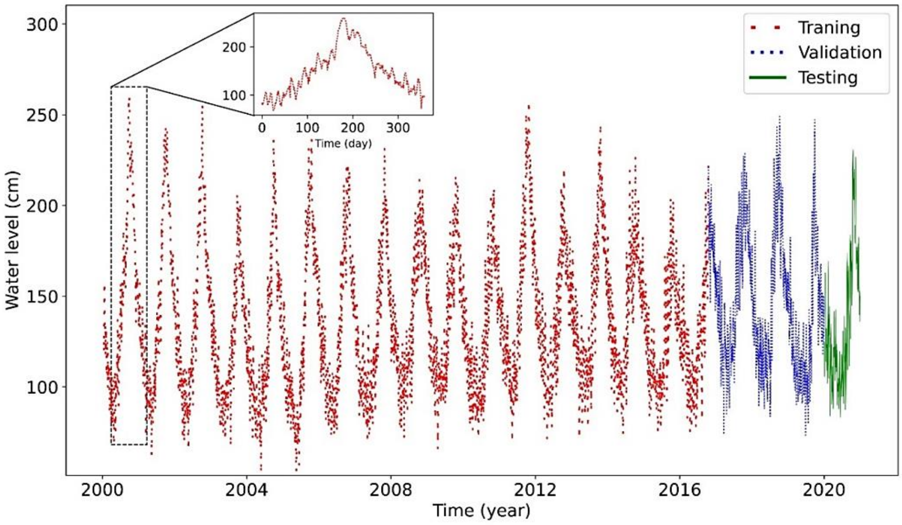

*Figure 2*. Daily maximum water level time series measured at Cao Lanh hydrological station

## 3. Methodology

The methodology used to construct SVR, DT, RF, LGBM, and LR in this study was divided into four main stages: (i) data collection and preprocessing; (ii) construction of prediction models; (iii) validation of models for one-step prediction; iv) testing of models for multi-step prediction (Fig. 3).

(i) Data collection and preprocessing: data on daily maximum, minimum, and average water levels from 2000 to 2020 were collected as input data for the prediction model. These data were divided into three parts: daily data from January 1, 2000, to October 18, 2016, were used for model construction/training; from October 19, 2016, to December 31, 2019, for model validation; from January 01, 2020, to December 31, 2020, for multi-step-ahead prediction testing (Fig. 2). Since univariate time series of water levels at a single gauging station is used for prediction, sequence length or sliding window width (i.e. the number of previous time steps used for prediction) need to be specified. Using this sequence length, input variables can be selected. The present study utilized ACF and PACF plots to determine this value. Besides, several normalization techniques were tested, such as Min-Max, Zscore, logistic, and LogNormal... However, the models obtained the highest accuracy with the min-max normalization technique.

(ii) Construction of prediction models: This research employs five distinct algorithms for constructing prediction models: Support Vector Regression (SVR), Decision Trees (DT), Random Forests (RF), Light Gradient Boosting Machine (LGBM), and Linear Regression (LR). These algorithms are characterized by specific hyperparameters, initially set to predefined values. The appropriate configuration of these parameters significantly influences the efficacy of machine learning algorithms. To optimize model performance, researchers can engage in hyperparameter tuning. This process systematically adjusts these settings to identify the most effective combination for the given dataset and prediction task. By referring to the studies of Probst et al., 2019, Nematzadeh et al., 2022., the hyperparameters of the algorithms used in this study for tunning are presented as in Table 1. Hyperparameter tuning is a problem that has a long history. Grid search and random search are two commonly used methods for coarse-tuning the hyperparameters of machine learning models. Random search is a fundamental improvement on grid search that generally gives better results than grid search (Yu Tong and Zhu Hong, 2020). On the other hand, k-fold cross-validation is often utilized to verify the generalization of the models (Gorriz et al., 2024). K-fold cross-validation is usually used with values K = 5 or K = 10 (Pachouly et al., 2022). The value of K was chosen as 5 for this work. Therefore, this study uses the random search method with 5-fold cross-validation and RMSE as a loss function to optimize the hyperparameters of the models.

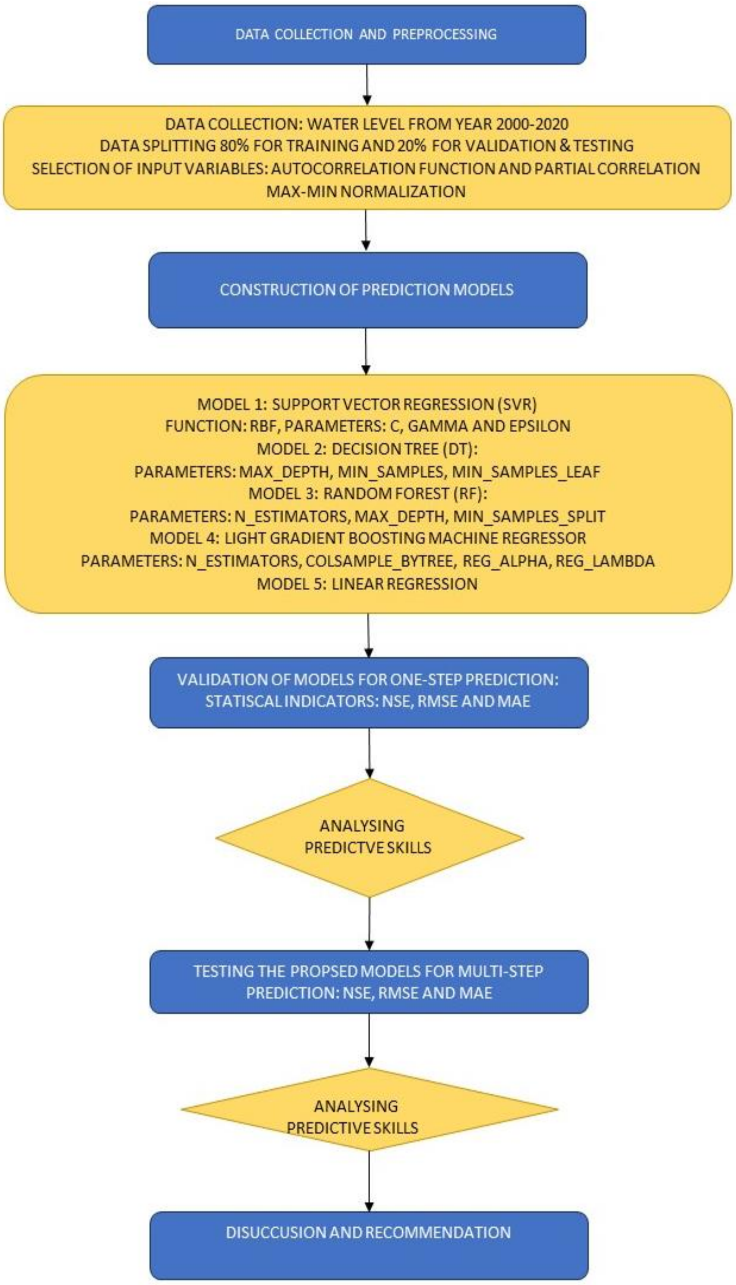
DATA COLLECTION AND PREPROCESSING
DATA COLLECTION: WATER LEVEL FROM YEAR 2000-2020
DATA SPLITTING 80% FOR TRAINING AND 20% FOR VALIDATION & TESTING
SELECTION OF INPUT VARIABLES: AUTOCORRELATION FUNCTION AND PARTIAL CORRELATION
MAX-MIN NORMALIZATION
CONSTRUCTION OF PREDICTION MODELS
MODEL 1: SUPPORT VECTOR REGRESSION (SVR)
FUNCTION: RBF, PARAMETERS: C, GAMMA AND EPSILON
MODEL 2: DECISION TREE (DT):
PARAMETERS: MAX_DEPTH, MIN_SAMPLES, MIN_SAMPLES_LEAF
MODEL 3: RANDOM FOREST (RF):
PARAMETERS: N_ESTIMATORS, MAX_DEPTH, MIN_SAMPLES_SPLIT
MODEL 4: LIGHT GRADIENT BOOSTING MACHINE REGRESSOR
PARAMETERS: N_ESTIMATORS, COLSAMPLE_BYTREE, REG_ALPHA, REG_LAMBDA
MODEL 5: LINEAR REGRESSION
VALIDATION OF MODELS FOR ONE-STEP PREDICTION:
STATISCAL INDICATORS: NSE, RMSE AND MAE
ANALYSING
PREDICTVE SKILLS
TESTING THE PROPSED MODELS FOR MULTI-STEP
PREDICTION: NSE, RMSE AND MAE
ANALYSING
PREDICTIVE SKILLS
DISUCCUSION AND RECOMMENDATION

