Hue University Journal of Science: Natural Science
Vol. 134, No. 1S-1, 141–155, 2025 (Special Issue: IFGTM 2025)

pISSN 1859-1388
eISSN 2615-9678

Salinity forecasting in the Vietnamese Mekong Delta: Evaluating the predictive power of machine learning approaches using multitemporal lag features

Le Van Quyen[1], Cuong Tuan Nguyen[1], Sameh A. Kantoush[2], Phan Cao Duong[3], Nguyen Thi Phuong Mai[4], Doan Van Binh[1]*

[1] Faculty of Engineering, Vietnamese-German University, Ho Chi Minh City, Vietnam
[2] Water Resources Research Center, Disaster Prevention Research Institute, Kyoto University, Goka-sho, Uji City, Kyoto, Japan
[3] Ireland's Centre for AI, School of Computer Science, University of Dublin, Belfield, Dublin 4, Ireland
[4] Department of Civil Engineering, Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam

* Correspondence to Doan Van Binh <binh.dv@vgu.edu.vn>

(*Received: 19 May 2025; Revised: 09 October 2025; Accepted: 19 December 2025*)

Abstract. Salinity intrusion poses a threat to water security, agriculture, and livelihoods in the Vietnamese Mekong Delta (VMD), particularly under the combined pressures of climate change and upstream hydrological developments. Accurate short- to mid-term salinity forecasts are essential for proactive water resource management. This study evaluates the performance of two machine learning models—Random Forest Regression (RFR) and Support Vector Regression (SVR)—for salinity forecasting using long-term observational data (1996–2023) from 44 monitoring stations in the VMD. To capture temporal dynamics, multitemporal lag features (1, 10, 20, 30, and 60 days) were generated from observed salinity records. Bayesian optimization and time-series cross-validation were used for model tuning. Results show that SVR performs best for short-term forecasts (1–3 days), achieving R^2 and *NSE* up to 0.927–0.928, *MAE* ≈ 0.824 g/L, and *RMSE* ≈ 1.858 g/L, while RFR provides more stable predictions over longer horizons (4–7 days), maintaining R^2/*NSE* values of 0.627 to 0.766 with lower errors. Additionally, the 20-day lag windows yielded the most accurate results, likely reflecting the influence of tidal cycles. These findings highlight the importance of selecting appropriate models and temporal features for various forecast horizons, providing a data-driven framework to enhance early warning systems and support adaptive water resource management in the VMD.

Keywords: Bayesian optimisation, machine learning, multitemporal lag features, salinity forecasting, Vietnamese Mekong Delta

1 Introduction

Salinity intrusion is a natural phenomenon commonly observed in coastal regions [1]. This refers to the inland intrusion of seawater through river estuaries. Under natural conditions, the extent and dynamics of salinity intrusion are primarily governed by local meteorological and hydrological factors, including tidal regimes, upstream river discharge, and precipitation. In low-lying deltaic regions, saline water penetrates more deeply inland, affecting a larger area.

In recent years, the manifestations of salinity intrusion have become increasingly pronounced under climate change and human activities. Numerous deltas worldwide have experienced the increasing severity of this phenomenon. For instance, the average salinity in the coastal areas of Bangladesh increased by approximately 26% over 35 years (1973–2009), adversely affecting various

localities [2], while in the neighboring Indian Bengal Delta, monitoring data show a sudden surge, with the salinity front shifting inland by roughly 20 km in the central delta zone [3], attributed to reduced freshwater discharge from the Ganges River, rising sea levels, and declining groundwater levels. These global observations underscore that salinity intrusion is not confined to a few regions but represents a widespread and escalating challenge for deltaic systems worldwide.

In that context, the Vietnamese Mekong Delta (VMD), one of the most vulnerable and densely populated deltas, has exhibited similar patterns, with salinity intrusion causing increasingly severe socio-economic impacts [4, 5]. During the 2015–2016 dry season, salinity intrusion damaged nearly 160,000 hectares of rice fields, directly affected 200,000 individuals, and caused indirect hardship to millions, resulting in estimated economic losses exceeding 1 trillion VND. A similar event in 2019 affected 100,000 hectares of crops and 320,000 people, leading to damage of 570 billion VND. In 2020, the salinity penetrated up to 100 km inland, severely disrupting 460,000 hectares of cropland and leaving 685,558 people across 10 provinces with insufficient freshwater for agriculture and domestic use. These recurring and increasingly severe impacts highlight the critical need for effective salinity forecasting tools to support proactive water resource management, agricultural planning, and livelihood protection in the region [6].

For forecasting salinity, conventional hydrodynamic and statistical models have traditionally been employed [7-9]. However, these approaches often encounter limitations, including high computational demands, extensive data requirements, and relatively low prediction accuracy in highly dynamic environments. Recently, machine learning (ML) techniques have emerged as powerful alternatives, demonstrating considerable potential in water resource forecasting because of their capability in handling nonlinear relationships, managing complex datasets, and delivering robust predictions with fewer assumptions and simplified parameterization [10]. This may lead to greater predictive accuracy even in highly dynamic water environments. Machine learning can infer relationships directly from large, diverse datasets without requiring extensive, predefined physical parameters typically required by traditional models. Furthermore, this data-driven approach significantly reduces computational costs, offering impractical real-time prediction capabilities for many computationally intensive hydrodynamic models.

Among ML methods, Random Forest Regression (RFR) and Support Vector Regression (SVR) are notably effective because of their adaptability to nonlinear phenomena, and successful application in various hydrological forecasting tasks [11, 12]. Previous studies have highlighted the strength of these models in predicting hydrological variables, including rainfall and stream flow. Although several studies have examined salinity intrusion in the VMD, most have been restricted to limited spatial scales or have addressed different research questions [13, 14], leaving a substantial gap in comprehensive forecasting approaches. Therefore, the potential of ML-based models for salinity intrusion forecasting – particularly those incorporating multiple temporal lags as input features to capture broader spatio-temporal dynamics – warrants further investigation in the context of the VMD.

Therefore, this study aims to bridge this knowledge gap by evaluating the predictive power of the RFR and SVR models using multitemporal lag features to forecast daily salinity intrusion in

pISSN 1859-1388
eISSN 2615-9678

the VMD. Specifically, we assess model performance across different temporal windows (lag features of 1, 10, 20, 30, and 60 days) and various forecast horizons (from 1 to 7 days ahead). The outcomes of this research are expected to provide valuable insights into selecting appropriate forecasting strategies and enhancing early warning systems, thereby supporting sustainable water resource management and climate adaptation strategies in the region.

