Everest Advances in Science and Technology (EAST), Vol. 2, No. 1, 2026 (Special Issue)
ISSN: 3102-0410 (Print), 3102-0429 (Online)
DOI: https://doi.org/10.3126/east.v2i2.99464

Suspended Sediment Concentration and Load Prediction in the West Rapti River Basin: A Comparative Study of Machine Learning Approaches

SUWAJ ARYAL[1*], RAM KRISHNA REGMI[2]

Abstract. It is very important to accurately forecast suspended sediment concentration for effective management of reservoir sedimentation and design of hydropower turbines in Siwalik river systems. This research compares the performances of six different machine learning algorithms (ANN, LSTM, MLP, XGBoost, RF, and SVR) in forecasting daily suspended sediment concentration (SSC) and sediment load (SSL) in the West Rapti River Basin, Nepal. Predictors used in this research are flow, rainfall, temperature, particle size distribution, antecedent precipitation indices, NDVI, soil moisture, hydrological flashiness indices, and autoregressive SSC lags. The dataset includes information for 8 water years (WY2016–WY2023), and it is chronologically split into three parts, with WY2022 being selected as the held-out validation period and WY2023 as the test period. It was found that tree and kernel-based algorithms have the best SSC forecast performances during the test phase, and RF ($R^2 = 0.646$) and XGBoost ($R^2 = 0.635$) are the best SSC forecasters. SSL forecast performance has greatly improved in most algorithms, but SVR yields the best SSL performance ($R^2 = 0.894$). The importance of the predictors was estimated using permutation (for ANN, LSTM, MLP, and the SVR kernel model), Gini/MDI (for RF), and Gain (for XGBoost). In all considered algorithms, autoregressive SSC lags are the most important predictor (importance between 48 and 95%), yielding 87–95% importance in the deep learning and kernel-based models. The elimination of the lag predictor decreases the performance of the best performing algorithm (RF) from 0.646 to 0.172 – slightly better than the performance of the sediment rating curve approach (0.155). It means that the current set of hydrometeorological predictors explains only about 17% of the direct variability in SSC. Ultimately, the moderate R^2 ceiling of around 0.65 across such structurally diverse models, combined with this massive drop in performance without the lag, proves that our ability to predict SSC is limited by the input data itself rather than the model architectures. This highlights a critical need to include actual sediment supply indicators in future studies.

Keywords: ANN, West Rapti River, LSTM, Machine Learning, Random Forest, Suspended Sediment, SVR, XGBoost.

[1,2] Institute of Engineering, Lalitpur, Nepal
E-mail: suwajaryal2016@gmail.com
* Corresponding author

Manuscript received: 12 June 2026; Revised: 15 July 2026; Accepted: 5 August 2026.

1. INTRODUCTION

Nepal has a theoretical hydropower potential of 83,000 MW which are possible due to its steep gradients and abundant river systems of the Siwalik region [1]. But it comes with the high amount of sediment. These regions are one of the most severe suspended sediment regimes on Earth. In Nepal, about 77%–86% of the total yearly rainfall occurs during the monsoon, confined to a period of just 120 days. The monsoon causes high SSC values, sometimes exceeding 100,000 mg/L, in the highly flashy Siwalik Rivers during monsoon flooding [2, 3]. Globally, Siwalik Rivers are among the largest contributors to oceanic suspended sediment flux. It has high sediment contribution relative to its global area coverage [4]. This high sediment flux threatens reservoir storage capacity, as it decreases the live storage of the reservoir affecting the energy production capacity of the project. It also affects the reservoir through accelerated sedimentation and causes progressive hydro-abrasion of turbine runners, runner buckets, and guide vanes. So, modelling accurate SSC and sediment load (SSL) prediction is an important engineering concern for sustainable hydropower operation, especially in Nepal [5].

The empirical Sediment Rating Curve (SRC) has historically served as the baseline prediction tool for sediment flowing in the river. It is the power law relation between the sediment amount and the discharge on the river. However, the SRC systematically underestimates sediment transport outside its calibration range, especially during extreme monsoon floods where values outside data ranges occur [6]. Machine learning (ML) models have increasingly been applied to suspended sediment prediction. These models offer flexible non-linear mappings which outperform the SRC on aggregate metrics [7, 8, 9, 10, 11].

Despite these statistics, a core critical gap remains. Most of the reported ML accuracy derives from autoregressive features (lagged SSC), a dependency that inflates performance metrics while restricting applicability to ungauged catchments or multi-day forecasting [12, 13]. Besides, the SSL is a much more direct measure of reservoir trapping efficiency and turbine erosion than SSC; nevertheless, a very limited number of comparative analyses have been carried out to test the ability to predict SSC and SSL through various structures. Within Nepal, previous studies of the West Rapti basin were focused on hydrological simulation [14], leaving ML-based daily SSC and SSL prediction untouched.

This study addresses these gaps by providing a comparison of six Machine Learning architectures applied to predict the daily SSC and SSL in the West Rapti river basin, based on a chronological expanding-window cross-validation framework. The specific objectives of this study are: (i) to compare predictive accuracy for SSC and SSL across six ML architectures against a standard baseline of Sediment Rating Curve; (ii) to quantify the autoregressive lag dependency in each model through permutation feature importance; (iii) to isolate the maximum predictive power of the physical hydrometeorological feature set by evaluating model performance under a no-lag condition; and (iv) to assess whether SSL prediction outperforms SSC prediction and identify the conditions under which this occurs.

2. STUDY AREA AND DATA

The research area is set in the Jalkundi station, which is an important station, part of the West Rapti River, found in the mid-western region of Nepal. As can be seen from the location map in Figure 2, the catchment area straddles the provinces of Lumbini and Karnali Pradesh. The focus of hydrology in this particular region will be concentrated around the gauging station (No. 360) of the Jalkundi Catchment.

The collected dataset spans Water Years (WY) 2016–2023, yielding 1,273 raw daily SSC records. The term water year represents the yearly cycle hydrological cycle starting from April instead of January to match the Nepalese calendar. Eight records (one per water year) were excluded from modelling because the autoregressive SSC lag-1 feature is not present for the first day of each year. This leaves 1,265 observations used in model training and evaluation. The chronological split was: WY2016–2021 for training (876 observations), WY2022 for validation (173 observations), and WY2023 for independent testing (216 observations). Table 1 contains the sources for data and the

derived features. Spearman rank correlation multicollinearity screening ($|r| < 0.80$ threshold) led to removing 7 of the original 20 variables due to high intercorrelation ($|r| \geq 0.80$): stream power VS ($|r| = 0.88$ with Q), SPI-30 ($|r| = 0.95$ with API), Q lag-1, Q lag-2, SSC lag-2, rainfall lag-1, and maximum daily rainfall. The final selected 13-feature set includes: discharge (Q), grain size (d_{50}), daily rainfall, temperature, soil moisture, NDVI, rate of discharge change (ΔQ), Q-ratio, antecedent precipitation index (API), 3-day rolling rainfall ($P_{3\text{paj}}$), cyclical month encoding (sin / cos), and SSC lag-1.