*Figure 3*. The methodological procedure used in this study

*Table 1.* Hyperparameters of the algorithms

| Algorithm | Hyperparameter | Type | Lower | Upper |
|---|---|---|---|---|
| SVR | C | numeric | -10 | 10 |
| | gamma | numeric | -10 | 10 |
| | epsilon | numeric | 0 | 1 |
| DT | max_depth | integer | 1 | 30 |
| | min_samples_split | integer | 1 | 60 |
| | min_samples_leaf | integer | 1 | 60 |
| RF | n_estimators | integer | 1 | 2000 |
| | max_depth | integer | 1 | 30 |
| | min_samples_split | integer | 1 | 60 |
| LGBM | n_estimators | Integer | 1 | 2000 |
| | colsample_bytree | numeric | 0 | 1 |
| | reg_alpha | numeric | -10 | 10 |
| | reg_lambda | numeric | -10 | 10 |

(iii) Model validation: after obtaining all the model's hyperparameters, several statistical indices, namely NSE, RMSE MAE, $R^{2,}$ and walk-forward validation strategy, were used to evaluate the performance of the prediction models.

(iv) Model testing: after the validation, the proposed models (SVR, DT, RF, LGBM, and LR) were operated in a recursive mode to predict the water levels for one, three, five, and seven days ahead and compare with the observed values was undertaken.

### *3.1. Support vector regression*

Support Vector Regression (SVR) is a supervised learning algorithm initially introduced by Vapnik et al. (1995) that addresses classification and regression tasks. The SVR process operates in two primary stages: Initially, it employs a kernel function to map the input data into an N-dimensional feature space. Subsequently, it constructs a hyperplane within this space to partition the data. This hyperplane effectively segments the space into distinct regions, each encompassing a specific data category. The algorithm then decomposes data by calculating the distance between each data point in the N-dimensional space and the constructed hyperplane. This approach allows SVR to effectively model complex relationships in the data while maintaining good generalization capabilities (Dehghani et al., 2020; Zhang et al., 2014).

SVR uses statistical principles to define boundaries between domains, reduce generalization errors relative to training errors, and improve model convergence acceleration. Kernel functions are used to convert data from two-dimensional space to multidimensional space. This spatial transformation makes it possible to define input-output relationships from the most complex to the simplest (Essam et al., 2022). The accuracy of the SVR algorithm can be improved by adjusting parameters such as C Gamma. C plays a role in removing anomalous data points during SVR model optimization. If this parameter has a significant value, the optimization process chooses a hyperplane that best separates all the data points. The gamma parameter determines the number of data points to construct the separating hyperplane. With small gamma values, data points far from the median are used in the median calculation.

### *3.2. Decision Tree*

The concept of decision trees has been present for a long time in statistics and machine learning. A decision tree is a supervised learning model widely used for classification and regression tasks. Unlike other supervised learning models, DT does not have a prediction equation; instead, we seek a decision tree that predicts the training data well and applies it to make predictions on the test set. A decision tree consists of a series of nodes connected by edges. In regression problems, each node represents a variable and a value of that variable. Each leaf node represents the final predictions. When inputting a new sample, the predicted value is calculated by traversing the tree from the root node to a leaf node and taking the prediction value at that node. DT is advantageous for its interpretability, yet it may be prone to overfitting, especially when the tree is too deep, or the number of leaf nodes is unlimited.

### *3.3. Random Forest*

The Random Forest algorithm, developed by Leo Breiman in 2001, is a robust and popular supervised learning method applicable to classification and regression tasks. This ensemble technique derives its name from its structure, consisting of multiple decision trees built on various subsets of the original dataset, obtained through bootstrap sampling. The algorithm operates in three main phases:

- Subset Selection: The model randomly selects a subset of data points and features from the original dataset for each decision tree. Specifically, it chooses n random records and m features from a dataset containing k total records.
- Tree Construction: The algorithm then builds individual decision trees for these randomly selected samples.
- Output Aggregation: Each decision tree produces its prediction. The final output is determined by combining these individual predictions. Classification tasks are typically done through majority voting, where the most common prediction among all trees is selected. For regression tasks, the average of all three predictions is usually used.

This approach allows Random Forest to leverage the strength of multiple decision trees while mitigating individual tree biases, resulting in a powerful and versatile machine-learning model.

### *3.4. Light Gradient Boosting Machine Regressor*

LightGBM (LGBM) is an advanced machine learning algorithm developed by Microsoft that falls under gradient boosting models. Its core concept is derived from the Gradient Tree Boosting (GTB) model, introduced by Friedman et al. in 2000. LGBM is renowned for its speed, distributed processing capabilities, and high performance, making it suitable for various machine-learning tasks, including ranking, classification, and regression (Fan J. et al., 2019).

The LGBM process begins with data preparation, where the training dataset is separated into input features and target values for regression purposes. The algorithm recommends using target values and metric characteristics. Initial parameters such as learning rate, tree count, maximum depth, and feature fraction are set, which can be fine-tuned to optimize performance. The model's construction and training involve creating a series of decision trees. Each tree is built using a gradient-based optimization technique to minimize loss functions. The model iteratively expands its ensemble of trees, adjusting predictions based on the loss function's gradient. Once trained, the model can be applied to new data points for prediction. LGBM employs a weighted sum approach to combine predictions from all trees in the ensemble. These weights are determined during training based on the loss function's gradients.

LGBM distinguishes itself through its innovative leaf-wise tree growth approach, departing from the conventional level-wise strategies employed by many algorithms. This method enables the tree to develop along the most promising paths, yielding a more profound yet more streamlined structure contributing to reduced training errors. LGBM incorporates a histogram-based technique for determining optimal splits to enhance computational efficiency. This approach discretizes continuous variables into bins rather than processing individual values. The result significantly accelerates the training phase and reduces memory requirements. These features collectively enable LGBM to handle complex datasets with improved speed and efficiency while maintaining high predictive accuracy. By balancing sophisticated modeling techniques with computational pragmatism, LGBM has established itself as a powerful and versatile tool in the machine learning domain.

### *3.5. Linear regression*

The linear regression model is the simplest and most basic statistical and predictive modeling model. It helps provide a linear relationship function between the predictors and the predictand. For data time-series predictive modeling, it can be described mathematically by the following equation:

$$x(t+k+1) = \beta_0 + \beta_1 x(t) + \cdots + \beta_k x(t+k) \quad (1)$$

where x(t+k+1) is the one-step-ahead predictand at time step t+k+1, whose value needs to be predicted; x(t) is the predictor observed at time step t; k is the sequence length; $\beta_i$ (i=0, k) are regression coefficients that need to be estimated from historical observations.

