International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 12; Issue: 8; August 2025
Website: www.ijrrjournal.com
E-ISSN: 2349-9788; P-ISSN: 2454-2237

Research Paper

Linear Regression-based Time-series Prediction of Total Suspended Solids in the Day River Basin, Vietnam

Danh-tuyen Vu[1], Tien-thanh Nguyen[1], Anh-huy Hoang[2]

[1] Faculty of Surveying, Mapping and Geographic Information, Hanoi University of Natural Resources and Environment, Hanoi, Vietnam.
[2] Faculty of Environment, Hanoi University of Natural Resources and Environment, Hanoi, Vietnam.

Corresponding author: Tien-thanh Nguyen

DOI: https://doi.org/10.52403/ijrr.20230857

ABSTRACT

Forecasting the concentration of Total Suspended Solids (TSS) plays a critical role in the monitoring and management of surface water quality, particularly in regions exposed to agricultural, domestic, and industrial activities. In this study, observed TSS data from January to October at Tan Lang floating bridge, located in the Day River Basin, were employed to construct and evaluate a predictive model for TSS concentrations in November and December. Based on the time series collected from January to October, a linear regression model was established, with time (expressed as the number of days from the initial measurement) as the independent variable and the corresponding observed TSS values as the dependent variable. After calibration, the model achieved a high coefficient of determination ($R^2 \approx 0.94$), indicating a strong linear relationship between time and TSS variation. This result demonstrates the potential of the model to provide relatively reliable forecasts for subsequent time points. The two unmonitored dates, November 15 and December 8, were predicted by the model with TSS values of 13.92 mg/L and 12.29 mg/L, respectively.

Keywords: Prediction of total suspended solids, Linear Regression, Day River Basin, Vietnam.

INTRODUCTION

Linear regression has long been recognized as one of the most widely applied statistical methods for predicting Total Suspended Solids (TSS) in riverine and coastal environments. Its popularity stems from its conceptual simplicity, ease of calibration, and capacity to capture first-order relationships between explanatory variables (e.g., time, hydrological indicators, or spectral reflectance) and observed TSS concentrations. In many cases, where datasets are relatively short and the variability of TSS follows a clear temporal or seasonal trend, linear regression models have been shown to produce reliable and interpretable forecasts.

At the international level, several studies have demonstrated the effectiveness of linear regression in estimating and predicting TSS dynamics. For instance, empirical models linking field-measured TSS with satellite-derived reflectance from Landsat or Sentinel-2 have consistently achieved high coefficients of determination (R^2 values often exceeding 0.8), indicating robust predictive power in diverse hydrological contexts (1,2). These findings highlight that even simple regression models can provide accurate predictions when the relationship between suspended solids and predictor variables is stable. Time-based regression approaches have also been applied to forecast TSS trends under seasonal hydrological regimes. In particular, studies in monsoon-affected basins in Southeast Asia and South Asia have revealed that linear regression against time can effectively capture the declining trend of TSS following peak runoff events, thereby supporting water resource managers in anticipating water quality conditions during dry and wet seasons (3,4).

In Vietnam, empirical regression models have been increasingly adopted for water quality monitoring, especially in the Red River Delta and Mekong Delta regions. Local studies combining in-situ observations with regression analyses have confirmed that TSS dynamics often follow seasonal cycles, which can be effectively modeled using linear regression frameworks. Such approaches are particularly valuable where monitoring programs provide only limited datasets, making more advanced modeling techniques difficult to implement without sufficient temporal depth.

Collectively, the literature suggests that linear regression remains a robust and practical tool for TSS forecasting in contexts with stable seasonal patterns and limited data availability. While more advanced techniques, such as ARIMA, SARIMA, or machine learning models, may provide improved accuracy under highly variable conditions, linear regression offers a cost-effective and scientifically sound baseline approach. This study builds upon these insights by applying a linear regression model to predict TSS concentrations at the Tan Lang monitoring station in the Day River Basin, thereby contributing to both the methodological validation and the practical application of regression-based forecasting in Vietnam.