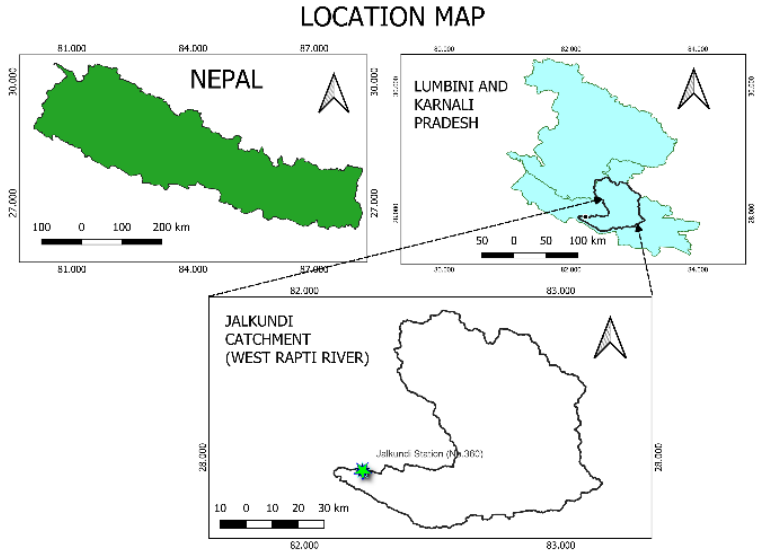

FIGURE 1. Location Map for Jalkundi Station

TABLE 1. Raw Data Sources and Derived Input Features

Source	**Raw Data**	**Derived Feature / Purpose**
DHM	Daily SSC	Target variable; SSC lag-1 (1 day sediment lag, autoregressive feature)
DHM	Daily Q (m^3/s)	Transport capacity: ΔQ; Q-ratio; stream power (VS)
CHIRPS	Precipitation	Rainfall; API (Antecedent Precipitation Index); P_3-day MA (3 days moving Average); SPI-30
Google Earth Engine	MODIS NDVI; SMAP; LST	Moderate Resolution Imaging Spectroradiometer, Normalized Difference Vegetation Index, Soil Moisture Active Passive, Land Surface Temp
SRTM DEM	30 m DEM	Channel slope S (for stream power)
Field (Primary)	Sieve analysis	Median grain size d_{50} (mm)

3. METHODOLOGY

3.1. **Machine Learning Models.** Six models spanning different learning paradigms were used in this study. Each model had the same 13-feature input selected after collinearity scanning. All models were hyperparameter-tuned using Optuna TPE sampler [16] for the expanding-window cross-validation objective. For input standardization, a robust scaler function was adopted.

3.1.1. *Artificial Neural Network (ANN).*) Artificial Neural Network (ANN): Fully-connected, feed-forward model, having input layer with a size of 224, a hidden layer with 16 neurons and an output layer of one neuron. For nonlinear activations Rectified Linear Unit (ReLU) is used. The optimization criterion is the minimization of Huber loss with $\delta = 0.00386$. Training is performed with Adam optimization method, gradient clipping and early stopping criteria for 60 and up to 500 epochs respectively.

3.1.2. *Long Short-Term Memory (LSTM).* Bi-directional LSTM with three layers, hidden states having increased dimension from 64 to 128 due to bi-directionality and an attention mechanism for a 30 days period[17]. LSTM is fed by a series of 30 sequential days with feature vectors of size 13, where SSC lag-1 represents the SSC on the previous day at each series step, and other 12 variables provide the hydro meteorological context of this particular day. SSC target is transformed with the $\log_{10}$ before training. Loss function is the Huber loss with $\delta = 0.8$, while the optimization is done using AdamW optimizer and ReduceLROnPlateau learning rate scheduler.

3.1.3. *Multilayer Perceptron (MLP).* Residual MLP architecture with projection layer (13 to 256 units) followed by 6 residual blocks (BatchNorm, LeakyReLU, dropout 0.217, skip connections). Trained with pure Huber data loss, serving as an ablation baseline for the residual architecture [18].

3.1.4. *Extreme Gradient Boosting (XGBoost).* 900 boosting iterations with histogram-based tree learning (max depth 8, subsample 0.611, colsample_bytree 0.963), L1 regularization ($\alpha = 2.025$), L2 regularization ($\lambda = 7.123$), learning rate 0.00568 [19].

3.1.5. *Random Forest (RF).* 100 decision trees trained on bootstrap samples with tree depth 10, minimum split 5, and maximum features fraction 0.7 at each node [20]. The final prediction for the Suspended Sediment Concentration (SSC) is calculated using the average of the tree outputs.

3.1.6. *Support Vector Regression (SVR).* Radial Basis Function kernel defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{1}$$

where $\gamma = \text{scale}$, Regularization $C = 8.0$, and margin $\varepsilon = 0.05$ [21].

3.2. **Sediment Load Derivation.** Daily suspended sediment load (SSL, t/day) is derived using predicted SSC and measured Discharge:

$$\text{SSL} = \text{SSC} \times Q \times 0.0864 \tag{2}$$

where Q is observed daily mean discharge (m^3/s) and 0.0864 is the unit conversion factor. . For the LSTM model, which applies $\log_{10}$ transformation to SSC targets, the Duan smearing correction factor [22] was applied during retransformation. It remove systematic back-transformation bias, which may aries during retransformation. And SSL is a post calculated product from SSC rather than a direct model predicted output, all models were trained and validated on SSC only. SSL performance is therefore evaluated exclusively at the test phase (WY2023). So, Computing SSL metrics at training or validation phases would be misleading, and having related to discharge, it seems as a data leaked model, which should be avoided.

3.3. **Validation Framework.** The walk-forward cross-validation is carried out in five expanding windows (five folds), where each fold adds a new year. In Fold 1, the training window is 2 years long (April 1, 2016 – March 31, 2017), each fold increasing by one year until Fold 5 (April 1, 2016 – March 31, 2021, 6 years). The next year following the training window is the validation set for hyperparameter tuning using Optuna, with the year after serving as the independent test set for that fold. The hyperparameter optimization is performed on different fold sets than the one used for CV-R^2 assessment in order to avoid optimization leakage into the CV statistics. This approach strictly preserves the temporal ordering and helps to avoid the data leakage due to the randomness of the random k-fold cross-validation on the SSC time series [23].

3.4. **No-Lag Ablation Experiment.** In order to evaluate the predictive power of the physical hydrometeorological features without the effect of the sediment memory, the most successful SSC model (Random Forest, $R^2 = 0.646$) is retrained with SSC lag-1 excluded from the 12-feature set. This ablation quantifies the maximum predictive capacity of the remaining input features i.e. discharge, rainfall, soil moisture, NDVI, temperature, grain size, and derived featured.

