Proceedings of the 11th World Congress on New Technologies (NewTech'25)
Paris, France - August, 2025
Paper No. ICCEIA 124
DOI: 10.11159/icceia25.124

GIS-Enhanced River Stage Prediction for Ungauged Basins: A Case Study of the Upper Ping River Basin, Thailand

Thirasak Panyaphirawat[1], Chana Sinsabvarodom[1,*], Damrongsak Rinchumphu[1], Pheerawat Plangoen[1], Oleg Gorbunov[1]
[1]Department of Civil Engineering, Faculty of Engineering, Chiang Mai University
239, Huay Kaew Road, Muang District, Chiang Mai, Thailand
thirasak_p@cmu.ac.th; chana.sinsabvarodom@cmu.ac.th; damrongsak.r@cmu.ac.th;
pheerawat.p@cmu.ac.th; oleg_gorbunov@cmu.ac.th

Abstract – Accurate prediction of river stage in ungauged basins is essential for flood forecasting and water resource management, particularly in regions with limited observational data. This study investigates the integration of Geographic Information Systems (GIS) enhanced reanalysis data with machine learning (ML) techniques to predict river stage levels in the Upper Ping River Basin, northern Thailand. Five years of hourly hydrometeorological variables from the ERA5-Land dataset, combined with observed river stage measurements from the P.1 Hydrological Station, were used to train and evaluate four ML models: Random Forest Regression (RFR), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP). Bayesian hyperparameter optimization and five-fold cross-validation were employed to ensure robust model training and evaluation. Among the models, XGB achieved the highest accuracy with an R^2 score of 0.9882, followed by RFR (0.9792), while SVR and MLP exhibited lower performance and higher sensitivity to data variability. Feature importance was further examined using SHapley Additive exPlanations (SHAP), revealing that runoff and soil moisture variables contributed most significantly to model performance, while wind, temperature, and even precipitation showed comparatively lower influence. A feature exclusion analysis confirmed that removing the top 14 ranked features substantially reduced prediction accuracy, underscoring their importance. The results highlight the potential of interpretable ML models combined with high-resolution GIS-based data for reliable river stage prediction in data-scarce regions. This framework offers promising applications in real-time flood forecasting and hydrological decision-making.

***Keywords*:** River Stage Prediction, Geographic Information Systems, ERA5-Land Hourly Reanalysis, Random Forest Regression, Extreme Gradient Boosting, Support Vector Regression, Multi-Layer Perceptron

1. Introduction

River stage, also referred to as water level or stage height, is the height of the water surface above a predefined reference point, typically a local datum at a specific location along a river [1]. It serves as a fundamental measurement in hydrology and is crucial in determining river discharge through rating curves, which establish a relationship between the river stage and the volume of water flowing in the river. Monitoring and understanding river stage is essential in hydrology due to its critical role in water resource management, environmental conservation, and hazard mitigation [2]. For the safe and reliable design of hydraulic structures, probabilistic assessment techniques particularly those based on extreme value analysis are employed to estimate hydrological extremes, thereby informing critical design parameters in civil engineering applications [3-5]. In recent years, researchers have integrated Geographic Information Systems (GIS) data with machine learning (ML) algorithms to enhance the understanding and simulation of hydrological processes across a wide range of applications. GIS offers robust tools for capturing and analysing spatiotemporal data, including topography, land use, soil characteristics, atmospheric conditions, precipitation, and weather patterns, while ML algorithms leverage these datasets to uncover complex, nonlinear relationships and generate accurate predictions. A wide range of ML techniques have been applied and compared with traditional hydrological models in tasks such as rainfall–runoff simulation [6, 7], sediment transport modelling [8], and flood risk assessment [9-11], demonstrating their potential for improved performance and real-time decision-making in water resource management.