2 Materials and methods

2.1 Study area

The Vietnamese Mekong Delta, encompassing approximately 40,000 km² in the southernmost part of Vietnam, represents the final segment of the Mekong River before it discharges into the East Vietnam Sea (Fig. 1). This region is characterized by a complex network of rivers and canals that support diverse and dynamic hydrological processes. Home to more than 19 million people across 13 provinces, the VMD is a vital hub for agriculture and aquaculture, contributing approximately 50% of the nation's rice output and 60% of its aquatic products [15].

Despite its productivity, the VMD is increasingly vulnerable to environmental challenges. Climate change-induced sea-level rise, upstream hydropower development, and land subsidence have intensified salinity intrusion, particularly during the dry season [16]. These intrusions threaten freshwater availability, agricultural sustainability, and the livelihoods of local communities. Addressing these challenges necessitates integrated water resource management and adaptive strategies to ensure the region's resilience and continued contribution to national food security.

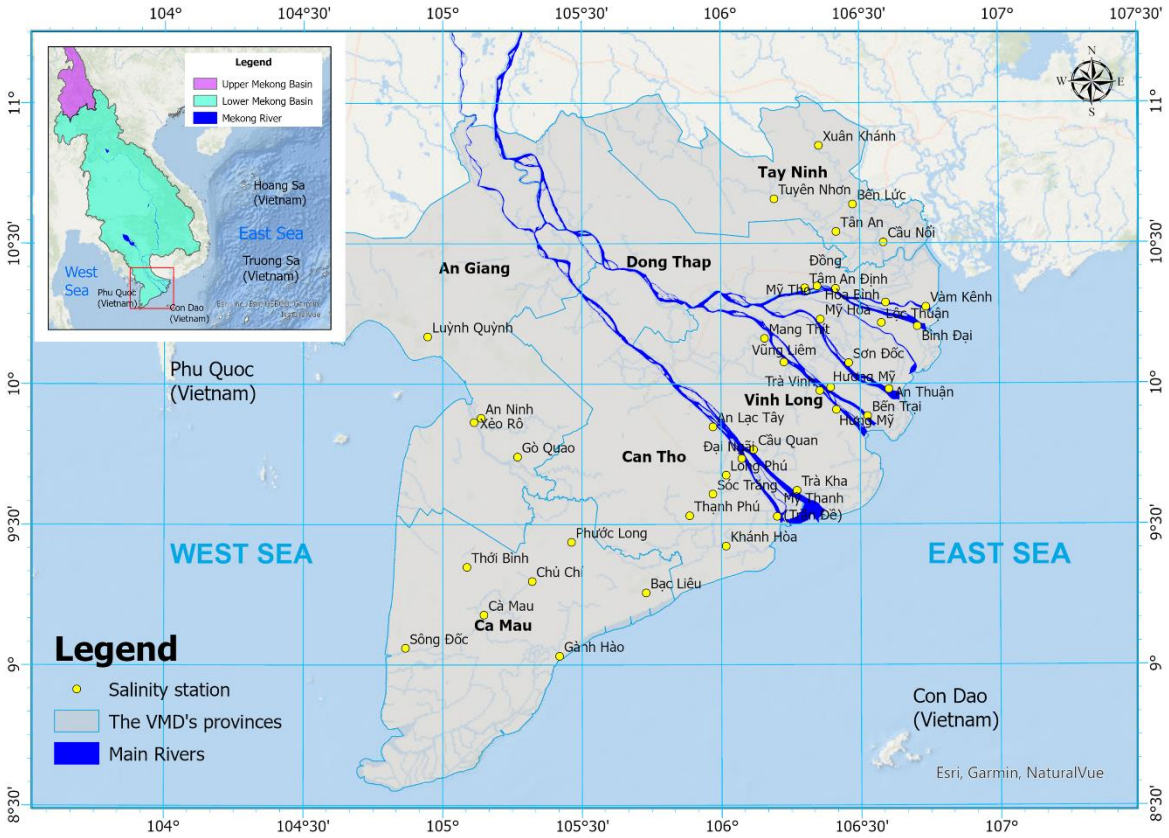

Fig. 1. Location map of Vietnamese Mekong Delta and salinity observation stations

2.2 Data collection and preprocessing

This study utilizes hourly salinity records collected from 44 hydrological monitoring stations located across the coastal regions of the VMD from 1996 to 2023. The stations are spatially distributed to cover key salinity-prone areas (see Fig. 1). The raw data were compiled from various sources and underwent several preprocessing steps (i.e., data cleaning, formatting, outlier removal, and data normalization) to ensure consistency and suitability for model development. Specifically, the daily maximum salinity values were extracted from the hourly records, formatted into time series data structures, and screened to remove outliers.

Importantly, salinity observations were predominantly available during the dry season (December to June), when saltwater intrusion is most severe. In contrast, salinity is typically not measured during the wet season, resulting in temporal gaps in the dataset.

Given the non-continuous nature of field measurements – usually recorded every two hours but not daily – substantial gaps (e.g., missing days or weeks) are present in the time series. To generate high-quality training data for machine learning models, only continuous data segments meeting predefined lookback window lengths (i.e., 1, 10, 20, 30, and 60 days) were retained. Importantly, no gap-filling or interpolation methods were applied; the analysis relies solely on the observed values. While this may reduce the total volume of usable data, it ensures that the dataset is free from uncertainties related to imputation and better reflects the original measurement conditions.

For ensuring comparability across input features and improving training dynamics, the salinity data were normalized by using *MinMaxScaler*. This min-max normalization technique linearly scales values to the [0, 1] range based on the minimum and maximum observed values of each variable. Multitemporal lag features were then generated to capture temporal dependencies at varying time intervals that are essential for forecasting tasks. Finally, the prepared dataset was partitioned into three subsets: 70% for model training, 20% for validation, and 10% for testing. This partitioning strategy supports robust model evaluation and helps prevent overfitting.

2.3 Machine learning models

This study evaluated the predictive performance of two state-of-the-art machine learning models: Random Forest Regression and Support Vector Regression, both of which have demonstrated high effectiveness in modelling nonlinear and multitemporal environmental phenomena.

Random forest regression

Random forest regression, introduced by Breiman, is an ensemble learning technique that builds various decision trees during training and outputs the average prediction of the individual trees [17]. Each tree is trained on a randomly sampled bootstrap subset of the data, and at each node, a random subset of features is considered for splitting. This randomized construction helps mitigate overfitting and improves model generalizability. Random forest regression has been widely applied in environmental modelling because of its robustness to noise, ability to capture complex nonlinear interactions, and tolerance to missing or correlated input features [18, 19]. In salinity forecasting, RFR is particularly suitable because it can model the intricate relationships between lagged hydrological variables.