3.5. **Performance Metrics.** The model performances were quantified using the Coefficient of Determination (R^2), as a main evaluation metrics, which is computed as: (R^2 / NSE):

$$R^2 = 1 - \frac{\sum(\text{obs}_i - \text{pred}_i)^2}{\sum(\text{obs}_i - \overline{\text{obs}})^2} \tag{3}$$

Which is also numerically equivalent to Nash–Sutcliffe Efficiency. The other evaluation metrics are Root Mean Square Error (RMSE), Percent Bias (PBIAS) and Standard Deviation ratio. The Standard Deviation Ratio quantifies variance compression, computed as:

$$\text{SD ratio} = \frac{\text{predicted SD}}{\text{observed SD}} \tag{4}$$

4. RESULTS AND ANALYSIS

4.1. **SSC Prediction Performance.** Table 2 shows test-year (WY2023) SSC performance metrics for all six models alongside the SRC baseline ($R^2 = 0.155$). SRC suboptimality for the Jalkundi case study is associated with the pronounced variability of the discharge-SSC relation at various sites in the Siwalik catchments. In particular, there is a notable variability of the SRC R^2 at various sites, with R^2 being as high as 0.8 in other sites in Nepal [24]. All ML models substantially outperform the SRC benchmark. Among the ML models, the tree-based ensemble approaches deliver the highest SSC performance. Specifically, the R^2 value for Random Forest is 0.646, RMSE is 439 mg/L, and PBIAS is −5.1%. XGBoost is very close to Random Forest in terms of R^2 (0.635), RMSE (446 mg/L), and PBIAS (−7%). Both LSTM and SVR deliver similar R^2 (0.611 and 0.610, correspondingly). Despite that these models have different learning paradigms, sequential time-series learning in LSTM and kernel-based learning in SVR, both of them deliver nearly equal performance. ANN delivers $R^2 = 0.570$, while the Residual MLP demonstrates the worst performance in SSC prediction ($R^2 = 0.477$). It may be associated with the fact that the Residual MLP has the most complex architecture (394,000 parameters) and cannot benefit from the limited 876-element training sample [25].

A core finding is that there is a convergence of six models with a wide range of different architectures into a narrow interval of R^2 from 0.477 to 0.646. Tree-based models, deep learning models, and kernel-based models, each of which has an inherently different inductive bias, achieve the same performance ceiling. Model-architecture independence implies that the SSC feature set is constraining the performance more than the architecture itself.

Among neural architectures, LSTM has the minimum PBIAS (−3.24%), while the ANN approach is able to preserve the mass balance well. However, SVR, which shows the highest PBIAS (−9.80%) among models with good performance, delivers the highest load prediction performance (see Section IV-B). SD ratio ranges from 0.673 (Residual MLP) to 0.773 (LSTM), implying that the most optimal model explains only 77% of the variability of Siwalik SSC, and thus all models underestimate the SSC variance. The universal negative PBIAS in all models (from −3.24% to −10.81%) is caused by the inherent underestimation of SSC peaks due to the rare high-concentration SSC cases in the training dataset.

TABLE 2. Test-Year (WY2023) SSC Performance for All Six Models and SRC Baseline

Model	R^2	**RMSE (mg/L)**	**PBIAS (%)**	**SD Ratio**	**CV Mean R^2**
SRC (Baseline)	0.155	–	–	–	0.155
Random Forest	0.646	439.1	−5.06	0.771	0.460
XGBoost	0.635	445.8	−3.77	0.771	0.431
LSTM	0.611	447.8	−3.24	0.773	0.144
SVR	0.610	460.9	−9.80	0.744	0.392
ANN	0.570	483.8	−7.50	0.716	0.332
Residual MLP	0.477	533.3	−10.81	0.673	$-0.922^{\dagger}$

†*Residual MLP CV mean $R^2 = -0.922$ is driven by catastrophic performance in Fold 1 (2 training years), the 394,000-parameter model hugely over fits the limited training data, resulting in predictions worsen than the training mean. Performance improves as we increase the data in later folds.*

The average R^2 value obtained via five-fold cross-validation (Table II, last column) demonstrates definite distinctions in generalization stability which cannot be observed via the single split analysis. The tree-based methods (RF: 0.460, XGBoost: 0.431) and SVR (0.392) demonstrate excellent cross-validation performance, which confirms the relevance of the WY2023 test results. However, the relatively weak cross-validation mean R^2 of the LSTM model (0.144) and the pathological cross-validation mean of the Residual MLP (−0.922, caused by the catastrophic performance on Fold 1 with just two training years) illustrate the high dependence of DL methods on the number of observations used for training. By comparison, the ANN demonstrates average instability with the mean R^2 value of 0.332.

4.2. **Sediment Load (SSL) Prediction Performance.** Table 3 presents SSL prediction metrics for the test year WY2023. Notably, a significant change in ranking among models accompanies a considerable increase in the SSL R^2 value in comparison with their SSC counterparts. Specifically, Support Vector Regression provides the highest SSL $R^2 = 0.894$ (RMSE = 7,297 t/day), followed by Random Forest ($R^2 = 0.892$, RMSE = 7,393 t/day) and XGBoost ($R^2 = 0.876$). The SSL $R^2 = 0.849$ is achieved by the LSTM approach, showing a considerable increase compared to its SSC $R^2 = 0.611$. ANN shows a similar SSL and SSC $R^2 = 0.570$, suggesting that there is no advantage of load calculation based on converted SSC. It should be stressed that the SSL R^2 value of the Residual MLP drops drastically down to 0.154, making this model significantly worse than all others in terms of load prediction and corresponds to a pathological SD ratio = 1.564. The systematic SSL improvement observed for almost all models (SSC $R^2 \approx 0.61$–0.65, SSL $R^2 \approx 0.85$–0.89) is explained by the equation used for load calculation: SSL = SSC × Q × constant. As discharge Q is included in model inputs and is quite predictable, a product convolution makes prediction errors less influential. High-SSC-low-Q events, where SSC is mispredicted, have a relatively small effect on the load integral, while high-discharge events with even moderate SSC prediction lead to the large load contribution. The exception is the Residual MLP, in which extremely overfitted SSC predictions become more exaggerated when multiplied by discharge and produce erroneous load predictions. From the perspective of engineering applications such as annual reservoir trap efficiency calculation and sediment budget computations, SSL R^2 values of 0.87–0.89 for tree-based models and SVR indicate good predictive capability. Annual load bias in Table III is estimated by PBIAS column: for one test year, PBIAS and annual load error are equivalent and equal to (cumulative predicted load − cumulative observed load)/cumulative observed load × 100. For all considered years, annual load biases stay within −6% for Random Forest (−5.94%) and XGBoost (−5.15%). For all six models, annual load biases do not exceed 15%; specifically, a failure of Residual MLP is shown by the near-zero SSL $R^2 = 0.154$, meaning that its load predictions are not correlated with the observed signal. The latter does not necessarily mean that the total load volume is inaccurate, but the structure of load prediction is completely distorted. The results suggest practical applicability of tree-based ensemble and SVR models for operational annual load

estimation. Conversely, this performance improvement fails when severe SSC overfitting produces extreme, unstructured predictions; in these cases, multiplication by discharge amplifies rather than smooths the errors, completely distorting the load prediction structure despite a deceptively low overall volume bias. Notably, the SSL ranking (SVR → RF → XGBoost → LSTM) differs from the SSC ranking (RF → XGBoost → LSTM → SVR). SVR's rise to first place in SSL reflects its characteristic behavior: the ε-insensitive tube loss suppresses predictions of extreme instantaneous SSC spikes, producing a smoother SSC series. While this lowers SSC R^2, the smoothed predictions integrate better over discharge to yield accurate cumulative loads, particularly when extreme SSC outliers (which dominate load totals) are predicted directionally correctly even if underestimated in magnitude.