However, predicting river stage of ungauged basins remains challenging due to the absence of in-situ observations and spatial heterogeneity. Traditional and physically-based models often struggle without calibration, while machine learning approaches like Long Short-term Memory (LSTM) require extensive training data and are sensitive to data quality and variability [12]. River systems are influenced by multiple interdependent factors, including rainfall, terrain characteristics, soil infiltration, land use, and climatic variability. Traditional hydrological models often struggle to fully capture these complexities. Another major issue is data availability and quality. Reliable river stage prediction depends on extensive, high-quality datasets that include past river flow, precipitation, and other meteorological variables. Many regions, especially in developing countries, lack dense monitoring networks, leading to data gaps and inaccuracies. Even when data is available, inconsistencies due to instrument errors, missing values, and sensor malfunctions can introduce uncertainties in the predictions [13]. In this study, we utilize data from ERA5-Land hourly reanalysis in conjunction with diverse machine learning techniques to determine the most effective models and key predictive features for river stage forecasting in ungauged basins.

2. Material and methods

2.1. Study Area

The Upper Ping River Basin is geographically located in northern Thailand, primarily covering Chiang Mai Province and partially extending into neighbouring provinces such as Lamphun and Tak. In this study, river stage data are obtained from the P.1 Hydrological Station, situated at latitude 18.7876 and longitude 99.0044. The station, owned and operated by the Royal Irrigation Department under the Ministry of Agriculture and Cooperatives, plays a vital role in flood prediction and mitigation efforts for Chiang Mai City. The catchment area of the station is approximately 6,350 km² [14] and comprises six distinct upper basins, each characterized by varied landscapes, including high mountainous regions, rolling hills, and expansive floodplains, with individual hydrological stations as shown in Figure 1. However, to simulate the conditions of ungauged basins, only data from the P.1 Station was used for prediction. The area is subjected to a tropical monsoon climate, distinctly marked by wet and dry seasons. Annual rainfall averages between 900 and 1,900 mm [15], predominantly occurring from June to September, accounting for approximately 80% of the total yearly precipitation [10]. This seasonal rainfall distribution profoundly influences the basin's hydrological dynamics, particularly concerning flood events and overall water resource management.

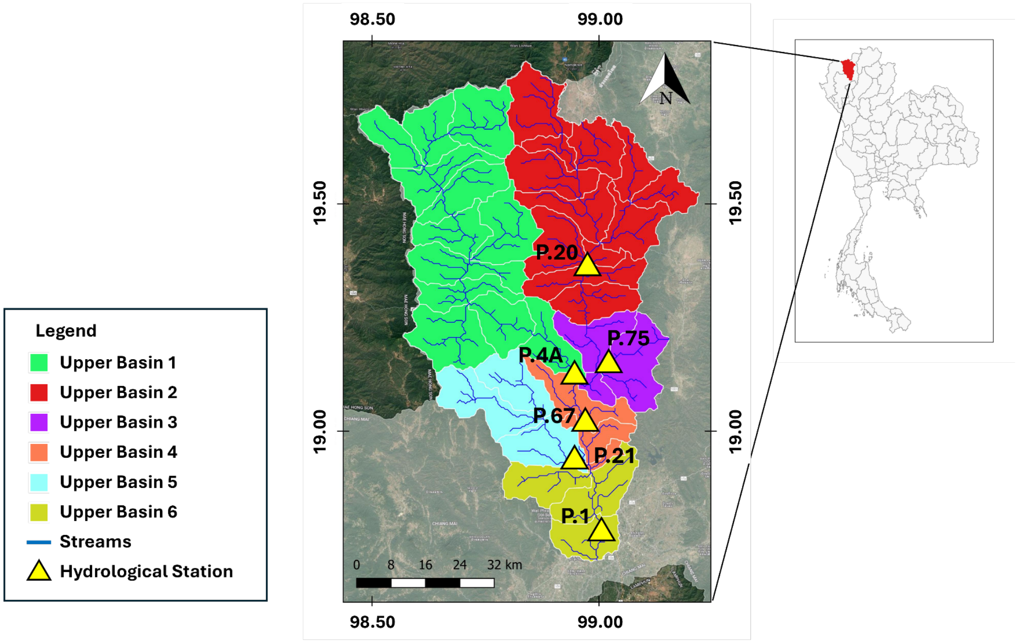

Fig. 1: Study Area.