This study implements RFR using the scikit-learn library. To maximize its predictive accuracy and control for model complexity, its crucial hyperparameters were systematically tuned. We employed Bayesian optimization for this task,

pISSN 1859-1388
eISSN 2615-9678

which is a sample-efficient method that intelligently searches for optimal parameter values by building a probabilistic model [20] of the objective function (e.g., R^2 score). For ensuring that the evaluation was robust and respected the temporal nature of the salinity data, this optimization was coupled with a time series cross-validation scheme by using five consecutive splits (n_splits = 5). This cross-validation approach is critical as it preserves the chronological order of observations in each split, preventing data leakage and yielding a more realistic estimate of the model's generalization performance on unseen future data [21].

Support vector regression

Support vector regression is a kernel-based method derived from the theory of support vector machines (SVMs), and adapted for regression tasks [22, 23]. Support vector regression seeks a function that approximates the target values within an ε-insensitive tube, while minimizing a regularized loss that controls model complexity. This structural risk minimization principle offers powerful generalizability, especially when the amount of training data is limited.

To handle the nonlinear relationships, the SVR model employs the *radial basis function* (RBF) kernel, which implicitly maps input features into a high-dimensional space where linear regression becomes feasible [24]. Evidence indicates that this kernel outperforms others in terms of efficiency and accuracy when applied to regression problems [25, 26]. Besides, the model's hyperparameters, including the penalty term C, the kernel width γ, and the margin ϵ, are also optimized via Bayesian optimization and time series cross-validation.

Support vector regression has proven effective in prior hydrological studies and is well-suited for capturing time-lagged dependencies and threshold effects in salinity dynamics [27, 28]. However, its computational cost increases with data size, and its performance is sensitive to parameter settings, necessitating careful tuning.

2.4 Evaluation metrics

For evaluating model performance, four widely accepted metrics are used: the root mean square error (*RMSE*), the mean absolute error (*MAE*), the coefficient of determination (R^2), and the Nash-Sutcliffe model efficiency coefficient (*NSE*) [7]. These metrics collectively measure the models' accuracy, precision, and explanatory power, enabling a comprehensive comparison of the forecasting scenarios across different lag intervals and prediction horizons. The root mean square error and *MAE* examine the difference between forecasted and observed values, ranging from 0 to positive infinity, with the lower values indicating better performance. The coefficient of determination ranges from 0 to 1.0, with higher values indicating a better model fit, whereas the *NSE* varies from negative infinity to 1.0, with a value of 1.0 representing the optimal fit [29]. Although R^2 has the advantage of being easily interpretable and consistently bounded between 0 and 1.0, the *NSE* provides a critical reference point at 0, where *NSE* values below 0 indicate that the means of the observed data serve as a better predictor than the model forecasting. Formulas to calculate those metrics are described as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}i - yi)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}i - yi|$$

$$R^2 = \left(\frac{n\sum_{i=1}^{n}(y_i\hat{y}_i) - \sum_{i=1}^{n}y_i\sum_{i=1}^{n}\hat{y}_i}{\sqrt{\left(n\sum_{i=1}^{n}y_i^2 - (\sum y_i)^2\right)\left(n\sum_{i=1}^{n}\hat{y}_i^2 - (\sum \hat{y}_i)^2\right)}}\right)^2$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(\hat{y}i - yi)^2}{\sum_{i=1}^{n}(\bar{y}i - yi)^2}$$

where n is the number of samples; $yi, \hat{y}i,$ and $\bar{y}$ are the observed, forecasted, and mean salinity values, respectively.

2.5 Modelling methodology

This study employed a data-driven modelling framework to forecast salinity levels in the VMD via machine learning techniques. The methodological process is structured into four main stages (Fig. 2): data collection, data processing, model training, and performance evaluation.

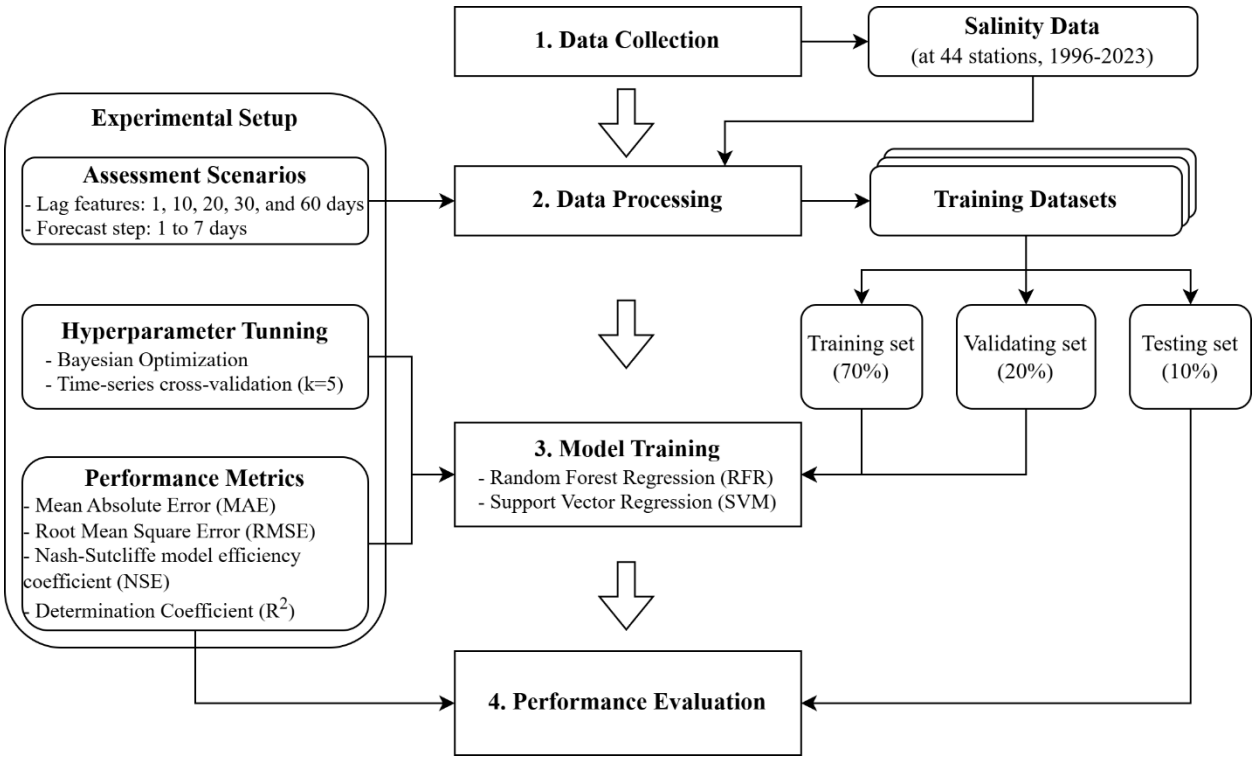

Fig. 2. Overall workflow for evaluating salinity predictive performance of RFR and SVR models

Step 1: Salinity observations from 44 monitoring stations across the VMD's coastal area, spanning from 1996 to 2023, were compiled as the primary dataset for model development and evaluation. The raw records included sub-daily measurements, which were aggregated into daily maxima to establish a consistent temporal resolution for the target variable.