TABLE 3. Test-Year (WY2023) Sediment Load (SSL) Performance for All Six Models

Model	**SSL R^2**	**SSL RMSE (t/day)**	**PBIAS (%)**	**SD Ratio**
SVR	0.894	7,297	−13.15	0.895
Random Forest	0.892	7,393	−5.94	1.005
XGBoost	0.876	7,889	−5.15	0.990
LSTM	0.849	9,045	−7.35	0.913
ANN	0.570	14,722	−5.72	1.309
Residual MLP	0.154	20,635	−6.05	1.564

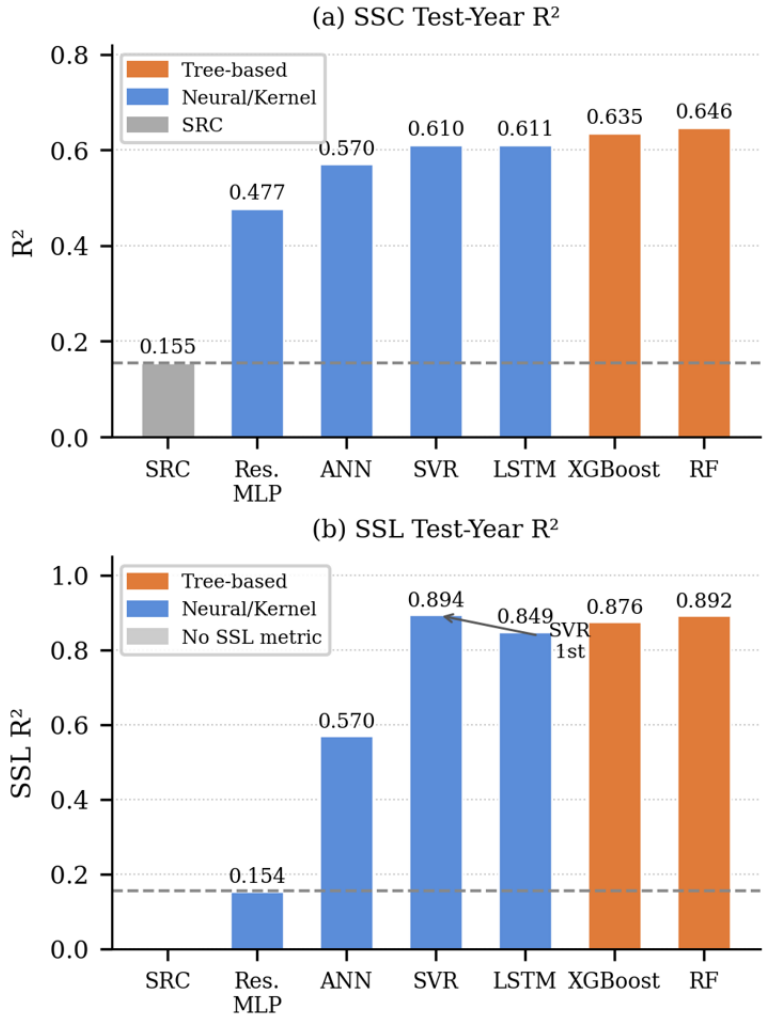

FIGURE 2. Performance of all six ML models for SSC prediction across Training and Testing phases

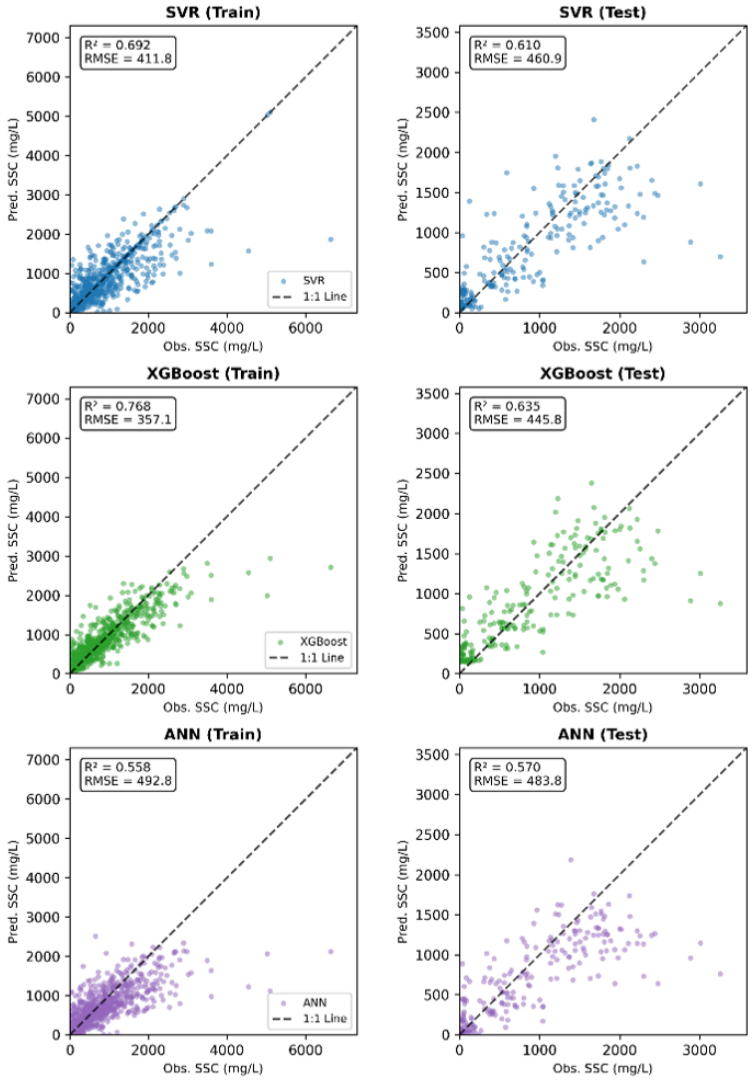

FIGURE 3. Dense Scatter Plot of SVR, XGBoost and ANN for SSC prediction.