2.2. Data

Five years of hourly river stage data, spanning from 2019 to early 2024, were obtained from the Royal Irrigation Department, Ministry of Agriculture and Cooperatives. As illustrated in Figure 2, the river stage demonstrates distinct seasonal and interannual variability, with pronounced peaks typically occurring during the monsoon seasons. The blue line represents the daily maximum gauge height, while the red dashed line indicates the critical flood threshold at 4.2 meters. During most of the observation period, water levels remained between 1.0 and 2.0 meters. However, several significant surges were recorded, including an extreme event during the 2022 monsoon season when the river level exceeded the flood threshold, resulting in inundation within the city.

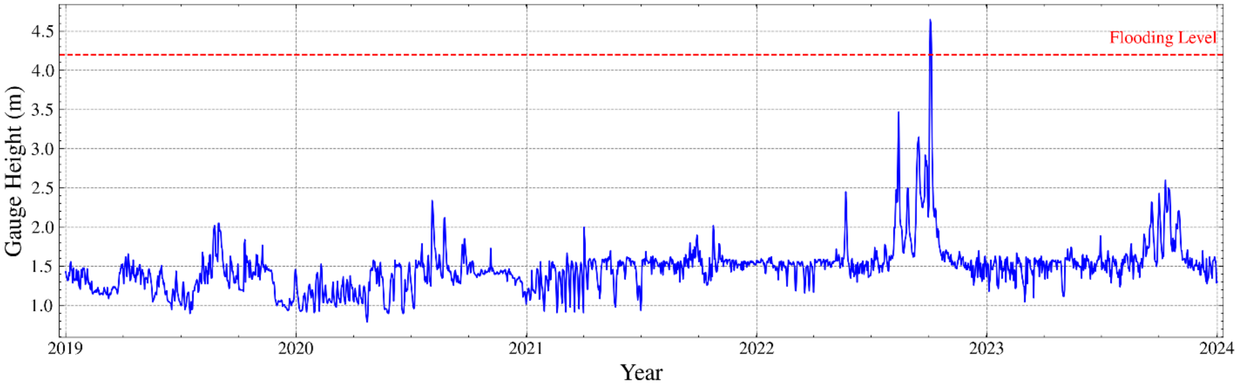

Fig. 2: P.1 Hydrological Station Daily Maximum Gauge Height.

ERA5-Land is a high-resolution, hourly reanalysis dataset developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), offering a consistent and detailed representation of land surface variables from 1950 to the present [16]. It comprises more than 60 variables at approximately 9 km spatial resolution (0.1° × 0.1°) and hourly temporal resolution, covering essential hydrometeorological fields such as multi-depth soil moisture and temperature, surface fluxes, radiation, precipitation, wind, and runoff. In this study, 34 variables relevant to the hydroclimatic characteristics of the study area were selected, excluding snow and lake-related parameters due to their limited relevance in the region. The selected variables were spatially aggregated within the study catchment using summary statistics tailored to their physical nature, precipitation was aggregated using a summation method, while temperature and similar variables were aggregated using the mean. These processed variables were subsequently employed as input features for the machine learning models. A detailed list of the ERA5-Land variables utilized is presented in Table 1. The dataset was partitioned into training/validation and testing sets, with 80% of the data allocated for training and validation, and the remaining 20% reserved for final testing. Within the training and validation set, five-fold cross-validation was employed to optimize model performance and prevent overfitting. Once the optimal hyperparameters were identified, the best-performing model configuration was evaluated on the unseen test set. The dataset spans from January 1, 2019, at 00:00 hours to December 31, 2023, at 23:00 hours, comprising a total of 43,824 hourly samples.

Table 1: ERA5-Land variables for input features.

Input Features (Unit)		
dewpoint_temperature_2m (K)	runoff_hourly (m)	u_component_of_wind_10m (m/s)
temperature_2m (K)	sub_surface_runoff_hourly (m)	v_component_of_wind_10m (m/s)
skin_temperature (K)	surface_runoff_hourly (m)	soil_temperature_level_1 (K)
forecast_albedo	total_evaporation_hourly (m)	soil_temperature_level_2 (K)
surface_pressure (Pa)	total_precipitation_hourly (m)	soil_temperature_level_3 (K)
surface_latent_heat_flux_hourly (J/m^2)	potential_evaporation_hourly (m)	soil_temperature_level_4 (K)