Step 2: Each station's time series was first processed independently to ensure temporal integrity. Outliers were removed; values were normalized, and the series were filtered to retain only continuous daily segments sufficient to accommodate the specified lag windows (1, 10, 20, 30, and 60 days) and forecast horizons (1 to 7 days). After this station-specific preprocessing, the datasets were chronologically partitioned into training (70%), validation (20%), and testing (10%) subsets. Corresponding subsets from all 44 stations were then concatenated, and lagged variables were generated, allowing each time-ordered sample from any station to serve as an independent instance for model training.

Step 3: Model training was performed via the training subset, and hyperparameter optimization was carried out via Bayesian optimization in conjunction with 5-fold time series cross-validation. This strategy enabled the models to capture salinity dynamics and temporal patterns adaptively. Each model was trained and evaluated across seven forecast horizons from 1 to 7 days.

pISSN 1859-1388
eISSN 2615-9678

Step 4: Model performance was quantitatively assessed via four statistical metrics: *MAE*, *RMSE*, *NSE*, and R^2. These metrics were applied to the testing dataset to evaluate the generalization ability of the models under varying forecast horizons and lag scenarios. The best validated model is used to predict salinity concentrations at the analyzed stations.

3 Results

3.1 Data exploration

The dataset comprises daily records of maximum salinity collected from 44 monitoring stations across the coastal areas of the VMD, with values ranging from 0.025 to 48.2 g/L (mean = 9.20 g/L; median = 6.3 g/L; SD = 8.89 g/L), as illustrated in Fig. 3a. The positively skewed distribution reflects substantial spatial and temporal variability, from persistently fresh conditions at upstream stations (<1 g/L) to extreme values exceeding 40 g/L at coastal sites (Fig. 3b). These high-end observations, although flagged as statistical outliers in boxplot diagnostics, are hydrologically meaningful and were retained to capture the full extent of seawater intrusion. Together, these statistics underscore the diverse hydrological regimes across the monitoring network and provide a robust basis for evaluating the predictive performance of machine learning models under varying conditions.

In addition to spatial variability, temporal configuration also plays a crucial role. In particular, the choice of lag window and forecast horizon directly affects the number of usable samples, which, in turn, constrains the model training and evaluation. Table 1 presents the number of samples available for model training, validation, and testing across different lag window configurations. Increasing the lag window reduced the usable samples, primarily because it excluded the time steps that lacked a continuous data record. However, the observations among the designed lag windows have no notable imbalance in model training, validation, and testing.

The configuration with a 1-day lag retained the largest sample size, with training samples ranging from 15,875 to 51,757, validation samples from 4,533 to 14,794, and testing samples from 2,280 to 7,417. This is the baseline scenario, where most of the original dataset is preserved. In contrast, expanding the lag window to 10, 20, and 30 days resulted in a substantial decline in available samples, with training data typically ranging from approximately 14,948 to 15,640 samples. The validation and testing subsets also showed minimal variation. When the lag was further extended to 60 days, the dataset size reduced to its smallest value, with training sets ranging from 13,475 to 14,144, validation sets ranging from 3,802 to 4,043, and testing sets ranging from 1,884 to 2,036.

Table 1. Number of available samples corresponding to the lag window

Lag window	Training sample	Validation sample	Testing sample
1 day	15,875–51,757	4,533–14,794	2,280–7,417
10 days	15,539–15,640	4,429–4,468	2,222–2,251
20 days	15,368–15,474	4,372–4,422	2,189–2,225
30 days	14,948–15,190	4,247–4,339	2,121–2,186
60 days	13,475–14,144	3,802–4,043	1,884–2,036