The merit of the LR model is that it is inexpensive to develop, and its structure and result are easily interpreted. The drawback is that it does not consider the nonlinear relationship between the predictand and predictors and the possible interactions among predictors. In addition, the assumption of a Gaussian distribution of predictive errors is not always valid.

### *3.6. Performance metrics*

In this study, the performance of the water level prediction model was evaluated using Nash-Sutcliffe Efficiency (NSE), coefficient of determination ($R^2$), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Previous studies have extensively applied these statistical indices (Nguyen, 2023; Nguyen et al., 2023b).

NSE, $R^{2,}$ and RMSE are considered popular indexes for assessing the quality of hydrological modeling. It measures the errors between the simulation values of the models and the observation values (Adnan et al., 2020; Adnan et al., 2021; Liu et al., 2020; Xu et al., 2022; Moriasi et al., 2015; Manh Van Le et al., 2023; Dam Duc Nguyen et al., 2023).

$$NSE = 1 - [\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{obs} - \overline{Y^{obs}}\right)^2}] \quad (2)$$

$$R^2 = \left[\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - \overline{Y^{obs}}\right)\left(Y_i^{sim} - \overline{Y^{sim}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i^{obs} - \overline{Y^{obs}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i^{sim} - \overline{Y^{sim}}\right)^2}}\right]^2 \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i^{sim} - Y_i^{obs})^2}{n}} \quad (4)$$

Where $Y_i^{obs}$ is the value of the observation water level at time i; $Y_i^{sim}$ is the prediction value at time i; $\overline{Y^{obs}}$ is the mean value of the observed water level; $\overline{Y^{sim}}$ is the mean value of the predicted water level; n is length of the time series used for evaluation.

MAE measures the average of errors in a set of predictions, regardless of their direction. It is the sample mean of the absolute difference between the prediction and the actual number of observations, where all differences are weighted equally (Kisi, 2010). The formula for calculating the MAE is as follows:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \quad (5)$$

Where $e_i$ is the average of the absolute errors; $y_i$ is the calculated/simulated water level at time i. $x_i$ is the actual water level measured at time i.

## 4. Results

### *4.1. Selection of input variables*

Regarding building models for forecasting time series with seasonal variability, determining sequence length (also called seasonal lag) is a nontrivial task. As one may infer from Fig. 2, bi-weekly lunar tidal cycles influence the water levels at Cao Lanh station. That means there are two spring and two neap tides in a lunar month. To confirm the existence of this seasonal cycle, plots of the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) of the three-time series (HmaxCL, HminCL, HtbCL) were analyzed. Fig. 4 represents the plots of ACF and PACF of daily maximum water level time series at Cao Lanh gauging station. The ACF plot shows that the time series has a seasonal cycle with an approximate value of 15 days.

On the other hand, the PACF plot shows the association between the observed value at time t and the observed value at time t-k (neglecting the observed values in between). Theoretically, any value of k that has a value of PACF lying outside of the significance bounds (shaded area) should be considered in a machine-learning model. However, the PACF value has the highest negative values at a window of 15 days. Therefore, the value of 15 is chosen as the sequence length.

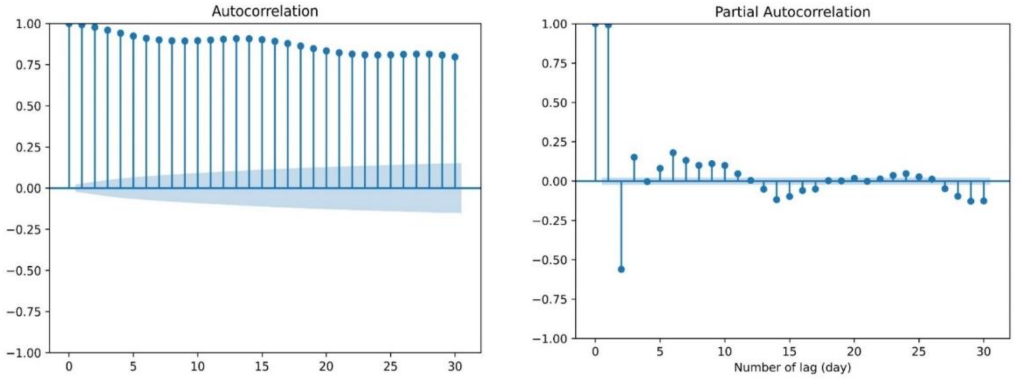

*Figure 4*. Plots of ACF (upper) and PACF (lower) for daily maximum water level time series

### *4.2. Validation of models for 1-day-ahead prediction*

Table 2 presents the performance metrics of the models proposed to predict the 1-day-ahead water level in the Cao Lanh station. The results show that in the case of HmaxCL, the SVR model outperforms the other models with (NSE, $R^2$, RMSE, and MAE) values of (0.967, 0.964, 6.58, 5). The following models, arranged in decreasing order of predictive skills, are LGBM (0.966, 0.962, 6.71, 5), LR (0.966, 0.960, 6.72, 5), DT (0.955, 0.956, 7.69, 6), RF (0.953, 0.953, 7.86, 6). Concerning HminCL, a similar order of model performances was obtained, except that LGBM is as good as the SVR. Concerning HtbCL, the SVR model retains its first place, followed by LR, LGBM, DT, and RF. In general, in all three cases (HmaxCL, HminCL, and HtbCL), the SVR model outperforms the other models for 1-day-ahead water level prediction. LGBM and LR perform equally well, and the predictive skill of the DT model is a little better than that of the worst, the RF model. If only MAE were considered, the five models could be separated into two groups concerning their predictive skills. The first group consists of SRV, LGBM, and LR models; the second includes DT and RF models. Furthermore, the predictive skills of all models concerning three data types are highest with daily mean water level (HtbCL). Meanwhile, the daily minimum water level (HminCL) is the most difficult to predict.

Figure 5 presents temporal patterns of the 1-day-ahead water level prediction results using HmaxCL, HminCL, and HtbCL and the corresponding observed values during the validation phase. The results show that, in general, the predicted water level time series closely follows the observed water level time series in the cases of HmaxCL, HminCL, and HtbCL for all models but RF. However, the predictive skills of all five models are not much different, as indicated by the similarities in the corresponding evaluation metrics in

Table 1. Models' predictions well captured the lunar cycle of all observed water level time series. The validation results in Fig. 5 also reveal that all five models are neither overfitted nor underfitted. Therefore, all five models are utilized in the testing phase to see if they can make accurate multi-step-ahead predictions.