surface_net_solar_radiation_hourly (J/m^2)	surface_sensible_heat_flux_hourly (J/m^2)	surface_net_thermal_radiation_hourly (J/m^2)
surface_solar_radiation_downwards_hourly (J/m^2)	leaf_area_index_high_vegetation (area fraction)	leaf_area_index_low_vegetation (area fraction)
volumetric_soil_water_layer_1 (volume fraction)	volumetric_soil_water_layer_2 (volume fraction)	volumetric_soil_water_layer_3 (volume fraction)
evaporation_from_bare_soil_hourly (m of water equivalent)	evaporation_from_the_top_of_canopy_hourly (m of water equivalent)	evaporation_from_vegetation_transpiration_hourly (m of water equivalent)
evaporation_from_open_water_surfaces_excluding_oceans_hourly (m of water equivalent)		

2.3. Machine Learning Models

In this study, we conduct a comparative analysis of four distinct machine learning regression models: Random Forest Regression (RFR) [17], Extreme Gradient Boosting (XGB) [18], Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP). These models were selected for their proven ability to handle complex, nonlinear relationships and their extensive use in real-world applications across diverse fields. RFR and XGB are ensemble-based methods that utilize multiple decision trees to enhance predictive accuracy. SVR employs kernel-based learning techniques to capture high dimensional patterns while maintaining strong generalization capabilities. MLP is a fundamental type of feedforward artificial neural network that is effective in modelling nonlinear relationships in multivariate datasets. Each machine learning model was optimized through hyperparameter tuning using a Bayesian optimization approach implemented via Optuna [19], combined with five-fold cross-validation to ensure robust and generalizable performance. The number and type of hyperparameters varied across models, reflecting differences in their structural complexity and learning mechanisms. Model performance was evaluated using the coefficient of determination (R^2 score) as shown in Eq. (1), which measures the proportion of variance explained by the model.

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

Where y_i denotes the observed (actual) value, $\hat{y}_i$ represents the predicted value generated by the model, $\bar{y}_i$ is the mean of the observed values, and n is the total number of observations. The model that achieved the highest predictive accuracy was subsequently selected for feature selection analysis to assess the relative importance of input variables.

3. Results and Discussion

3.1. Model Performance Evaluation

To evaluate the predictive performance of the selected machine learning models, we conducted hyperparameter tuning and integrated with five-fold cross-validation method. Each model was optimized through 50 iterative trials to ensure a comprehensive exploration of the hyperparameter space and robust generalization. Table 2 summarizes the optimal hyperparameters identified during this optimization process, along with their corresponding R^2 scores and computational times. The predictive performance ranking of the models based on their R^2 scores is as follows: XGB (0.9882), RFR (0.9792), SVR (0.8464), and MLP (0.7575). In terms of computational efficiency, MLP required the least optimization time, followed by SVR, XGB, and RFR, respectively. Among the evaluated models, the ensemble-based methods, RFR and XGB, consistently outperformed SVR and MLP, demonstrating superior capability in capturing the complex, nonlinear relationships within the dataset. However, these ensemble methods also required longer computational times, with RFR taking the longest, 46 minutes and 49 seconds, to complete the optimization process. It is noteworthy that achieving satisfactory results with SVR necessitated standard scaling of input features, while MLP required standard scaling of both input features and labels. Figure 3 illustrates the variability and distribution of R² scores for each model across the five cross-validation folds. Both XGB and RFR exhibited narrow interquartile ranges and fewer outliers, indicating high consistency and robustness across data splits. In contrast, SVR and MLP demonstrated greater variability, suggesting increased sensitivity to training data partitioning and underscoring the necessity for careful data preprocessing to ensure reliable performance. To maintain visual clarity and comparability in the plot, two extreme outlier trials from SVR and eleven from MLP, each yielding negative R² scores, were excluded, as their inclusion would have distorted the scale due to their significant deviation from the majority of results.

Table 2: Hyperparameters tuning results for machine learning models.