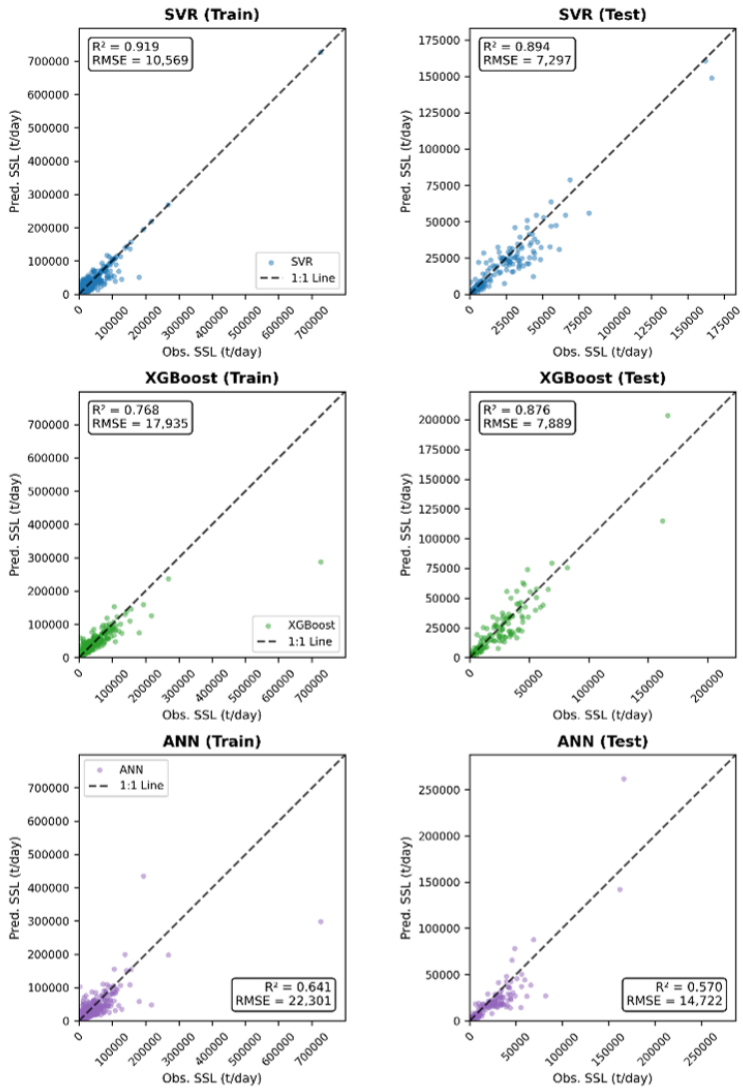

FIGURE 4. Scatter Plot for SVR, XGBoost and ANN for SSL prediction.

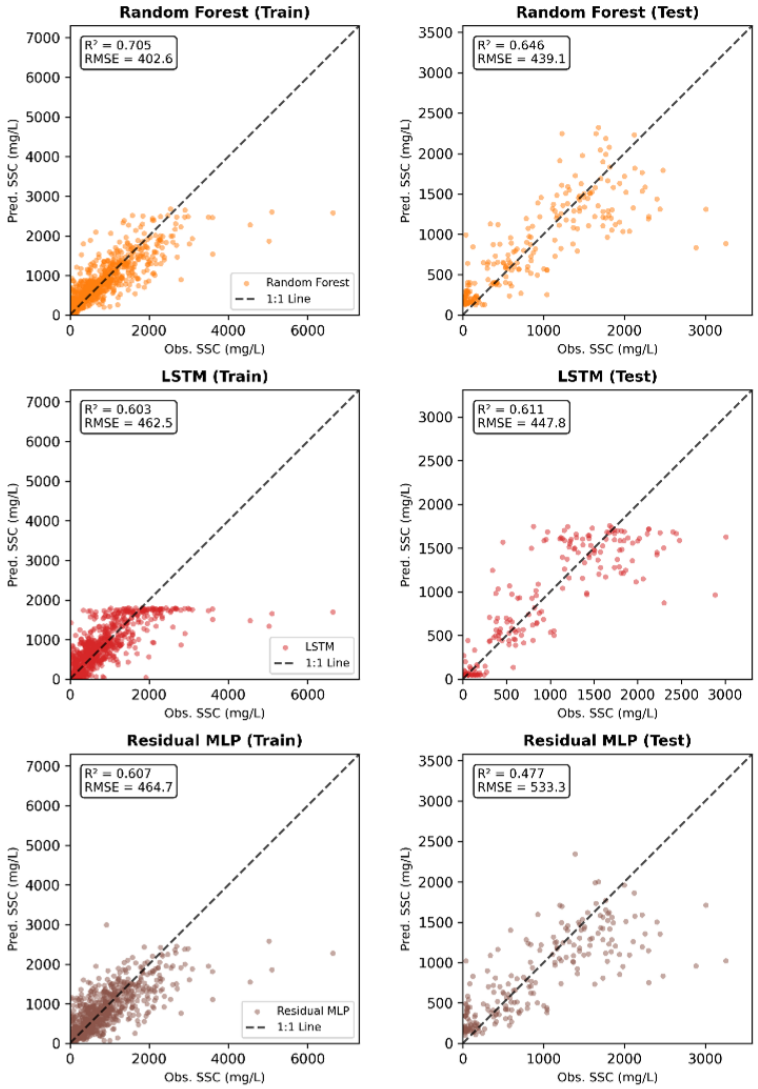

FIGURE 5. Dense Scatter Plot for Random Forest, LSTM and Residual MLP for SSC prediction.

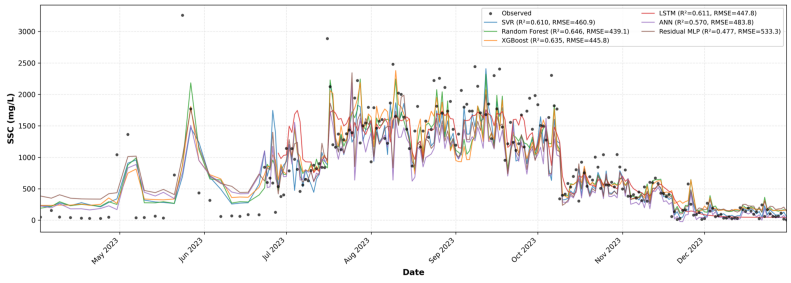

FIGURE 6. Test Year Hydrograph for SSC prediction.

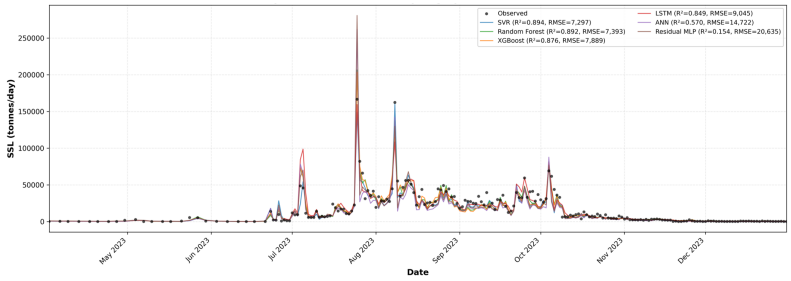

FIGURE 7. Suspended Sediment Load Hydrograph for test year for all six models.