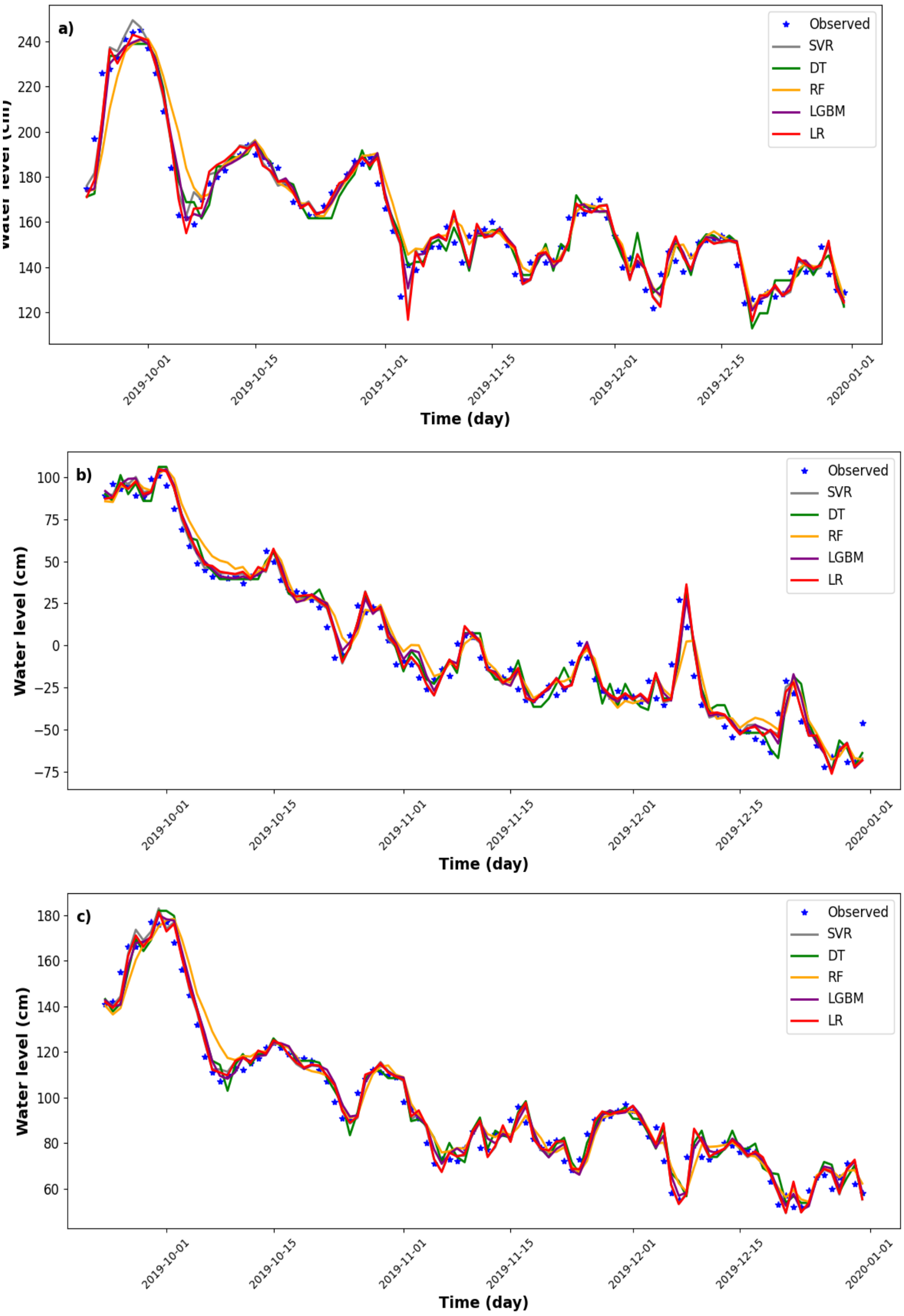

*Figure 5*. The 1-day-ahead water level predictions for HmaxCL (a), HminCL (b), and HtbCL (c) produced by SVR, DT, RF, LGBM, LR in the validation phase

*Table 2*. The performance metrics of the proposed models in the validation phase

| | **NSE** | **$R^2$** | **RMSE** | **MAE** |
|---|---|---|---|---|
| SVR model | | | | |
| HmaxCL | 0.967 | 0.964 | 6.58 | 5 |
| HminCL | 0.977 | 0962 | 8.63 | 6 |
| HtbCL | 0.987 | 0.979 | 4.72 | 4 |
| DT model | | | | |
| HmaxCL | 0.955 | 0.956 | 7.69 | 6 |
| HminCL | 0.972 | 0.948 | 9.53 | 7 |
| HtbCL | 0.980 | 0.971 | 5.73 | 4 |
| RF model | | | | |
| HmaxCL | 0.953 | 0.953 | 7.86 | 6 |
| HminCL | 0.972 | 0.950 | 9.49 | 7 |
| HtbCL | 0.978 | 0.970 | 6.10 | 5 |
| LGBM model | | | | |
| HmaxCL | 0.966 | 0.962 | 6.71 | 5 |
| HminCL | 0.977 | 0.959 | 8.61 | 6 |
| HtbCL | 0.985 | 0.976 | 5.01 | 4 |
| LR model | | | | |
| HmaxCL | 0.966 | 0.960 | 6.72 | 5 |
| HminCL | 0.976 | 0.962 | 8.77 | 6 |
| HtbCL | 0.986 | 0.979 | 4.78 | 4 |

## 5. Testing of models for one, three, five, and seven-day-ahead predictions

After validating the proposed models, these models were used to predict the water level for one, three, five, and seven days lead times using independent testing data series. Table 3 presents the performance of the models proposed to predict the water level in the Cao Lanh station. In general, model predictive accuracy decreases as the lead time increases (Table 3 and Fig. 6).

*Table 3.* Performance of the models measured by MAE for one, three, five, and seven days lead times

| **Time series** | **1-day forecast** | **3-day forecast** | **5-day forecast** | **7-day forecast** |
|---|---|---|---|---|
| SVR model | | | | |
| HmaxCL | 5 | 8 | 11 | 14 |
| HminCL | 7 | 10 | 14 | 17 |
| HtbCL | 4 | 6 | 9 | 11 |
| DT model | | | | |
| HmaxCL | 6 | 8 | 10 | 11 |
| HminCL | 8 | 11 | 13 | 16 |
| HtbCL | 5 | 8 | 11 | 13 |
| RF model | | | | |
| HmaxCL | 6 | 8 | 12 | 14 |
| HminCL | 8 | 11 | 14 | 17 |
| HtbCL | 5 | 7 | 10 | 12 |
| LGBM model | | | | |
| HmaxCL | 5 | 8 | 12 | 14 |
| HminCL | 7 | 10 | 15 | 17 |
| HtbCL | 4 | 6 | 9 | 12 |
| LR model | | | | |
| HmaxCL | 5 | 8 | 12 | 14 |
| HminCL | 7 | 9 | 13 | 16 |
| HtbCL | 4 | 6 | 9 | 11 |