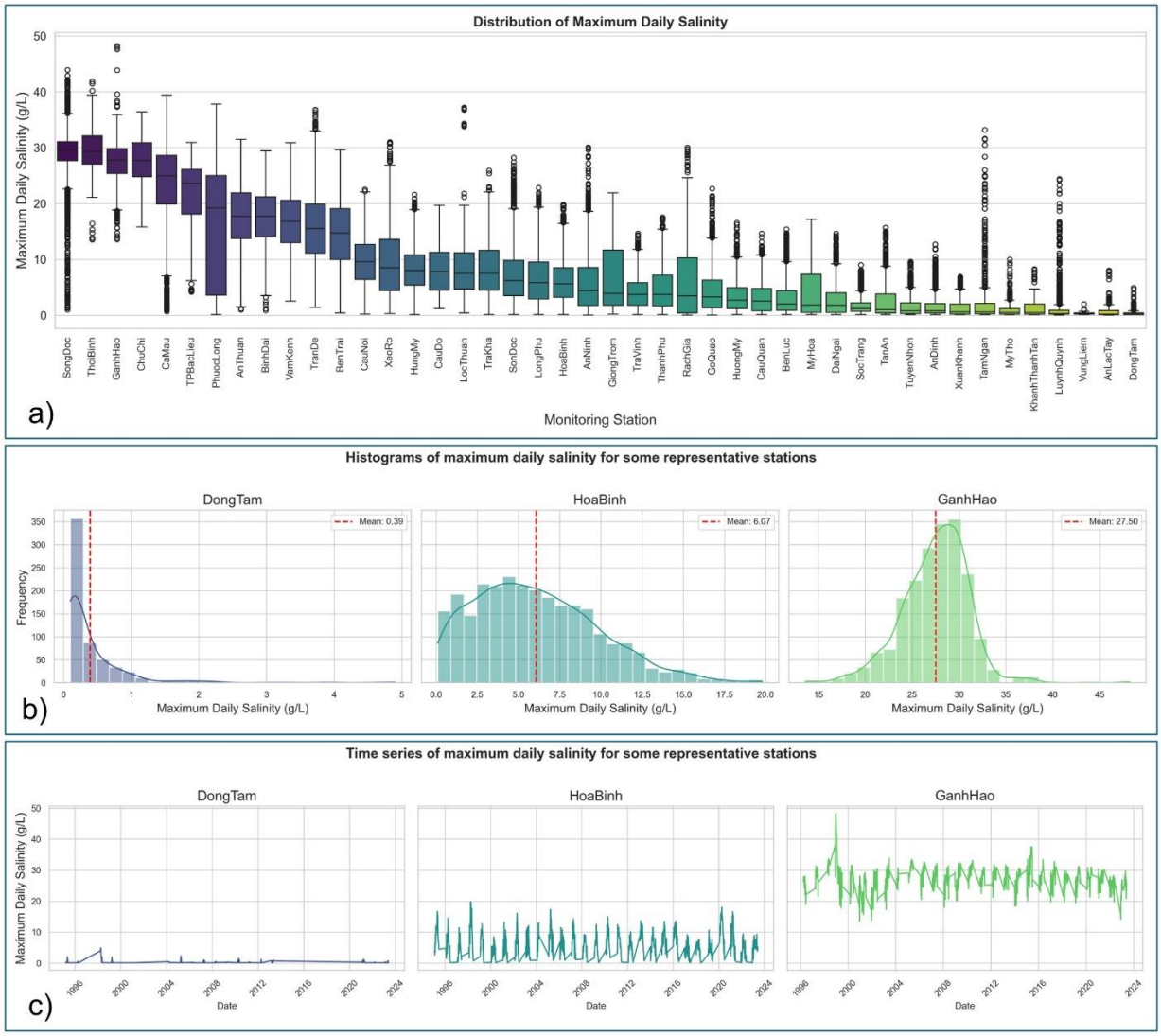

Fig. 3. Distribution of maximum daily salinity across monitoring stations in study areas (a); histograms (b); time series data (c) of representative upstream (Dong Tam), transitional (Hoa Binh), and coastal (Ganh Hao) stations, illustrating distinct spatial patterns

3.2 Model performance evaluation

The predictive performance of the RFR and SVR models was assessed across a range of lag features and forecast horizons to investigate the influence of the look-back window length and prediction step on forecasting accuracy. Seventy modelling scenarios were evaluated, corresponding to combinations of five lag windows and seven forecast steps (i.e., 35 scenarios for each model).

During the training phase, both models consistently achieved high R^2 and *NSE* across all lag-horizon combinations, indicating robust fitting to the observed data. However, their error dynamics varied substantially with the forecast length. The SVR model demonstrated superior predictive accuracy for shorter forecasts (1–2 days) when longer lag windows (20–60 days) were applied. Under these conditions, the lowest training errors were obtained, with minimum *MAE* and *RMSE* values of 0.787 g/L and 1.385 g/L,

pISSN 1859-1388
eISSN 2615-9678

respectively. In contrast, the RFR exhibited greater robustness over extended forecast horizons. Although slightly less accurate for the short-term predictions, RFR maintains higher R^2/*NSE* and smaller error magnitudes as the lead time increases. For instance, at a 60-day lag and a 7-day forecast step, RFR achieved an R^2 of 0.93 and *NSE* of 0.928, with *RMSE* of 1.802 g/L and *MAE* of 1.145 g/L, outperforming SVR (*RMSE* = 2.874 g/L and *MAE* = 1.736 g/L). Overall, the highest training performance for both models was observed at a forecast step of 1 day, where SVR and RFR achieved R^2 values of 0.955 (lag window 30 days) and 0.952 (lag window 60 days), respectively.

In the testing phase, both models also exhibited strong generalizability. For shorter forecast steps (1–2 days ahead), R^2 and *NSE* values exceeded 0.85 in multiple scenarios, with the *MAE* typically remaining around 1.0 g/L. The best test performance ($R^2 \approx NSE \approx 0.928$) was recorded at a lag of 1 and a step of 1 for both models (Fig. 4). As the forecast horizon extended, a gradual decline in R^2 and *NSE* was observed. The SVR maintained an advantage in certain mid-range scenarios—for example, attaining R^2 and *NSE* of 0.813 and 0.806, respectively, at a lag of 10 days and a forecast step of 3 days. However, RFR showed more consistent performance over longer forecast steps (4–7 days), particularly at a lag of 30 days, where R^2 and *NSE* remained above 0.62 even at step 7. In contrast, the performance of SVR deteriorated more noticeably at extended horizons, especially when longer lag features (e.g., 60 days) were used.

Notably, R^2 and *NSE* witnessed nearly identical patterns across the lag-horizon combinations, confirming their strong agreement in evaluating model efficiency. The coefficient of determination emphasizes the proportion of variance explained, whereas *NSE* provides a more sensitive measure of predictive reliability relative to the mean observation. In both senses, RFR consistently outperforms SVR for medium to long horizons, while SVR slightly surpassed RFR in very short forecasts. This balance implies that while SVR effectively captures shorter fluctuations, RFR offers greater generalization and stability, making it more suitable for operational salinity forecasting over multiple-day horizons.