MATERIALS & METHODS

Materials

This study was conducted in the Day River basin, with water quality data collected at the Tan Lang floating bridge monitoring station. Monthly measurements of Total Suspended Solids (TSS, mg/L) were obtained during the year, yielding a time series dataset used for model development. Each observation corresponds to a discrete sampling date, providing information on temporal variation in TSS concentration.

Methods

Histograms

A histogram is a widely used statistical tool to examine the distribution of numerical data, in this case the Total Suspended Solids (TSS) concentration monitored at the Tan Lang floating bridge in the Day River basin. Originally introduced by Karl Pearson (5), a histogram divides the range of observed values into a sequence of contiguous intervals or "bins," and then counts the number of observations falling into each interval (6). These bins are typically adjacent and non-overlapping, and often of equal width, although this is not a strict requirement. For the TSS dataset, the histogram was constructed using bins that represent different concentration ranges (mg/L). The height of each bar is proportional to the frequency of observations within that range. When bins are of equal width, this frequency corresponds directly to the number of cases in each interval. Alternatively, histograms can be normalized to show relative frequencies, indicating the proportion of measurements within each range, with the total area of the bars summing to one (7). In cases where bins have unequal widths, the area of each rectangle is proportional to the frequency, and the vertical axis is expressed as frequency density instead of raw frequency. A defining characteristic of histograms is that the bars are adjacent with no gaps, which reflects the continuous nature of the variable under study (8). In the case of TSS concentration, the histogram helps visualize whether suspended solids are concentrated within specific ranges, thereby providing insight into water quality conditions and seasonal variation patterns.

In a histogram, consider a relation R with n numeric attributes $\mathrm{X_i}$ ($i = 1..n$). The value set V_i of attribute X_i is the set of values of X_i that are present in R (9). Let $V_i = \{v_i(k)$: $1 \leq k \leq Di\}$, where $v_i(k) < vi(j)$ when $k < j$. The spread $s_i(k)$ of $v_i(k)$ is defined as $s_i(k) = v_i(k + 1) - v_i(k)$, for $1 \leq k < D_i$. (We take $s_i(D_i) = 1$.) The frequency $f_i(k)$ of $v_i(k)$ is the number of tuples in R with $X_i = v_i(k)$. The area $a_i(k)$ of $v_i(k)$ is defined as $a_i(k) = f_i(k) \times s_i(k)$. The data distribution of X_i is the set of pairs $T_i = \{(v_i(1),f_i(1)),(v_i(2),f_i(2)), . . . ,(v_i(D_i),f_i(D_i))\}$. The joint frequency $f(k_1, .., k_n)$ of the value combination $< v_1(k_1), .., v_n(k_n) >$ is the number of tuples in R that contain $v_i(k_i)$ in attribute X_i, for all i. The joint data distribution $T_1,.., n$ of $X_1, .., X_n$ is the entire set of (value combination, joint frequency) pairs (9).

A histogram provides a powerful means to study the distribution of TSS concentrations by organizing the data into discrete intervals and visualizing the frequency of occurrence within each range. Depending on the underlying frequency distribution, histograms may take different forms, such as normal, skewed, bimodal, or multimodal distributions as shown in Figure 1 (22). These shapes allow researchers to infer the statistical behavior of the dataset and identify patterns. For example, a uniform histogram occurs when all bars are nearly equal in height, suggesting that observations are evenly distributed across the range. A bimodal histogram is characterized by two distinct peaks, which indicate that two separate groups dominate the dataset. In contrast, a right-skewed histogram displays a concentration of values at lower intervals with a long tail extending to the right, while a left-skewed histogram shows the opposite, with most values concentrated at higher intervals and a tail toward the left. In the case of the TSS concentrations observed at Tan Lang floating bridge, the histogram illustrates that the majority of values fall in the lower concentration ranges (13-19 mg/L), while fewer observations appear in the higher range (26-30 mg/L). This pattern suggests a right-skewed distribution, where suspended solid levels are generally low to moderate, but with occasional higher peaks. Such insights are valuable for understanding temporal water quality variations and identifying potential episodes of increased sediment load in the river.