Model	Hyperparameters	Description	Optimization Range	Best Setup	R^2 Score	Optimization Time (s)
RFR	n_estimators	Number of trees in the forest	100 to 3,000	2,307	0.9792	46 m 49 s
	max_depth	Maximum depth of each tree	3 to 20	19		
	max_features	Fraction of features to consider for splitting	0.1 to 1.0	0.8436		
	min_samples_leaf	Minimum samples required at a leaf node	1 to 10	2		
XGB	learning_rate	Learning rate	0.001 to 0.1	0.0319	0.9882	19 min 1 s
	n_estimators	Number of boosting rounds (trees)	100 to 3,000	2,248		
	max_depth	Maximum depth of each tree	3 to 20	16		
	min_child_weight	Minimum sum of instance weight for splitting	1 to 10	2		
	gamma	Minimum loss reduction to split	0 to 5	0.00017		
	subsample	Fraction of samples used per tree	0.5 to 1.0	0.6788		
	colsample_bytree	Fraction of features used per tree	0.5 to 1.0	0.8183		
SVR	C	Penalty for misclassification	0.1 to 10	9.9367	0.8464	9 min 18 s
	epsilon	Margin of tolerance around predicted value	0.01 to 1.0	0.0335		
	gamma	Radial Basis Function (RBF) influence	0.0001 to 0.1	0.0655		
MLP	n_layers	Number of hidden layers	1 to 20	3	0.7575	8 min 37 s
	n_units_l	Number of neurons per layer	10 to 1,000	848, 636, 196		
	activation	Activation function	identity, logistic, tanh, relu	relu		
	learning_rate_init	Initial learning rate	0.0001 to 0.1	0.00146		

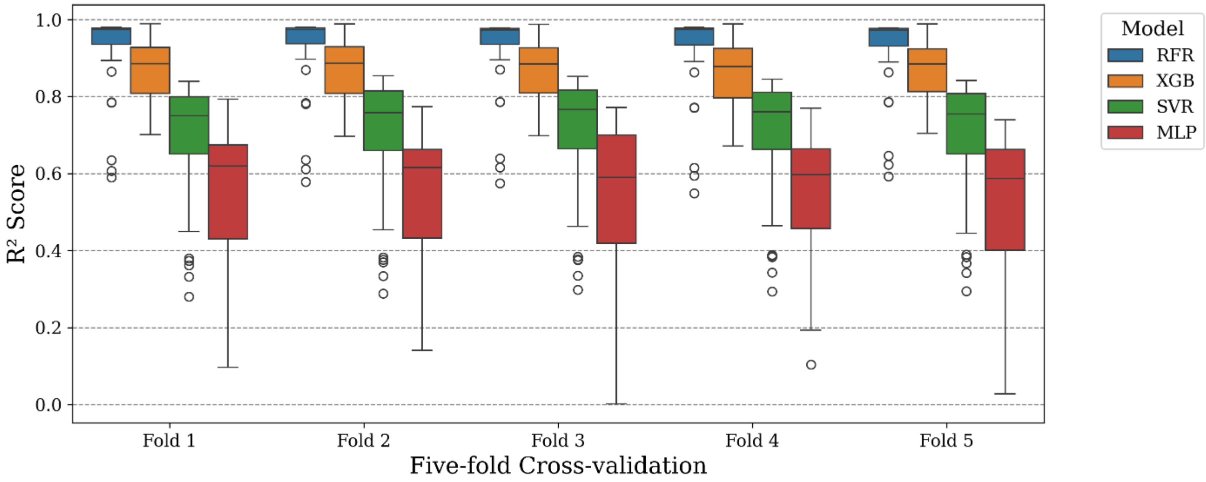

Fig. 3: Hyperparameter tuning results from optimization trials.

3.2. Feature Importance Evaluation

Following the comparative evaluation of model performance, XGB model was selected for the subsequent feature importance analysis, owing to its superior predictive accuracy and relatively efficient computational time. To assess the contribution of individual features, the model was initially trained using all 34 input variables. Feature importance was then quantified using SHapley Additive exPlanations (SHAP) [20], a robust, model-agnostic interpretability framework that provides locally accurate explanations of feature impact on the model's predictions. Figure 4 displays the mean absolute SHAP values for each of the 34 features, revealing that variables associated with runoff and soil moisture, consistent with established hydrological processes, dominate the upper ranks. In contrast, features related to wind components, solar radiation, temperature, and evaporation rank lower, suggesting a comparatively weaker influence on the model's predictive capability. Notably, precipitation exhibits a surprisingly low impact on the model's performance.