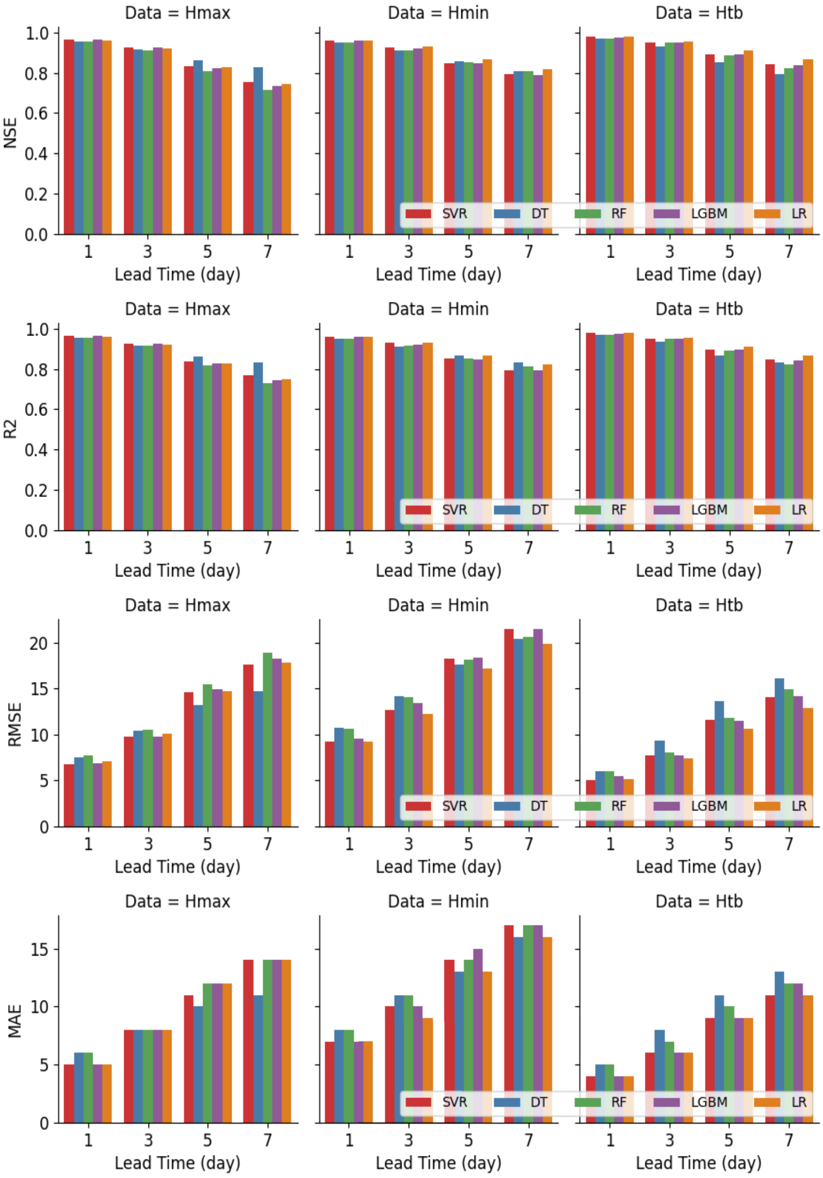

*Figure 6*. Comparisons of performance metrics of the five prediction models with different lead times

Particularly, for the SVR model, in the case of HmaxCL, the value of MAE increases from 5 in 1-day-ahead prediction to 8, 11, and 14 for 3-day-ahead, 5-day-ahead, and 7-day-ahead predictions, respectively. For HminCL, the value of MAE increases from 7 to 10, 14, and 17 for 1-day-ahead, 3-day-ahead, 5-day-ahead, and 7-day-ahead predictions, respectively; and increases from 4 in 1-day-ahead prediction to 6, 9 and 11 for HtbCL for 3-day-ahead, 5-day-ahead and 7-day-ahead predictions, respectively. Concerning the 1-day-ahead and 3-day-ahead predictive skills results similar to those of the corresponding ones in the validation phase were obtained. In other words, two models can be distinguished: the first group consists of SRV, LGBM, and LR models, and the second group includes DT and RF models. However, this picture becomes different when longer lead time predictions (i.e., 5- and 7-day lead times) are considered. The DT outperforms the other models in predicting HmaxCL and HminCL. In contrast, for predicting HtbCL, the SRV, LR, and LGBM retain their first places. Concerning predictability, the daily minimum water level remains the most difficult to predict, and the daily mean water level is the easiest.

Figure 7 visually presents the temporal dynamics of the 7-day-ahead water level predictions for HmaxCL, HminCL, and HtbCL using the SVR, DT, RF, LGBM, and LR models. The forecasted water level using the DT model closely follows the observed water level in the three use cases of HmaxCL, HminCL, and HtbCL. This is the best model to reproduce the dynamic pattern of the observed water levels, especially for the peaks and troughs. That is the reason why the predictive skill of the DT model is highest concerning daily maximum and minimum water level predictions with longer lead times (i.e., 5 and 7 days ahead).

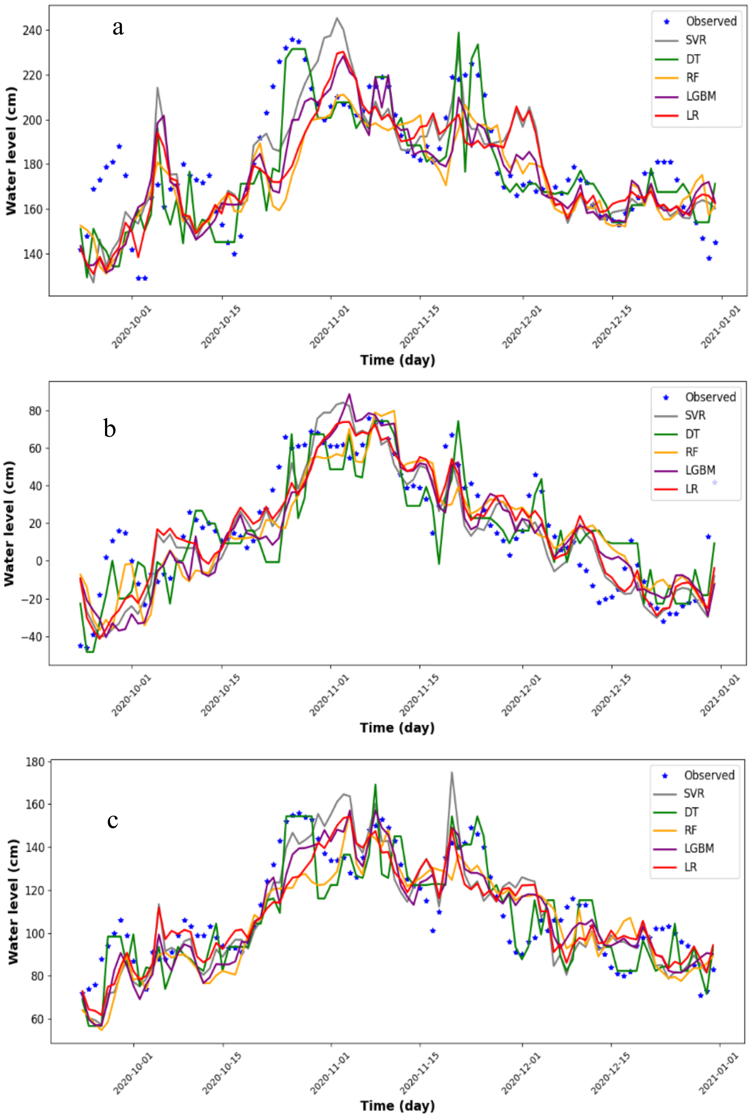

*Figure 7.* The 7-day-ahead water level predictions for HmaxCL (a), HminCL (b), and HtbCL (c) produced by SVR, DT, RF, LGBM, LR in the testing phase

## 6. Discussions

### 6.1. Variations in machine learning performance across different study regions

Machine learning (ML) application in water level forecasting has been extensively studied across various regions, revealing significant differences in model performance and optimal input selection for each case study. Our research on the Tien River in the Mekong Delta identified Support Vector Regression (SVR) as the most suitable model. However, other studies have yielded different results: a combination of statistical ML and ARIMA models proved most effective for the Red River (Thi-Thu-Hong Phan and Xuan Hoai Nguyen 2020)., multiple linear regression performed best for Lake Erie (Qi Wang and Song Wang, 2020), Gaussian Process Regression showed superior results for the Durian Tunggal River (Ahmed et al., 2022), and Random Forest was the preferred model for the Upo Wetland in South Korea (Choi et al., 2019). These diverse outcomes underscore the importance of conducting region-specific studies to determine the most appropriate model for each area. The variability in results demonstrates that no single ML model is universally optimal for all regions, highlighting the necessity of tailored approaches that account for local hydrological conditions and data characteristics. This emphasizes the critical need for comparative analysis when developing water level prediction systems for new areas rather than assuming that a model successful in one region performs equally well in another.