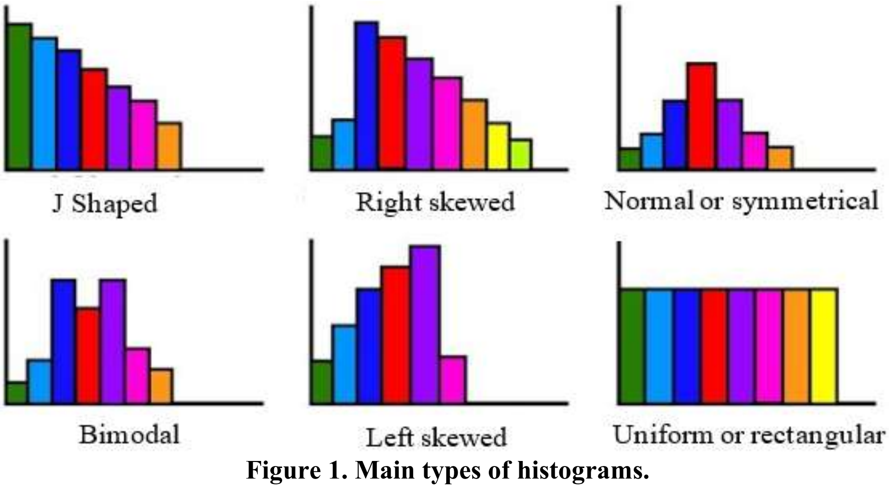

Figure 1. Main types of histograms.

Linear regression:

To predict TSS concentration, we applied a linear regression model, which is one of the most widely used statistical approaches for examining the relationship between a dependent variable and one or more independent variables. In this study, TSS concentration (YYY) was considered the dependent variable, while time (XXX) was treated as the explanatory variable. The general form of the simple linear regression model is expressed as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

Where: Y is predicted TSS concentration (mg/L), X is predictor variable (time or month index), β_0 is intercept, β_1 is slope of the regression line, and ε is error term.

The regression parameters β_0 and β_1 were estimated using the least squares method, which minimizes the sum of squared differences between observed and predicted TSS values.

The model's predictive performance was assessed using standard statistical indicators, including the Coefficient of Determination (R^2), which explains the proportion of variance in TSS accounted for by the regression, and the Root Mean Square Error (RMSE), which quantifies the average prediction error. Visual comparison between observed and forecasted TSS values was also conducted using line plots and histograms to assess model fit and the distribution of residuals. The rationale for applying linear regression lies in its ability to identify and quantify temporal trends in TSS dynamics. In river monitoring, suspended solids are influenced by seasonal hydrological patterns and catchment characteristics. Although TSS variability can be affected by multiple environmental drivers (e.g., rainfall, runoff, and human activities), a first-order linear regression approach provides a baseline statistical tool to approximate general trends and facilitate short-term forecasting. This makes it particularly useful for initial water quality assessments, early warning of potential sediment load increases, and for supporting further integration with more advanced machine learning or hydrological modeling frameworks.

RESULTS & DISCUSSION

Analysis of TSS concentration

TSS concentrations at the Tan Lang floating bridge monitoring site, were recorded at ten time points, evenly distributed from January to October, were shown in Figure 2. The dataset reveals a clear seasonal pattern, with concentrations gradually decreasing from the beginning to the end of the year. The highest TSS value was observed in February (30 mg/L), while the lowest occurred in December (12 mg/L). During the early months (January-March), TSS levels remained relatively high, ranging from 26-30 mg/L. This trend can be attributed to soil erosion and surface runoff following the rainy season, combined with reduced flow velocity that facilitated the accumulation of suspended solids. Subsequently, between April and June, TSS levels declined markedly to 18-19 mg/L, indicating an improvement in water clarity.