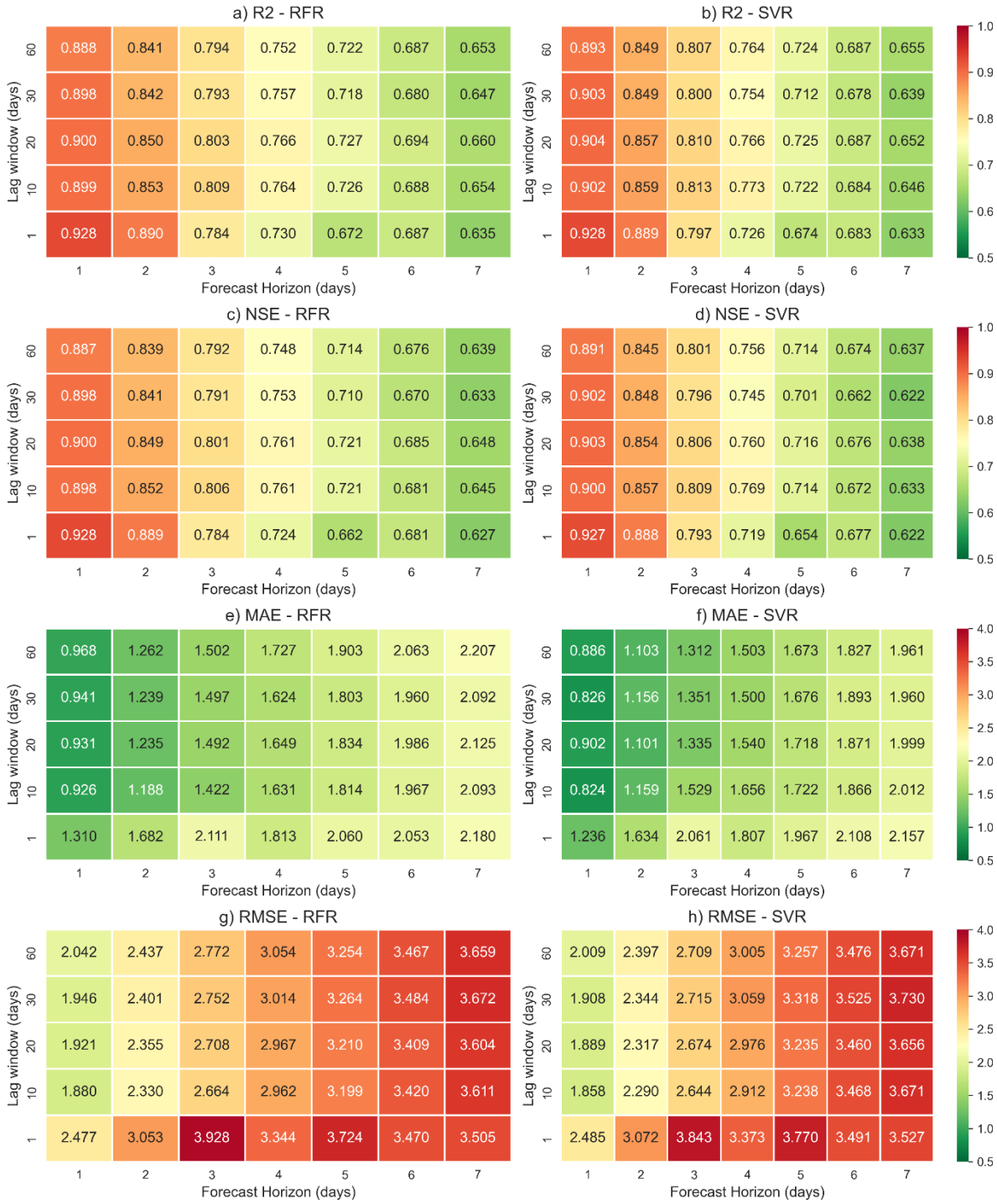

Fig. 4. Performance heatmaps of RFR and SVR models across different lag windows and forecast horizons Each cell represents the performance score for a specific combination of a lag window (y-axis) and a forecast horizon (x-axis). The color intensity corresponds to the metric value, with higher R^2/*NSE* and lower *MAE*/*RMSE* indicating better model performance.

pISSN 1859-1388
eISSN 2615-9678

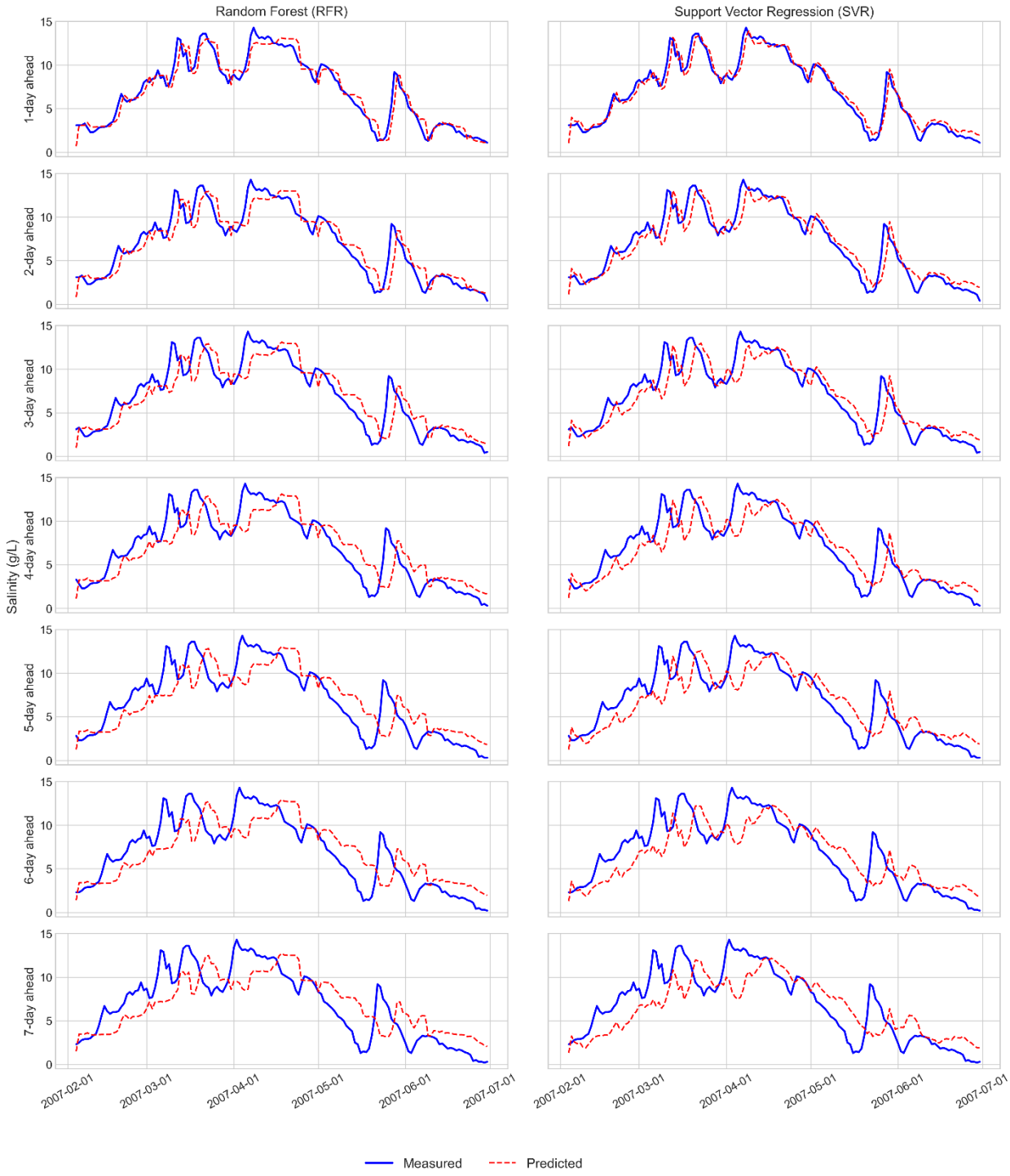

Fig. 5. Multistep salinity forecasts at Hoa Binh station via RFR and SVR with a 20-day lag window
Each subplot represents one forecast horizon, with the observed salinity (blue line) and predicted salinity (red dashed line) plotted over time (i.e., illustrated over a representative period from the testing dataset). Salinity is expressed in g/L. Hoa Binh station represents a transitional site where the 20-day lag gave the best performance.