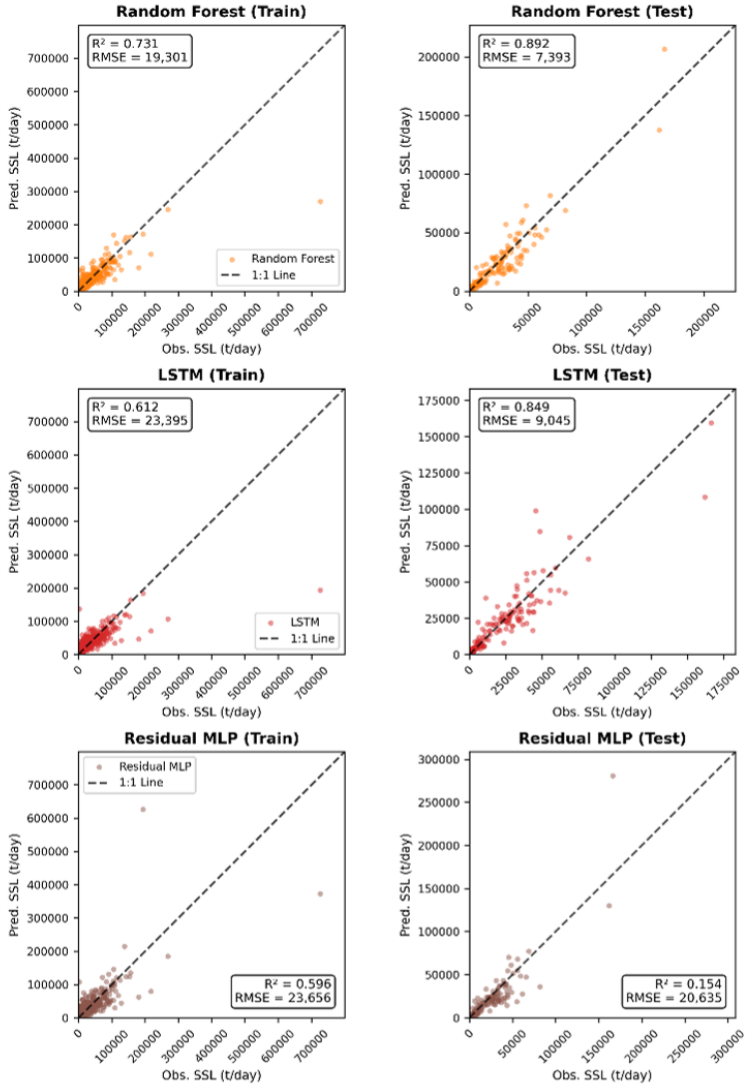

FIGURE 8. Scatter Plot for Random Forest, LSTM and Residual MLP for SSC prediction.

4.3. **Feature Importance and Lag Dependency.** Table 5 presents the normalized SSClag-1 feature importance for all six models. It shows that the autoregressive lag term dominates predictive contribution. It varies with model, but have high importance were seen on each of the models. SSC,lag-1 is the dominant predictor in every model, ranked first across all six architectures. Direct numerical comparison of importance magnitudes across models should be interpreted with caution, as Gini/MDI (RF), Gain-based (XGBoost), and permutation importance are not on the same scale; the consistent first-place rank across all six methods is the primary conclusion. Deep learning and kernel models attribute 87–95% of predictive importance to this single feature. Tree-based models show lower but still dominant lag importance (XGBoost: 48%, Random Forest: 69%), reflecting their ensemble structure's greater capacity to distribute information gain across multiple features. In XGBoost, discharge-related features (Q, stream power VS, ΔQ) account for the remaining $\sim 52\%$ of predictive gain, which partially explains why tree-based models maintain better cross-validation stability: they distribute learning more broadly across physical drivers rather than concentrating it in the autoregressive term. The physical interpretation is that SSClag-1 encodes the basin's sediment supply state: previous-day SSC captures progressive exhaustion of available sediment during sustained monsoon events and the replenishment between events, a process well documented in fluvial catchments exhibiting clockwise discharge–SSC hysteresis [26]. The models are therefore primarily functioning as autoregressive predictors rather than process-based hydrometeorological models [27], [28].

Table 4. Observed and Predicted Suspended Sediment Load (tons) for Calibration, Validation, and Testing Water Years

Phase	**Observed**	**SVR**	**RF**	**XGBoost**	**LSTM**	**ANN**	**Res. MLP**
Calibration	16,303,105	15,574,898	16,102,905	15,828,163	14,793,994	14,216,262	15,986,680
Validation	5,105,651	3,787,259	4,486,353	4,435,403	3,727,612	3,611,269	4,270,256
Testing	3,548,393	3,180,928	3,337,512	3,365,573	3,288,254	3,345,260	3,333,368

Table 5. SSC lag-1 Normalized Feature Importance and Method by Model

Model	**SSC lag-1 Importance**	**Rank**	**Method**
LSTM	0.951	1st	Sequence Permutation
Residual MLP	0.892	1st	Permutation
ANN	0.890	1st	Permutation
SVR	0.872	1st	Permutation
Random Forest	0.686	1st	Gini (MDI)
XGBoost	0.483	1st	Gain-Based

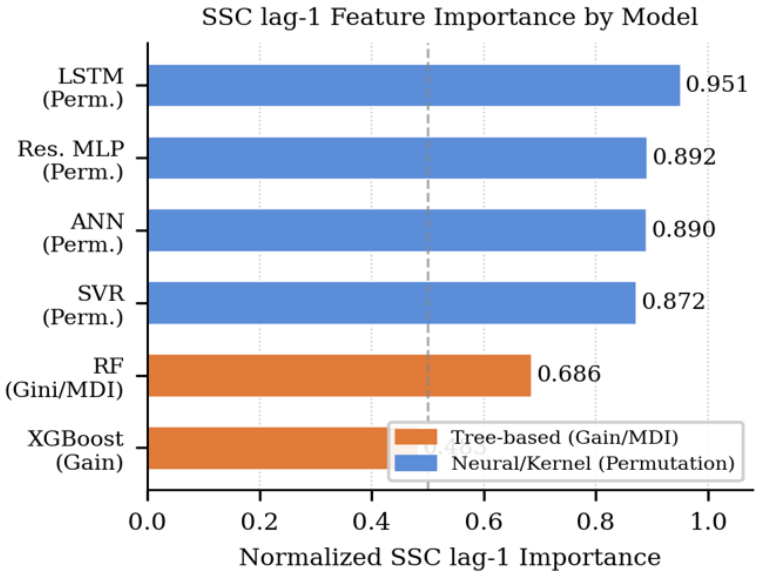

FIGURE 9. Lag feature (SSC lag 1) importance in different models

4.4. **No-Lag Condition: Maximum Physical Feature Predictive Power.** Table 6 presents the results of the no-lag ablation experiment. It shows that Removing SSC lag-1 from the best-performing Random Forest model produces a catastrophic failure of models, especially on validation and test phase. The training R^2 remains stable (0.709). it tells that the lag feature contributes only generalizable information rather than remembering the memories. But for validation and test phase, the results were memories based rather than prediction, which is the main reason the model failed on those two sets. Coefficient of determination (R^2) reduces to 0.172 from 0.646 after eliminating the lag-1 of SSC, implying a reduction of 73.4% in explanatory power. The performance without lag effect ($R^2 = 0.172$) is only slightly higher than the baseline SRC model ($R^2 = 0.155$) that uses discharge alone. This is symptomatic of the fact that the full 12 predictor hydrometeorological combination (discharge, precipitation, temperature, soil moisture, NDVI, grain size, ΔQ, API, P3, Q-ratio, and monthly cycle) contributes just 0.017 variance to the explained variance ratio more than the single predictor SRC, calculated as $0.172 - 0.155$. The PBIAS deteriorates to -23.06%.

This result establishes a critical bound: the maximum predictive capacity of the current physical feature set for direct SSC variance explanation is approximately 17%. The remaining performance gap between the no-lag result (17%) and the full-model result (65%) is entirely attributable to the autoregressive lag term, confirming that model accuracy is driven by sediment memory rather than physical process representation. The ceiling at $R^2 \approx 0.65$ cannot be broken by switching model architectures; it requires expanding the physical feature set to include sediment supply drivers not captured by current inputs. for example, real-time indicators of upstream landslide activity, bank erosion events, or satellite-derived turbidity proxies that directly measure sediment availability.