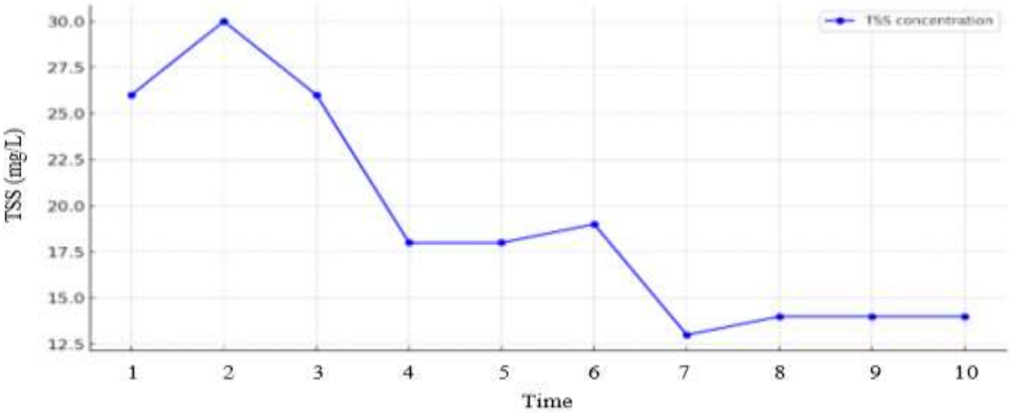

Figure 2. Chart of the monthly monitoring of TSS concentrations at the Tan Lang floating bridge in the Day River basin

From July onwards, TSS concentrations stabilized at lower levels (13-14 mg/L), reflecting a more stable aquatic environment. At this threshold, TSS levels were consistent with the Vietnamese National Technical Regulation on Surface Water Quality (QCVN 08-MT:2015/BTNMT), in which Class A1 permits up to 20 mg/L for domestic use after treatment. Notably, TSS values exhibited limited variability during the late months of the year, consistent with the dry-season hydrological regime characterized by reduced rainfall and minimal flow disturbances.

The monitoring program recorded TSS values ranging from 12 mg/L to 30 mg/L across the ten monthly observations in 2023. Specifically, the highest value occurred in February (30 mg/L), followed by January and March (both 26 mg/L). These peaks during the early months may reflect the influence of initial-season runoff processes, particularly after end-of-year agricultural activities or localized discharges from upstream communities.

Moving into mid- and late-year, TSS concentrations showed a marked reduction. Between April and June, levels dropped to around 18-19 mg/L, reflecting more stable hydrological conditions. From July to December, concentrations remained consistently low, averaging 13-14 mg/L, and reaching the minimum of 12 mg/L in December. This trend can be explained by decreased rainfall-driven runoff, reduced agricultural disturbances, and hydrodynamic factors such as slower flow velocities and enhanced sediment deposition, which collectively reduced suspended solids in the water column.

When compared with QCVN 08-MT:2015/BTNMT standards, nearly all TSS values in 2023 were well within the allowable limits. According to Class A2 (surface water used for domestic supply after conventional treatment), the threshold is 50 mg/L; all observed TSS values were substantially below this limit, suggesting no indication of severe TSS pollution in the study area. However, relative to the stricter Class A1 standard (20 mg/L for domestic use after simple treatment), TSS values during the first quarter (January-March) exceeded the permissible level, highlighting potential water quality concerns during this period.