### 6.2. Data quality/quantity, overfitting issue

Predicting water level is considered one of the complicated tasks in tidal rivers where the complicated topography, hydrodynamics, and human interventions co-exist. Hydrodynamic modeling often requires high-quality, detailed field observed data (Kong et al., 2023; Li et al., 2023; Siddique-E-Akbor et al., 2011). However, hydrological station networks have been poorly distributed, particularly in developing countries like Vietnam. Therefore, statistical models and machine learning have received attention from the scientific community. This study developed and compared statistical and machine learning models to predict the water level in the Cao Lanh station in the Tien River. This study's results are considered an alternative tool to support decision-makers or managers in distributing water resources to develop the economy.

This study used RMSE as the function objective for SVR, DT, LGBM, and LR or the machine learning model in general. Therefore, considering the overfitting problem is necessary to improve the prediction model's quality (Nguyen et al., 2020; Van Phong et al., 2020). We used several techniques, such as adjusting the model parameters and limiting the search boundary to reduce its effects. Several studies have pointed out that including more training data, such as rainfall, evaporation, and river flow, can reduce the effects of the overfitting problem. However, collecting enough training data is a big challenge for the hydrological field because the available data is still limited (Nguyen et al., 2020).

### 6.3. Model complexity

Model complexity is a critical consideration when data availability is limited. Utilizing overly complex models heightens the risk of overfitting. This issue arises when a model captures not only the underlying patterns but also the biases present in the training dataset. Therefore, these models cannot be generalized to new datasets. In addition, the use of models that are too simple leads to the problem of underfitting. This problem occurs when the models are too simple and cannot learn the entire training data. Therefore, model selection is essential and challenges the scientific community

(Tran and Kim, 2022). This study tried five machine learning and statistical models: SVR, DT, LGBM, RF, and LR. SVM was justified as a more appropriate algorithm than the RF, DT, LGBM, and LR model. Besides the advantage of being efficient in high dimensions, SVR is also efficient in cases where the dimension of the space is larger than the number of training samples.

Moreover, for the decision, SVM does not use all the training samples, only a part (the support vectors). Consequently, these algorithms require less memory (Gu et al., 2015; Ma et al., 2003). The RF model was second class because RF can effectively handle regression and classification tasks with high accuracy. Known for its ability to estimate missing values, this method maintains good accuracy even in incomplete data. Additionally, it facilitates the assessment of the importance or contribution of variables to the model, making the analysis more intuitive and informative (Ao et al., 2019; Langsetmo et al., 2023). The DT model was third class because DT facilitates decision-making by adopting a structured and targeted approach, thus representing a primary advantage of this method.

Furthermore, Building a decision tree via DT is a quick process that requires few resources, making this tool particularly effective for data analysis (Almuallim et al., 2002; Quinlan, 1987). LightGBM is considered more efficient than other gradient-boosting algorithms on decision trees. Because it generates more complex decision trees through a leaf-split rather than a level-split approach, a crucial factor in achieving better accuracy, this method can sometimes result in overfitting. Therefore, in the proposed model, LightGBM was ranked fourth (Li et al., 2024; Wang et al., 2021). The accuracy of the LR model is less than that of the other two models because the LR method directly uses previously observed water level data to train the model and forecast water levels for the following days. Therefore, this method is limited in solving nonlinear water-level forecasting problems (Jadhav and Channe, 2016).

### *6.4. Future study for climate change and human activities*

This study successfully predicted the water level in the Cao Lanh Station in the Dong Thap province in Vietnam. In the context of climate change and changing human activities such as dam construction, machine learning/statistical models can effectively predict these contexts. Their prediction skills can be improved if data related to climate change and changes in human activities are sufficient. However, data collection in developing countries, particularly Vietnam, is complicated due to data availability and sharing policies. Moreover, one of the significant challenges using machine learning/model statistics is the extrapolation problem, i.e., the models cannot predict the water level outside of the training data. Various studies have pointed out that integrating machine learning/statistical and optimization models is crucial and widely applied to solve these problems (Nguyen et al., 2023a). In future research, we integrate the individual models with the advanced optimization algorithms to improve the skill of water level prediction.

## 7. Conclusions

This study presents a comprehensive comparative analysis of various machine learning (ML) models for water level forecasting in the Tien River region of the Mekong Delta. The research yielded several significant findings. The proposed models - Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT), Light Gradient Boosting Machine (LGBM), and Logistic Regression (LR) - demonstrated successful water level predictions for three, five, and seven days in advance. Among these, the SVR model consistently outperformed the others across all scenarios, followed by RF, DT, and LGBM.

The study underscores the critical importance of input variable selection in water level prediction. Results indicate that HmaxCL (maximum water level) yields the most accurate predictions across all five models, followed by HtbCL (average water level) and HminCL (minimum water level). This case study highlights the significant potential of machine learning approaches in water level prediction. These methods are valuable tools for decision-makers and water resource managers, particularly in climate change and upstream dam construction. Future studies should consider incorporating additional input variables, such as rainfall and evaporation data, to enhance prediction accuracy.

In conclusion, this research not only demonstrates the effectiveness of machine learning in water level forecasting but emphasizes the importance of model selection and input variable choice. The findings provide a solid foundation for improved water resource management in the Mekong Delta region and offer promising avenues for future research in hydrological forecasting.

## References

Adnan 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 hydrometeorological data as inputs. Journal of Hydrology, 586, 124371. https://doi.org/10.1016/j.jhydrol.2019.124371.

Adnan R.M., Liang Z., Parmar K.S., Soni K., Kisi O., 2021. Modeling monthly streamflow in mountainous basin by MARS., GMDH-NN and DENFIS using hydroclimatic data. Neural Computing and Applications, 33, 2853–2871. https://doi.org/10.1007/s00521-020-05164-3.

Ahmed A.N., Yafouz A., Birima A.H., Kisi O., Huang Y.F., Sherif M., Sefelnasr A., El-Shafie A., 2022. Water level prediction using various machine learning algorithms: A case study of Durian Tunggal river., Malaysia. Engineering Applications of Computational Fluid Mechanics, 16, 422–440. https://doi.org/10.1080/19942060.2021.2019128.

Almuallim H., Kaneda S., Akiba Y., 2002. Development and applications of decision trees. Expert Systems. Elsevier, 53–77. https://doi.org/10.1016/B978-012443880-4/50047-8.

Ao Y., Li H., Zhu L., Ali S., Yang Z., 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174, 776–789. https://doi.org/10.1016/j.petrol.2018.11.067.

Baek S-S., Pyo J., Chun J.A., 2020. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. Water, 12, 3399. https://doi.org/10.3390/w12123399.

Choi C., Kim J., Han H., Han D., Kim H.S., 2019. Development of water level prediction models using machine learning in wetlands: A case study of Upo wetland in South Korea. Water, 12, 93. https://doi.org/10.3390/w12010093.

Chua L.H.C., 2012. Considerations for data-driven and physically based hydrological models in flow forecasting. IFAC Proc, 45, 1025–1030.

Dam Duc Nguyen, Hai Phu Nguyen, Dung Quang Vu, Indra Prakash, Binh Thai Pham, 2023. Using GA-ANFIS machine learning model for forecasting the load bearing capacity of drivenpiles. Journal of Science and Transport Technology, JSTT 2023, 3(2), 26–33. https://doi.org/10.58845/jstt.utt.2023.en.3.2.26-33.

Dehghani R., Torabi Poudeh H., Younesi H., Shahinejad B., 2020. Daily streamflow prediction using support vector machine-artificial flora SVM-AF hybrid model. Acta Geophysica, 68, 1763–1778. https://doi.org/10.1007/s11600-020-00472-7.