TABLE 6. Random Forest SSC Performance: With vs. Without SSC lag-1 (WY2023 Test Year)

Condition	**Train R^2**	**Val R^2**	**Test R^2**	**RMSE (mg/L)**	**PBIAS (%)**
With SSC lag-1	0.705	0.345	0.646	439.1	−5.06
Without SSC lag-1	0.709	0.178	0.172	671.2	−23.06
SRC Baseline	–	–	0.155	–	–

5. DISCUSSION

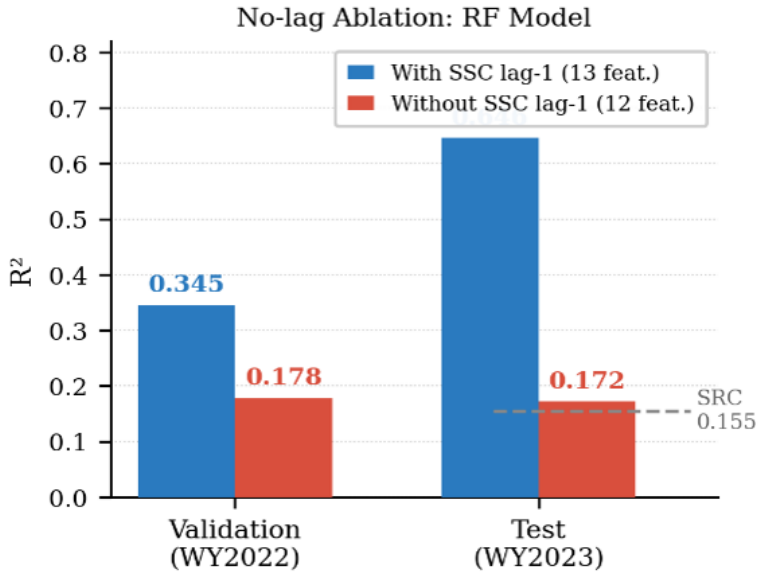

FIGURE 10. No lag performance (SSC lag 1) analysis for random forest model

The most operationally significant finding of this study is the divergence between SSC and SSL prediction quality. The SSC R^2 plateaus at approximately 0.65, whereas SSL R^2 for tree-based models and SVR reaches 0.87–0.89. For reservoir sedimentation and turbine abrasion management, sediment load is the primary concern rather than the sediment concentration. So, from these perspectives, it seems that the machine learning model has practical usefulness. The no-lag experiment provides the most direct evidence of the feature ceiling hypothesis. The hydrometeorological feature set as constituted (discharge, rainfall, temperature, soil moisture, NDVI, grain size, and their derived indices) is a transport-capacity-oriented set: it describes the capacity of the river to carry sediment in terms of stream power and flow energy [29], but not the availability of sediment to be transported. In supply-limited Siwalik catchments where SSC peaks are generated by episodic stochastic events (landslides, bank failures, debris flows during monsoon events), transport capacity features are necessary but not sufficient. The 17% direct predictive power from physical features is consistent with the supply-dominated transport regime of the Siwalik geology, where sediment supply variability, driven by random hillslope failures, is fundamentally unpredictable from the current feature set [30]. One more implication that arises from the absence of the lag effect involves the nature of the influence of the variables included in the dataset on suspended sediment concentration (SSC) at Jalkundi sub-catchment. The 13 variables were selected according to the assumption that transport capacity (discharge, rainfall, antecedent precipitation index, rate-of-change of discharge, stream power, soil moisture content, Q-ratio) and proxies for erosion processes (NDVI, grain size d_{50}) constitute the physical processes underlying daily SSC. However, the no-lag model, which takes into account all 12 of these variables, provides $R^2 = 0.172$, while the single predictor SRC taking into account just the discharge gives $R^2 = 0.155$. The contribution of the 11 additional features is just the additional increment of 0.017 in R^2. Rainfall, NDVI, soil moisture, API, and ΔQ, which would be independent predictors of erosion potential and antecedent state of the catchment if the transport capacity determined the mechanisms, do not make any contribution at the daily time scale. It is the basin-specific conclusion: discharge and rainfall are not the main drivers of the daily SSC variability at the Jalkundi station, contrary to the assumption that formed the basis of feature engineering. In a transport capacity limited system, higher discharge and more rainfall would reliably increase sediment mobilization, giving

independent information for the prediction without any lag. The minimal gain from 11 additional features beyond Q proves that the Jalkundi sub-catchment is not a transport capacity limited system at the daily time scale. Instead, SSC generation at this station is controlled by stochastic sediment supply events such as landslides, bank collapse, and debris flows resulting from Siwalik Hills instability, which are decoupled from discharge and rainfall. Hence, the improvement in the SSC prediction would require the inclusion of the direct indicators of sediment supply dynamics, but not more precise measures of the transport capacity. The cross-validation stability study draws attention to an important point in choosing the right model that might be lost on the single-split test performance scores. The tree-based ensembles display stable generalization for all five folds (CV R^2: RF 0.460, XGBoost 0.431). On the other hand, LSTMs demonstrate low CV mean (0.144) while the Residual MLP demonstrates a significantly negative CV mean (-0.922). This fact means that deep-learning models require caution in application to situations where fewer than about 4–5 water years of available data exist in the West Rapti watershed. While deep models can reach good performance scores for the last fold (6 years of training), they have unstable performance under the lack-of-data scenario, which characterizes new measurement points in the Siwalik chain. It is in line with Kratzert et al. [31] who showed that LSTM could model the long-term dependencies in hydrological time series, and higher performance corresponds to higher volume of training dataset. The model ranking reversal between SSC and SSL (SVR rises to first for load; Residual MLP collapses for load) has practical implications for model selection. If the operational objective is annual load estimation for reservoir management, SVR and RF are the preferred choices. If the objective is operational SSC monitoring with the lowest systematic bias, LSTM's PBIAS of -3.24% makes it the best choice among neural architectures. XGBoost offers the best overall compromise: second in SSC accuracy, third in SSL accuracy, and among the most cross-validation-stable architectures. Finally, the scope of conclusions that can be drawn from this study should be stated explicitly. The analysis is based on a single gauging station (Jalkundi, West Rapti basin) over an 8-year record. While the results provide consistent and internally coherent evidence, they are not sufficient on their own to draw definitive conclusions about optimal model selection for Siwalik SSC prediction in general. Results may differ at stations with different geology, catchment size, or monsoon exposure. Furthermore, performance metrics (R^2, NSE, RMSE, and PBIAS) are necessary but not sufficient evidence for operational deployment. A model performing well on these metrics in a retrospective study does not automatically qualify for real-time operational use, which additionally requires uncertainty quantification, robustness to data quality degradation, multi-year out-of-sample validation, and stakeholder consultation. The findings presented here should be understood as directional evidence to guide model selection and identify critical feature gaps, not as a prescription for immediate adoption.

6. CONCLUSION

This study systematically compared six machine learning architectures for daily suspended sediment concentration (SSC) and suspended sediment load (SSL) prediction in the West Rapti River basin, Nepal.