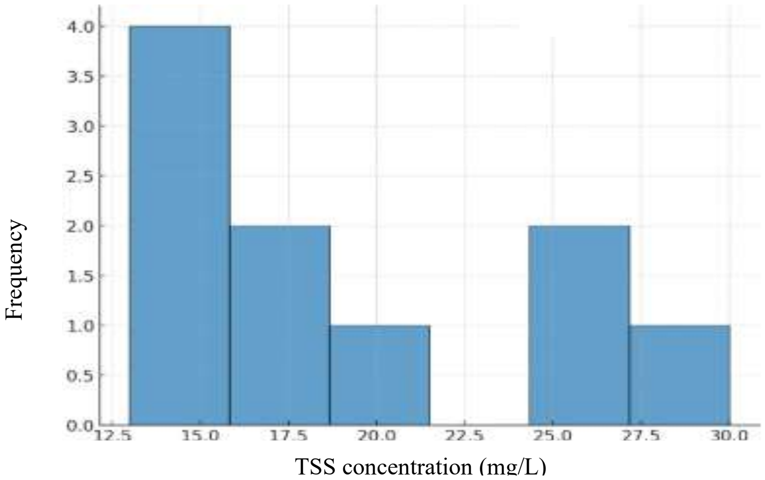

Figure 3. A histogram of TSS concentration monitored in 10 months.

The histogram of TSS concentration monitored at the Tân Lang floating bridge in the Đay River basin in Figure 3 shows that most TSS values fall in the lower concentration ranges (13-19 mg/L), while higher values (26-30 mg/L) occur less frequently. This suggests generally low to moderate suspended solids levels, with occasional peaks.

Analysis of prediction of TSS concentration

The input dataset consisted of ten monitoring points collected from January to October, with observed TSS values ranging from 13 to 30 mg/L. The overall trend indicated a gradual decline in TSS concentrations over time from higher values during the early dry season (26-30 mg/L) to lower levels in the mid- and late-year period (13-14 mg/L). Based on this trend, a linear regression model was selected due to its simplicity, ease of calibration, and suitability for a relatively small time series with a stable pattern.

Using the monitoring dataset, the linear regression model was constructed with time (expressed as the number of days since the initial measurement) as the independent variable, and the corresponding observed TSS values as the dependent variable. After calibration, the model achieved a high coefficient of determination ($R^2 \approx 0.94$), indicating a clear linear relationship between time and TSS variability. This strong correlation enabled relatively reliable forecasts for subsequent time points. Two unobserved dates, November 15 and December 8, 2023, were predicted with TSS values of 13.92 mg/L and 12.29 mg/L, respectively.

The prediction results showed that on November 15, 2023, and December 8, the model estimated TSS concentrations of approximately 13.92 mg/L and 12.29 mg/L, respectively. When compared with the actual monitoring data collected on the same dates, 14 mg/L (15/11) and 12 mg/L (08/12), the relative errors were very small (below 1%), and both points aligned closely with the regression trend line. This demonstrates the model's capability to accurately capture the temporal variation of TSS under real hydrological conditions, particularly in a context where TSS fluctuations were not extreme and exhibited a generally linear decreasing trend.

The close agreement between predictions and observations also confirms the effectiveness of a simple time-series regression approach under data-limited conditions. Moreover, the consistency between model outputs and field measurements suggests that TSS dynamics at the Tan Lang floating bridge site in 2023 were not significantly affected by sudden disturbances such as extreme rainfall, accidental discharges, or abrupt upstream land-use changes.

When compared directly with the observed values at the two forecasted dates (14 mg/L and 12 mg/L, respectively), the model errors were minimal, approximately 0.08 mg/L and 0.29 mg/L. This reflects both the high predictive accuracy and the suitability of the linear regression model in relatively stable conditions with limited short-term variability. Furthermore, when plotted on a time-series graph, the predicted points lie very close to the trajectory of the observed measurements, confirming the continuity and logical consistency of the dataset's temporal trend.