4 Discussion

4.1 Optimal lag features

The results indicate that both models achieved their highest predictive performance when using lag windows of 20 days. In contrast, shorter configurations such as 1-day and 10-day lags yielded strong performance for short-term forecasts (e.g., next-day predictions), but their accuracy declined substantially as the prediction horizon increased. This degradation is likely due to insufficient historical context in shorter lags, which limits the model's ability to generalize temporal patterns in salinity fluctuations – a phenomenon also noted in hydrological forecasting studies [30, 31].

Interestingly, the 20-day lag window, which roughly corresponds to one tidal cycle, provided accurate next-day predictions and sustained performance across multiple lead times. Although accuracy gradually declined as the forecast step increased, the reduction was smooth and stable, suggesting that this lag range strikes a suitable balance between information richness and noise suppression. Similar findings were reported by Ma et al., who demonstrated that incorporating historical data in the range of 15 to 30 days significantly improved flood forecasting accuracy via a deep learning framework [32]. Chen et al. also highlighted the effectiveness of incorporating tidal-scale temporal features in river level forecasting under typhoon conditions [33]. These studies reinforce that lag windows aligned with dominant hydro-tidal cycles can enhance the model's ability to capture periodic dynamics and improve the robustness of multistep prediction.

Surprisingly, extending the lag window from 30 to 60 days—initially assumed to provide more comprehensive historical information—did not improve model performance. The accuracy gains were marginal and comparable with those from the 1-day and 10-day configurations. This outcome suggests that excessive lag depth may introduce redundant or noisy information, thereby diluting relevant signals and potentially compromising the model's predictive ability [32]. In addition, as shown in Table 1, a longer lag substantially reduces the available samples, which may limit the models' capacity to learn robust patterns. Thus, the combined effects of redundant information and reduced sample size likely explain the lack of improvement with the longest lag window.

4.2 Comparative model performance

Results from Section 3.2 also showed that in this study, SVR exhibited greater performance in short-term forecasts (1–3 days), whereas RFR demonstrated better stability over extended prediction horizons (4–7 days). This distinction aligns with findings from prior hydrological studies. For example, Bargam et al. reported that SVR achieved lower *RMSE* and higher *NSE* scores than did RFR when applied to daily streamflow prediction in a data-scarce basin, suggesting SVR's ability to capture high-resolution, short-term dynamics with fewer training samples [34].

The superior short-term performance of SVR may be attributed to its kernel-based learning mechanism, which is well-suited for capturing local nonlinearities in time series data without overfitting. However, as the forecast horizon increases, SVR's reliance on recent observations may limit its capacity to model broader temporal dependencies and compound errors across steps. In contrast, RFR, as an ensemble of decision trees, benefits from its ability to generalize over longer historical patterns and mitigate overfitting by averaging across multiple estimators. These differences in model behavior have also been observed in applications such as drought or streamflow forecasting, where SVR typically excels

pISSN 1859-1388
eISSN 2615-9678

in short-term prediction. In contrast, RFR maintains greater stability over extended timeframes [35, 36].

The salinity forecasting results at the Hoa Binh station with a 20-day lag window (Fig. 5) clearly illustrate the behavioral differences between the RFR and SVR models across various forecast horizons. In short-term forecasts (Steps 1–3), SVR can follow observed trends, particularly during sharp increases and decreases in data. However, as the forecast horizon extends (Steps 4–7), SVR tends to smooth the predictions and attenuate amplitude variations. In contrast, RFR maintains the overall trend and demonstrates greater temporal stability. These findings are consistent with the quantitative results presented in Section 4.2, where SVR outperforms RFR in short-term forecasting because its kernel-based learning mechanism effectively captures local nonlinearities. Conversely, RFR, with its ensemble of decision trees and stronger generaliration capacity, performed more robustly over longer prediction windows.

This pattern is not unique to Hoa Binh; similar observations were made at stations such as Vam Kenh, Tra Kha, and Ben Trai. Support vector regression maintained relatively high short-term accuracy at locations with substantial variability or data gaps, whereas RFR exhibited greater resilience and consistency over extended forecast steps.

4.3 Limitations and future works

Despite the encouraging performance of both machine learning models in forecasting salinity across the VMD, this study is subject to certain limitations. First, the training dataset was constructed by aggregating data from all stations, which may have led to model confusion regarding station-specific patterns, potentially reducing predictive accuracy. This limitation arose from the lack of sufficient data at individual stations to train the ML separately. In future work, we plan to develop more advanced ML and deep learning models to examine further the benefits and drawbacks of using combined versus individual datasets. Second, the models were trained on processed data without applying imputation or decomposition techniques. While this was a deliberate choice to assess model performance under minimal assumptions, it undoubtedly constrained the predictive potential of both models.

Future studies will focus on improving model performance by strategically selecting training data tailored to each station's characteristics. Additionally, appropriate preprocessing techniques should be applied to exploit better the temporal and statistical features of the salinity data. Integrating other hydrometeorological variables, such as rainfall, streamflow, and water levels, will also be explored. Finally, the authors intend to investigate deep learning models and ensemble techniques to increase forecasting accuracy and robustness.

5 Conclusions

This study evaluated the salinity forecasting performance of random forest regression and support vector regression by using multitemporal lag features across different forecast horizons in the Vietnamese Mekong Delta. The results revealed that both models achieved high predictive accuracy, particularly with a 20-day lag window. Across the four evaluation metrics (R^2, *NSE*, *MAE*, and *RMSE*), support vector regression demonstrated superior short-term forecasting ability (1–3 days), whereas random forest regression exhibited more stable and reliable performance at longer forecast horizons (4–7 days). These differences are consistent with the inherent characteristics of each algorithm: support vector

regression's strength in capturing local nonlinearities versus random forest regression's robustness in generalizing over extended temporal contexts.

Extending the lag window to 30–60 days failed to improve accuracy, suggesting excessive historical data may introduce noise. The findings underscore the importance of selecting appropriate lag structures for effective multistep forecasting. While both models performed well when aggregated data were used, further improvements can be achieved by incorporating station-specific training, feature preprocessing, and integrating additional hydrometeorological variables. Overall, the proposed machine learning framework offers a promising data-driven approach for supporting early warning systems and adaptive water resource management in the Vietnamese Mekong Delta.

Acknowledgment

This research is funded by the National Foundation for Science & Technology Development (NAFOSTED) under grant number 105.08-2023.40.