DHI Water and Environment, 1999. MIKE 11 Reference Manual.

Do H.X., Le M.H., Pham H.T., Le T.H., Nguyen Q.B., 2022. Identifying hydrologic reference stations to understand changes in water resources across Vietnam - a data-driven approach. Vietnam Journal of Earth Sciences, 44(1), 144–164. https://doi.org/10.15625/2615-9783/16980.

Essam Y., Huang Y.F., Ng J.L., Birima A.H., Ahmed A.N., El-Shafie A., 2022. Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms. Scientific Reports 12, 1–26. https://doi.org/10.1038/s41598-022-07693-4.

Fan J., Ma X., Wu L., Zhang F., Yu X., Zeng W., 2019. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural Water Management, 225, 105758. https://doi.org/10.1016/j.agwat.2019.105758.

Ghaith M., Siam A., Li Z., El-Dakhakhni W., 2019. Hybrid hydrological data-driven approach for daily streamflow forecasting. Journal of Hydrologic Engineering, 25(2), 04019063. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001866.

Gorriz J.M., Segovia F., Ramirez J., Ortiz A., Suckling J., 2024. Is k-fold cross validation the best model selection method for machine learning? arXiv:2401.16407 [stat.ML]. https://doi.org/10.48550/arXiv.2401.16407.

Gu B., Sheng VS., Wang Z., Ho D., Osman S., Li S., 2015. Incremental learning for v-support vector regression. Neural networks, 67, 140–150. https://doi.org/10.1016/j.neunet.2015.03.013.

Herath M., Jayathilaka T., Hoshino Y., Rathnayake U., 2023. Deep machine learning-based water level prediction model for Colombo flood detention area. Applied Sciences, 13, 2194. https://doi.org/10.3390/app13042194.

Hicks F.E., Peacock T., 2005. Suitability of HEC RAS for flood forecasting. Canadian Water Resources Journal, 30(2), 159–174. https://doi.org/10.4296/cwrj3 002159.

Jadhav SD., Channe H., 2016. Comparative study of K-NN., naive Bayes and decision tree classification techniques. International Journal of Science and Research IJSR, 5, 1842–1845. https://www.ijsr.net/getabstract.php?paperid=NOV153131.

Kim B., Sanders B.F., Famiglietti J.S., Guinot V., 2015. Urban flood modeling with porous shallow-water equations: a case study of model errors in the presence of anisotropic porosity. J. Hydrol, 523, 680–692.

Kim D., Han H., Wang W., Kim H.S., 2022. Improvement of deep learning models for river water level prediction using complex network method. Water, 14, 466. https://doi.org/10.3390/w14030466.

Kisi O., 2010. Wavelet regression model for short-term streamflow forecasting. Journal of hydrology, 389, 344–353. https://doi.org/10.1016/j.jhydrol.2010.06.013.

Kong L., Li Y., Yuan S., Li J., Tang H., Yang Q., Fu X., 2023. Research on water level forecasting and hydraulic parameter calibration in the 1D open channel hydrodynamic model using data assimilation. Journal of Hydrology, 625, 129997. https://doi.org/10.1016/j.jhydrol.2023.129997.

Langsetmo L., Schousboe J.T., Taylor B.C., Cauley J.A., Fink H.A., Cawthon P.M., Kado D.M., Ensrud K.E., Group OFiMR., 2023. Advantages and disadvantages of random forest models for prediction of hip fracture risk versus mortality risk in the oldest old. JBMR plus, 7, e10757.

Leo Breiman., 2001. Random Forest. https://link.springer.com/article/10.1023/A:1010933404324.

Li G., Zhu H., Jian H., Zha W., Wang J., Shu Z., Yao S., Han H., 2023. A combined hydrodynamic model and deep learning method to predict water level in ungauged rivers. Journal of Hydrology, 625, 130025. https://doi.org/10.1016/j.jhydrol.2023.130025.

Li L., Jun K.S., 2022. A hybrid approach to improve flood forecasting by combining a hydrodynamic flow model and artificial neural networks. Water, 14, 1393. https://doi.org/10.3390/w14091393.

Li S., Dong X., Ma D., Dang B., Zang H., Gong Y., 2024. Utilizing the LightGBM Algorithm for Operator User Credit Assessment Research. arXiv preprint arXiv, 240314483. https://doi.org/10.48550/arXiv.2403.14483.

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. 10.1109/ACCESS.2020.2993874.

Ma J., Theiler J., Perkins S., 2003. Accurate on-line support vector regression. Neural computation, 15, 2683–2703. https://doi.org/10.1162/089976603322385117.

Manh Van Le, Indra Prakash, Dam Duc Nguyen, 2023. Predicting load-deflection of composite concrete bridges using machine learning models. Journal of Science and Transport technology, JSTT, 3(4), 44–52. https://doi.org/10.58845/jstt.utt.2023.en.3.4.44-52.

Moriasi D.N., Gitau M. W., Pai N., Daggupati P., 2015. Hydrologic and water quality models: performance measures and evaluation criteria. Transactions of the ASABE (American Society of Agricultural and Biological Engineers), 58(6), 1763–1785. http://dx.doi.org/10.13031/trans.58.10715.

Mosavi A., Ozturk P., Chau K.-W., 2018. Flood prediction using machine learning models: literature review. Water, 10, 1536.

Nanda T., Sahoo B., Chatterjee C., 2019. Enhancing real-time streamflow forecasts with wavelet-neural network based error-updating schemes and ECMWF meteorological predictions in Variable Infiltration Capacity model. Journal of Hydrology, 575, 890–910. https://doi.org/10.1016/j.jhydrol.2019.05.051.

Narsimlu B., Gosain A.K., Chahar B.R., Singh S.K., Srivastava PK., 2015. SWAT model calibration and uncertainty analysis for streamflow prediction in the Kunwari River Basin., India., using sequential uncertainty fitting. Environmental Processes, 2, 79–95. https://doi.org/10.1007/s40710-015-0064-8.

Nematzadeh S., Kiani F., Torkamanian-Afshar M., Aydin N., 2022. Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases. Computational Biology and Chemistry, 97, 107619. https://doi.org/10.1016/j.compbiolchem.2021.107619.

Nguyen H.D., 2023. Daily streamflow forecasting by machine learning in Tra Khuc River in Vietnam. Vietnam Journal of Earth Sciences 45(1), 82–97. https://doi.org/10.15625/2615-9783/17914.

Nguyen H.-D., Pham V.-D., Nguyen Q.-H., Pham V.-M., Pham M.H., Vu V.M., Bui Q.-T., 2020. An optimal search for neural network parameters using the Salp swarm optimization algorithm: a landslide application. Remote Sensing Letters, 11, 353–362. https://doi.org/10.1080/2150704X.2020.1716409.

Nguyen H.D., Van C.P., Do A.D., 2023a. Application of hybrid model-based deep learning and swarm-based optimizers for flood susceptibility prediction in Binh Dinh province. Vietnam Earth Science Informatics, 1–21. https://doi.org/10.1002/gj.4885.

Nguyen H.D., Van C.P., Nguyen Q.-H., Bui Q.-T., 2023b. Daily streamflow prediction based on the long short-term memory algorithm: a case study in the Vietnamese Mekong Delta. Journal of Water and Climate Change, 14, 1247–1267. https://doi.org/10.2166/wcc.2023.419.

Nguyen P.N.B., Phan V.T., Trinh T.L., Tangang F.T., Cruz F., Boon S.J., Juneng L., Chung J.X., Aldrian, 2022. Projected future changes in drought characteristics over Southeast Asia. Vietnam Journal of Earth Science, 44(1), 127–143. https://doi.org/10.15625/2615-9783/16974.