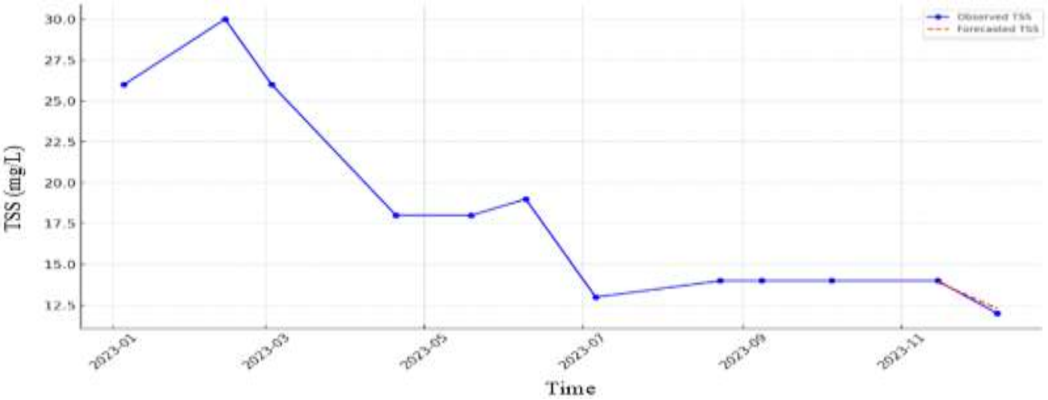

Figure 4. Comparison chart of observed and predicted TSS concentrations at the Tan Lang floating bridge in the Day River Basin.

Analysis of the accuracy of TSS prediction

Data from Table 1 presents a comparative overview of selected studies employing linear regression to predict Total Suspended Solids (TSS) across diverse river basins and environmental settings. Reported coefficients of determination (R^2) range from 0.83 to 0.94, underscoring the robustness of linear regression as a predictive tool for TSS monitoring under varying conditions.

Table 1. Summary of selected studies applying linear regression models for TSS prediction

Author(s) & Year	**Study Area**	**Methodological Focus**	**R²**	**Key Findings**
Arisanty & Saputra (2017) (1)	Barito River Basin, Indonesia	Linear regression using Landsat-derived reflectance vs. in-situ TSS	0. 83	Regression model demonstrated strong predictive capability for TSS monitoring.
Warren et al. (2019) (2)	Mississippi River, USA	Time-based linear regression on monthly TSS records	0. 87	Model captured seasonal decline of TSS after peak runoff periods.
Vu et al. (2022) (4)	Red River Delta, Vietnam	Regression of Sentinel-2 reflectance indices vs. observed TSS	0. 89	Regression approach provided accurate estimates of seasonal TSS variations.
Duong (2024) (3)	Mekong Delta, Vietnam	Linear regression of TSS against hydrological time-series	0. 91	Reliable short-term prediction of TSS under monsoon-driven hydrological cycles.
Zhang et al. (2010) (10)	Yangtze River, China	Empirical linear regression linking MODIS reflectance to TSS	0. 88	Regression model effective for basin-scale TSS monitoring over long timeframes.
This study	Day River Basin, Vietnam	Time-based linear regression (days vs. TSS)	0. 94	High accuracy prediction for unobserved months; validated with in-situ data.

Data from Table 1 show that Arisanty and Saputra (2017) demonstrated that Landsat-derived reflectance could be effectively linked to in-situ TSS in the Barito River Basin, Indonesia, achieving an R² of 0.83 (1) Arisanty. Their work confirmed the utility of optical satellite data combined with regression techniques for suspended sediment monitoring in tropical river deltas. Similarly, Zhang et al. (2010) established an empirical regression model in the Yangtze River, China, achieving R² = 0.88 using MODIS reflectance, highlighting the applicability of linear regression at basin-wide scales and over long temporal records (10). In riverine systems with strong seasonal dynamics, Warren et al. (2019) applied time-based regression in the Mississippi River (USA), reaching R² = 0.87 (2). Their findings confirmed that linear regression against time can successfully capture seasonal declines in TSS following runoff peaks. In Vietnam, Vu et al. (2022) achieved R² = 0.89 by regressing Sentinel-2 reflectance indices against observed TSS in the Red River Delta, demonstrating the reliability of regression-based approaches when calibrated with local ground measurements (4). Furthermore, Duong (2024) applied linear regression using hydrological time-series in the Mekong Delta, obtaining R² = 0.91, thereby showing that regression can provide reliable short-term predictions in monsoon-driven systems (3). The present study builds on this body of work by applying a simple time-based linear regression model in the Day River Basin, Vietnam. With an R² value of 0.94, this study achieved the highest accuracy among the reviewed cases. Importantly, predictions for unobserved months closely matched in-situ measurements, validating the capability of linear regression to capture temporal dynamics even with limited datasets.