First, each machine learning model performed better than the Sediment Rating Curve (SRC) baseline for SSC prediction, with tree-based ensemble methods reaching an architectural performance ceiling ($R^2 = 0.47$–0.65). Second, most models performed significantly better on SSL prediction than SSC prediction, with the exception of the Residual MLP, where high parameter density amplified upstream errors during load integration. Third, ablation tests revealed that physical feature sets explained less than half of total variance; the additional 47% performance boost relied entirely on autoregressive SSC lag-1 terms, meaning these models function as statistical predictors requiring continuous in-situ monitoring. Finally, future frameworks should incorporate domain-specific knowledge—such as landslide detection algorithms or satellite turbidity proxies—to align with the theory-guided data science paradigm and break the current predictive ceiling.

7. ACKNOWLEDGMENT

The authors extend their sincere appreciation to the Department of Hydrology and Meteorology (DHM), Government of Nepal, for providing the hydrologic and sediment data of the West Rapti River. The authors would also like to thank the Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal for their academic assistance. Furthermore, the author would like to express their heartfelt thanks to his family members for their constant support and motivation.

8. DATA AVAILABILITY

Data on suspended sediment concentration and discharge recorded on a daily basis were sourced from the Department of Hydrology and Meteorology (DHM), Government of Nepal (DHM Station 360, Jalkundi). Availability of such data can be arranged following an official request made to DHM according to their data sharing protocol (`https://www.dhm.gov.np`). The daily CHIRPS version 2.0 precipitation data set can be freely accessed at `https://www.chc.ucsb.edu/data/chirps`. The MODIS NDVI (MOD13Q1) and SMAP soil moisture data sets were downloaded using Google Earth Engine (`https://earthengine.google.com`). The main grain-size data set, collected in the field, is available from the corresponding author on request.

References

[1] H. M. Shrestha, "Facts and figures about hydropower development in Nepal," *Hydro Nepal: J. Water, Energy Environ.*, vol. 20, pp. 1–5, 2017.

[2] S. R. Kansakar, D. M. Hannah, J. Gerrard, and G. Rees, "Spatial pattern in the precipitation regimes of Nepal," *Int. J. Climatol.*, vol. 24, no. 13, pp. 1645–1659, 2004.

[3] E. J. Gabet, D. W. Burbank, B. Pratt-Sitaula, J. Putkonen, and B. Bookhagen, "Modern erosion rates in the High Himalayas of Nepal," *Earth Planet. Sci. Lett.*, vol. 267, no. 3–4, pp. 482–494, 2008.

[4] D. E. Walling and D. Fang, "Recent trends in the suspended sediment loads of the world's rivers," *Global Planet. Change*, vol. 39, no. 1–2, pp. 111–126, 2003.

[5] B. S. Thapa, B. Thapa, and O. G. Dahlhaug, "Empirical modelling of sediment erosion in Francis turbines," *Energy*, vol. 41, no. 1, pp. 386–391, 2012.

[6] R. I. Ferguson, "River loads underestimated by rating curves," *Water Resour. Res.*, vol. 22, no. 1, pp. 74–76, 1986.

[7] N. AlDahoul, Y. Essam, P. Kumar et al., "Suspended sediment load prediction using long short-term memory neural network," *Sci. Rep.*, vol. 11, p. 7826, 2021.

[8] Z. Mohammadi-Raigani, H. Gholami, A. Mohamadifar, A. N. Samani, and B. Pradhan, "Using an interpretable deep learning model for the prediction of riverine suspended sediment load," *Environ. Sci. Pollut. Res.*, vol. 31, pp. 32480–32493, 2024.

[9] H. A. Afan, A. El-shafie, W. H. M. W. Mohtar, and Z. M. Yaseen, "Past, present and prospect of an AI-based model for sediment transport prediction," *J. Hydrol.*, 2016.

[10] O. Kisi, "Suspended sediment estimation using neuro-fuzzy and neural network approaches," *Hydrol. Sci. J.*, vol. 50, no. 4, pp. 683–696, 2005.

[11] M. Zounemat-Kermani, O. Kisi, M. Adamowski, and J. Adamowski, "Evaluation of data driven models for river suspended sediment concentration modeling," *J. Hydrol.*, vol. 535, pp. 457–472, 2016.

[12] C. W. Dawson et al., "A comparative study of ANN techniques for river stage forecasting," in *Proc. IEEE IJCNN*, Montreal, 2005, pp. 2666–2670.

[13] H. Lamane, L. Mouhir, R. Moussadek, B. Baghdad, O. Kisi, and A. El Bilali, "Interpreting machine learning models based on SHAP values in predicting SSC," *Int. J. Sediment Res.*, vol. 40, no. 1, pp. 91–107, 2025.

[14] S. Neupane and A. Pandey, "Hydrological modeling of West Rapti River Basin of Nepal using SWAT model," in *Advances in Water Resources Engineering and Management*, Springer, 2021.

[15] J. L. Mugnier et al., "The Siwaliks of western Nepal: I. Geometry and kinematics," *J. Asian Earth Sci.*, vol. 17, no. 5–6, pp. 629–642, 1999.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD*, 2019.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[21] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, vol. 9, 1997.
[22] N. Duan, "Smearing estimate: A nonparametric retransformation method," *J. Amer. Statist. Assoc.*, vol. 78, no. 383, pp. 605–610, 1983.
[23] T. Cerqueira, L. Torgo, F. Smital, and B. Ženko, "Evaluating time series forecasting models: An empirical study on performance estimation methods," *Mach. Learn.*, vol. 109, pp. 1997–2028, 2020.
[24] V. Dahal, S. Kunwar, S. Bhandari et al., "Analyzing sedimentation patterns in the Naumure Multipurpose Project (NMP) reservoir using 1D HEC-RAS modeling," *Sci. Rep.*, vol. 14, p. 22134, 2024.
[25] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," *World J. Gastroenterol.*, vol. 28, no. 5, pp. 605–607, 2022.
[26] C. E. M. Lloyd, J. E. Freer, P. J. Johnes, and A. L. Collins, "Testing an improved index for analysing storm discharge–concentration hysteresis," *Hydrol. Earth Syst. Sci.*, vol. 20, pp. 625–632, 2016.
[27] G. E. P. Box, "Box and Jenkins: Time Series Analysis, Forecasting and Control," Palgrave Macmillan, 2013.
[28] J. Quilty and J. Adamowski, "Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models," *J. Hydrol.*, vol. 571, pp. 209–235, 2019.
[29] C. T. Yang and C. C. S. Song, "Theory of minimum rate of energy dissipation," *J. Hydraul. Div. ASCE*, vol. 105, no. 7, pp. 769–784, 1979.
[30] A. El Bilali, Y. Brouziyne, O. Attar, H. Lamane, A. Hadri, and A. Taleb, "Physics-informed machine learning algorithms for forecasting sediment yield," *Environ. Sci. Pollut. Res.*, 2024.
[31] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using LSTM networks," *Hydrol. Earth Syst. Sci.*, vol. 22, pp. 6005–6022, 2018.
[32] A. Karpatne et al., "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2318–2331, 2017.