Özdoğan-Sarıkoç G., Dadaser-Celik F., 2024. Physically based vs. data-driven models for streamflow and reservoir volume prediction at a data-scarce semi-arid basin. Environ Sci Pollut Res, 31, 39098–39119. https://doi.org/10.1007/s11356-024-33732-w.

Pachouly J., Ahirrao S., Kotecha K., Selvachandran G. Abraham A., 2022. A systematic literature review on software defect prediction using artificial intelligence: datasets, data validation methods, approaches, and tools. Engineering Applications of Artificial Intelligence, 111, 104773. https://doi.org/10.1016/j.engappai.2022.104773.

Pan M., Zhou H., Cao J., Liu Y., Hao J., Li S., Chen C.-H., 2020. Water level prediction model based on GRU and CNN. Ieee Access, 8, 60090–60100. 10.1109/ACCESS.2020.2982433.

Park K., Jung Y., Seong Y., Lee S., 2022. Development of deep learning models to improve the accuracy of water levels time series prediction through multivariate hydrological data. Water, 14, 469. https://doi.org/10.3390/w14030469.

Peters N.E., Freer J., Beven K., 2003. Modelling hydrologic responses in a small forested catchment Panola Mountain., Georgia., USA: a comparison of the original and a new dynamic TOPMODEL. Hydrological Processes, 17, 345–362. https://doi.org/10.1002/hyp.1128.

Phan T.-T.-H., Nguyen X.H., 2020. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. Advances in Water Resources, 142, 103656. https://doi.org/10.1016/j.advwatres.2020.103656.

Probst P., Boulesteix A.-L., Bischl B., 2019. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research, 20, 1934–1965.

Qi Wang, Song Wang, 2020. Machine learning-based water level prediction in lake Erie. Water, 12(100), 2654. https://doi.org/10.3390/w12102654.

Quinlan J.R., 1987. Generating production rules from decision trees. ijcai. Citeseer, 304–307.

Sampurno J., Vallaeys V., Ardianto R., Hanert E., 2022. Integrated hydrodynamic and machine learning models for compound flooding prediction in a data-scarce estuarine delta, Nonlin. Processes Geophys., 29, 301–315. https://doi.org/10.5194/npg-29-301-2022.

Siddique-E-Akbor A., Hossain F., Lee H., Shum C., 2011. Inter-comparison study of water level estimates derived from hydrodynamic–hydrologic model and satellite altimetry for a complex deltaic environment. Remote Sensing of Environment, 115, 1522–1531.

Thi Thu Hong Phan, Xuan Hoai Nguyen, 2020. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. Advances in Water Resources, 142, 103656.
https://doi.org/10.1016/j.advwatres.2020.103656.

Tran V.N., Kim J., 2022. Robust and efficient uncertainty quantification for extreme events that deviate significantly from the training dataset using polynomial chaos-kriging. Journal of Hydrology 609, 127716.
https://doi.org/10.1016/j.jhydrol.2022.127716.

Van Phong T., Ly H.-B., Trinh P.T., Prakash I., Btjvjoes P., 2020. Landslide susceptibility mapping using Forest by Penalizing Attributes FPA algorithm based machine learning approach. Vietnam J. Earth Sci., 42, 237–246. https://doi.org/10.15625/0866-7187/42/3/15047.

Vapnik V., Guyon I., Hastie T., 1995. Support vector machines. Mach Learn, 20, 273–297. https://link.springer.com/article/10.1007/BF00994018.

Vijendra Kumar, Naresh Kedam, Kul Vaibhav Sharrma, Darshan J. Mehta, Tommaso Caloiero, 2023. Advanced Machine Learning Techniques to Improve Hydrological Prediction: A comparative Analysis of Streamflow Prediction Models. Water, 15(14), 2572. https://doi.org/10.3390/w15142572.

Vinh Ngoc Tran, Jongho Kim, 2019. Quantification of predictive uncertainty with a metamodel: toward more eficient hydrologic simulations. Stochastic environmental research and risk assessment, 33, 1453–1476. https://doi.org/10.1007/s00477-019-01703-0.

Vinh Ngoc Tran, Jongho Kim, 2022. Robust and efficient uncertainty quantification for extreme events that deviate significantly from the training dataset using polynomial chaos-kriging. Journal of Hydrology, 609, 127716. https://doi.org/10.1016/j.jhydrol.2022.127716.

Vinh Ngoc Tran, M. Chase Dwelle, Khachik Sargsyan, Valeriy Y. Ivanov, Jongho Kim, 2020. A novel modeling framework for computationally efficient and accurate real time ensemble flood forecasting with uncertainty quantification. Advancing earth and space sciences. Water Resources Research, 56(3), e2019WR025727.
https://doi.org/10.1029/2019WR025727.

Vinh Ngoc Tran, Valeriy Y. Ivanov, Donghui Xu, Jongho Kim, 2023a. Closing in on hydrologic predictive accuracy: combining the strengths of high-fidelity and physics-agnostic models. Advancing earth and space sciences, Geophysical research letters, 50(17), e2023GL104464. https://doi.org/10.1029/2023GL104464.

Vinh Ngoc Tran, Valeriy Y. Ivanov, Giang Tien Nguyen, Tran Ngoc Anh, Phuong Huy Nguyen, Dae-Hong Kim, Jongho Kim, 2024. A deep learning modeling framework with uncertainty quantification for inflow-outflow predictions for cascade reservoirs. Journal of Hydrology, 629, 130608. https://doi.org/10.1016/j.jhydrol.2024.130608.

Vinh Ngoc Tran, Valeriy Y. Ivanov, Jongho Kim, 2023b. Data reformation - A novel data processing technique enhancing machine learning applicability for predicting streamflow extremes. Advances in Water Resources, 182, 104569. https://doi.org/10.1016/j.advwatres.2023.104569.

Wang F., Cheng H., Dai H., Han H., 2021. Freeway short-term travel time prediction based on lightgbm algorithm. IOP Conference Series: Earth and Environmental Science. IOP Publishing. 012029. 10.1088/1755-1315/638/1/012029.

Wang Q., Wang S., 2020. Machine learning-based water level prediction in Lake Erie. Water, 12, 2654. https://doi.org/10.3390/w12102654.

Wunsch A., Liesch T., Broda S., 2018. Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). Journal of Hydrology, 567, 743–758. https://doi.org/10.1016/j.jhydrol.2018.01.045.

Xu W., Chen J., Zhang XJ., 2022. Scale effects of the monthly streamflow prediction using a state-of-the-art deep learning model. Water Resources Management, 36, 3609–3625. https://doi.org/10.1007/s11269-022-03216-y.

Yu Tong, Zhu Hong, 2020. Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv, 2003.05689.

Zhang F., Dai H., Tang D., 2014. A conjunction method of wavelet transform-particle swarm optimization-support vector machine for streamflow forecasting. Journal of Applied Mathematics, 910196. https://doi.org/10.1155/2014/910196.

Zhao G., Pang B., Xu Z., Xu L., 2020. A hybrid machine learning framework for real-time water level prediction in high sediment load reaches. Journal of Hydrology, 581, 124422. https://doi.org/10.1016/j.jhydrol.2019.124422.

Zhu S., Hrnjica B., Ptak M., Choiński A., Sivakumar B., 2020. Forecasting of water level in multiple temperate lakes using machine learning models. Journal of Hydrology, 585, 124819. https://doi.org/10.1016/j.jhydrol.2020.124819.