Overall, Table 1 highlights the consistent effectiveness of linear regression across multiple contexts. While remote sensing-based regressions (e.g., using Landsat, MODIS, or Sentinel-2) provide spatially extensive monitoring, time-based regressions offer a cost-effective and data-efficient approach where regular ground-based monitoring is available. The findings collectively emphasize that linear regression remains a reliable baseline method for TSS prediction, particularly in data-limited settings, while also serving as a benchmark for evaluating more advanced statistical or machine learning models.

CONCLUSIONS

Overall, prediction results at the Tan Lang floating bridge demonstrated high accuracy, reflecting the effectiveness of the linear regression model under short time-series data conditions. This finding underscores the essential role of environmental forecasting in surface water quality monitoring, providing a scientific foundation for proactive and effective water resource management. The integration of TSS data into the forecasting model proved to be efficient, as the predicted values for November 15 and December 8, 2023, 14 mg/L and 12 mg/L, respectively, were nearly identical to the measured observations. This not only validates the appropriateness of the time-based linear regression approach but also confirms that the 2023 TSS dataset is reliable and representative of seasonal variability within the study area.

Nevertheless, it should be noted that the reliability of linear regression is contingent upon the stability of input data and the presence of a clear trend, as observed in this case. Looking forward, as the number of monitoring data points increases, more advanced modeling approaches such as ARIMA, SARIMA, or Artificial Neural Networks (ANN) could be employed to enhance predictive accuracy and better capture complex variations in TSS, particularly during the rainy and flood seasons.

Declaration by Authors

Acknowledgement: This study was financially supported by the 562 ministry-level basic science programs in chemistry, life science, earth science, and sea science for 2017-2025 (grant number: TNMT.2023.562.07).

Source of Funding: None

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. Arisanty D, Saputra AN. Remote sensing studies of suspended sediment concentration variation in Barito Delta. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing; 2017. p. 12058.
2. Warren MB, Bullard SA. First elucidation of a blood fluke (Electrovermis zappum n. gen., n. sp.) life cycle including a chondrichthyan or bivalve. Int J Parasitol Parasites Wildl. 2019; 10:170–83.
3. Duong HH, Phong ND, Ha TL, Tang TD, Trinh TN, Nguyen TM, et al. Application of Machine Learning to Forecast Drought Index for the Mekong Delta. 2024;
4. Thuy VTT, Ha NTT, Koike K, Thao NTP, Trung PN, Thanh DX. Estimating Water Content and Grain Size of Intertidal Flat Sediments Using Visible to Shortwave-Infrared Reflectance and Sentinel 2A Data: A Case Study of the Red River Delta, Vietnam. IEEE J Sel Top Appl Earth Obs Remote Sens. 2022; 15:2696–708.
5. Pearson K. X. Contributions to the mathematical theory of evolution. —II. Skew variation in homogeneous material. Philos Trans R Soc London(A). 1895;(186):343–414.
6. Howitt D, Cramer D. Introduction to statistics in psychology. Pearson education; 2008.
7. Freedman DA. Statistical models: theory and practice. cambridge university press; 2009.
8. Stangor C. Research methods for the behavior science. Cengage Learning; 2011.
9. Ioannidis Y. The history of histograms (abridged). In: Proceedings 2003 VLDB Conference. Elsevier; 2003. p. 19–30.
10. Zhang Y, Lin S, Liu J, Qian X, Ge Y. Time-series MODIS image-based retrieval and distribution analysis of total suspended matter concentrations in Lake Taihu (China). Int J Environ Res Public Health. 2010;7(9):3545–60.

How to cite this article: Danh-tuyen Vu, Tien-thanh Nguyen, Anh-huy Hoang. Linear regression-based time-series prediction of total suspended solids in the Day River Basin, Vietnam. *International Journal of Research and Review*. 2025; 12(8): 482-489. DOI: *https://doi.org/10.52403/ijrr.20250857*

******