References

1. De Costa GS, Kojiri T, Porter M. Salinity intrusion: its characteristics and impact - cases in the Asia Pacific region. In: Lee JHW, Lam KM, editors. 4th International Symposium on Environmental Hydraulics. Hong Kong, China; 2005. p. 2027-32.
2. Mahmuduzzaman M, Uddin Z, Nuruzzaman A, Rabbi F, Ahmed S. Causes of Salinity Intrusion in Coastal Belt of Bangladesh. International Journal of Plant Research. 2014;2014:8-13.
3. Sherin VR, Durand F, Papa F, Islam AKMS, Gopalakrishna VV, Khaki M, et al. Recent salinity intrusion in the Bengal delta: Observations and possible causes. Continental Shelf Research. 2020;202:104142.
4. Biswas KK, Sarder BC, Saika U, Hoque MA-A. Environmental and Socio-economic Impacts of Salinity Intrusion in the Coastal Area: A Case Study on Munshigong Union, Shymnagor, Satkhira. Jahangirnagar University Environmental Bulletin. 2013;2:41-9.
5. Dung TD. Drought and salinity intrusion in the Mekong Delta: What are the fundamental solutions? Health & Agricultural Policy Research Institute. 2024.
6. Minderhoud P, Coumou L, Erban L, Middelkoop H, Stouthamer E, Addink E. The relation between land use and subsidence in the Vietnamese Mekong Delta. Science of the Total Environment. 2018;634:715-26.
7. Qiu C, Zhu J-R. Influence of seasonal runoff regulation by the Three Gorges Reservoir on saltwater intrusion in the Changjiang River Estuary. Continental Shelf Research. 2013;71:16-26.
8. Andrews SW, Gross ES, Hutton PH. Modeling salt intrusion in the San Francisco Estuary prior to anthropogenic influence. Continental Shelf Research. 2017;146:58-81.
9. Gong W, Lin Z, Zhang H, Lin H. The response of salt intrusion to changes in river discharge, tidal range, and winds, based on wavelet analysis in the Modaomen estuary, China. Ocean & Coastal Management. 2022;219.
10. Ahmed AA, Sayed S, Abdoulhalik A, Moutari S, Oyedele L. Applications of machine learning to water resources management: A review of present status and future opportunities. Journal of Cleaner Production. 2024;441:140715.
11. Tran TT, Pham NH, Pham QB, Pham TL, Ngo XQ, Nguyen DL, et al. Performances of Different Machine Learning Algorithms for Predicting Saltwater Intrusion in the Vietnamese Mekong Delta Using Limited Input Data: A Study from Ham Luong River. Water Resources. 2022;49(3):391-401.
12. Dang LTT, Ishidaira H, Nguyen KP, Souma K, Magome J. Short-term salinity prediction for coastal areas of the Vietnamese Mekong Delta using various machine learning algorithms: a case study in Soc Trang Province. Applied Water Science. 2025;15(4):79.
13. Hoai PN, Quoc PB, Thanh Thai T. Apply Machine Learning to Predict Saltwater Intrusion in the Ham Luong River, Ben Tre Province. VNU Journal of Science: Earth and Environmental Sciences. 2022;38(3).

pISSN 1859-1388
eISSN 2615-9678

14. Hai T, Vu Van N, Hùng H, Tuan N, Dang T, Lam, et al. Assessing and Forecasting Saline Intrusion in the Vietnamese Mekong Delta Under the Impact of Upstream flow and Sea Level Rise. Journal of Environmental Science and Engineering B. 2019;8:174-85.
15. Cgiar Research Program on Climate Change A, Food Security - Southeast A. Assessment Report: The drought and salinity intrusion in the Mekong River Delta of Vietnam. Hanoi: CGIAR; 2016.
16. Loc HH, Low Lixian M, Park E, Dung TD, Shrestha S, Yoon Y-J. How the saline water intrusion has reshaped the agricultural landscape of the Vietnamese Mekong Delta, a review. Science of The Total Environment. 2021;794:148651.
17. Breiman L. Random forests. Machine learning. 2001;45:5-32.
18. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63:3-42.
19. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.
20. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Journal of Electronic Science and Technology. 2019;17(1):26-40.
21. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice: OTexts; 2018.
22. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Advances in neural information processing systems. 1996;9.
23. Vapnik V. The nature of statistical learning theory: Springer science & business media; 1999.
24. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing. 2004;14:199-222.
25. Ramedani Z, Omid M, Keyhani A, Shamshirband S, Khoshnevisan B. Potential of radial basis function based support vector regression for global solar radiation prediction. Renewable and Sustainable Energy Reviews. 2014;39:1005-11.
26. Li M, Liu Y-H, editors. Learning interaction force model for endodontic shaping with support vector regression. Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006 ICRA 2006; 2006: IEEE.
27. Yu P-S, Chen S-T, Chang IF. Support vector regression for real-time flood stage forecasting. Journal of Hydrology. 2006;328(3):704-16.
28. Guillou N, Chapalain G, Petton S. Predicting sea surface salinity in a tidal estuary with machine learning. Oceanologia. 2023;65(2):318-32.
29. Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology. 1970;10(3):282-90.
30. Zheng R, Sun Z, Jiao J, Ma Q, Zhao L. Salinity Prediction Based on Improved LSTM Model in the Qiantang Estuary, China. Journal of Marine Science and Engineering. 2024;12(8).
31. Moges DM, Virro H, Kmoch A, Cibin R, Rohith RAN, Martínez-Salvador A, et al. Streamflow Prediction with Time-Lag-Informed Random Forest and Its Performance Compared to SWAT in Diverse Catchments. Water. 2024;16(19).
32. Ma K, He D, Liu S, Ji X, Li Y, Jiang H. Novel time-lag informed deep learning framework for enhanced streamflow prediction and flood early warning in large-scale catchments. Journal of Hydrology. 2024;631.
33. Chen Y-C, Yeh H-C, Kao S-P, Wei C, Su P-Y. Water Level Forecasting in Tidal Rivers during Typhoon Periods through Ensemble Empirical Mode Decomposition. Hydrology. 2023;10(2).
34. Bargam B, Boudhar A, Kinnard C, Bouamri H, Nifa K, Chehbouni A. Evaluation of the support vector regression (SVR) and the random forest (RF) models accuracy for streamflow prediction under a data-scarce basin in Morocco. Discover Applied Sciences. 2024;6(6).
35. Sharma B, Goel NK. Streamflow prediction using support vector regression machine learning model for Tehri Dam. Applied Water Science. 2024;14(5).
36. Abbes A, Inoubli R, Rhif M, Farah I. Combining deep learning methods and multi-resolution analysis for drought forecasting modeling. Earth Science Informatics. 2023;16:1-10.