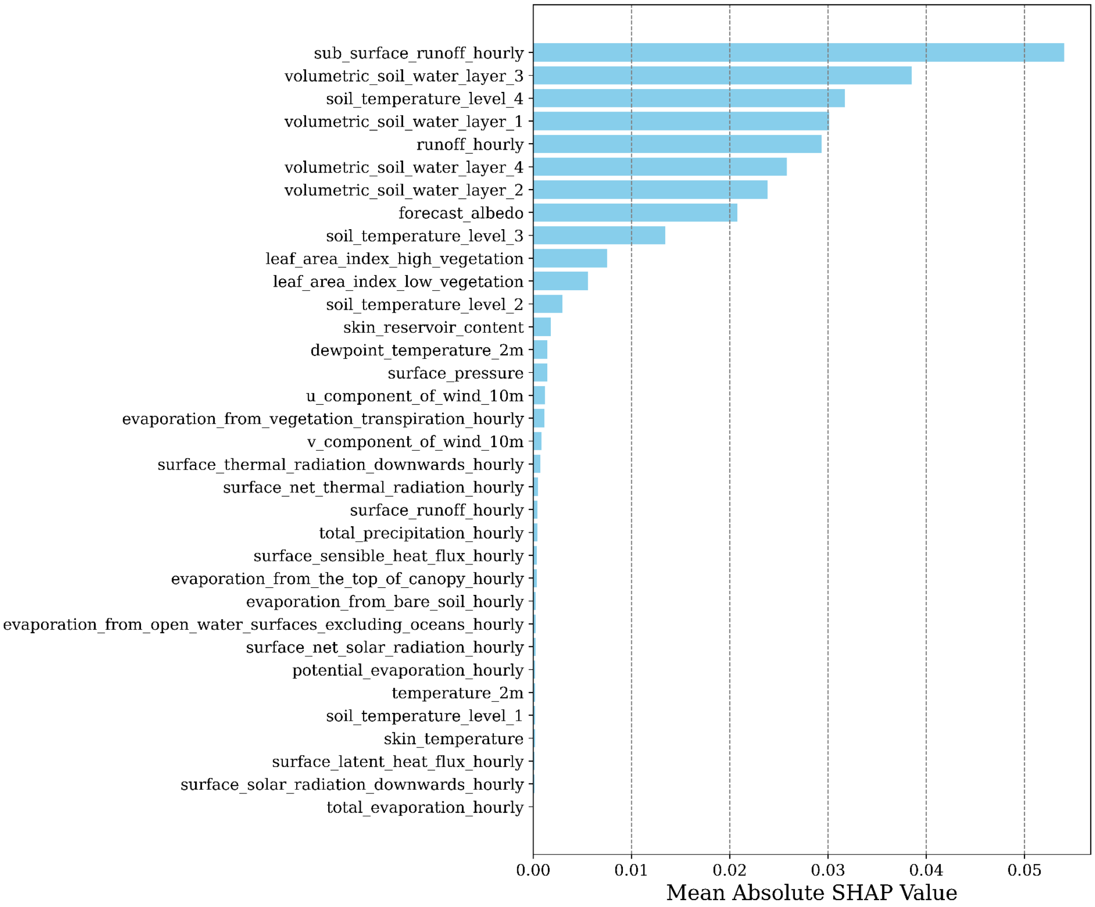

Fig. 4: Feature importance ranked by mean absolute SHAP value.

To evaluate feature importance, we conducted experiments under four distinct scenarios. In Scenario 1, the model was trained using all 34 input features. In Scenario 2, the top four features—those with the highest SHAP values—were removed. Scenario 3 involved excluding the top 14 features, while Scenario 4 retained only the 10 lowest-ranking features after removing the top 24. Figure 5 presents a comparative analysis of the actual river stage height versus the predicted height for each scenario, alongside their corresponding R² scores. Scenario 1 achieved the highest R² score of 0.9882, followed closely by Scenario 2 with 0.9778. In contrast, Scenario 3 and Scenario 4 obtained R² scores of 0.4838 and 0.2669, respectively. These results clearly demonstrate that the top 14 features are critical to the model's accuracy, as their exclusion leads to a substantial deterioration in predictive performance.

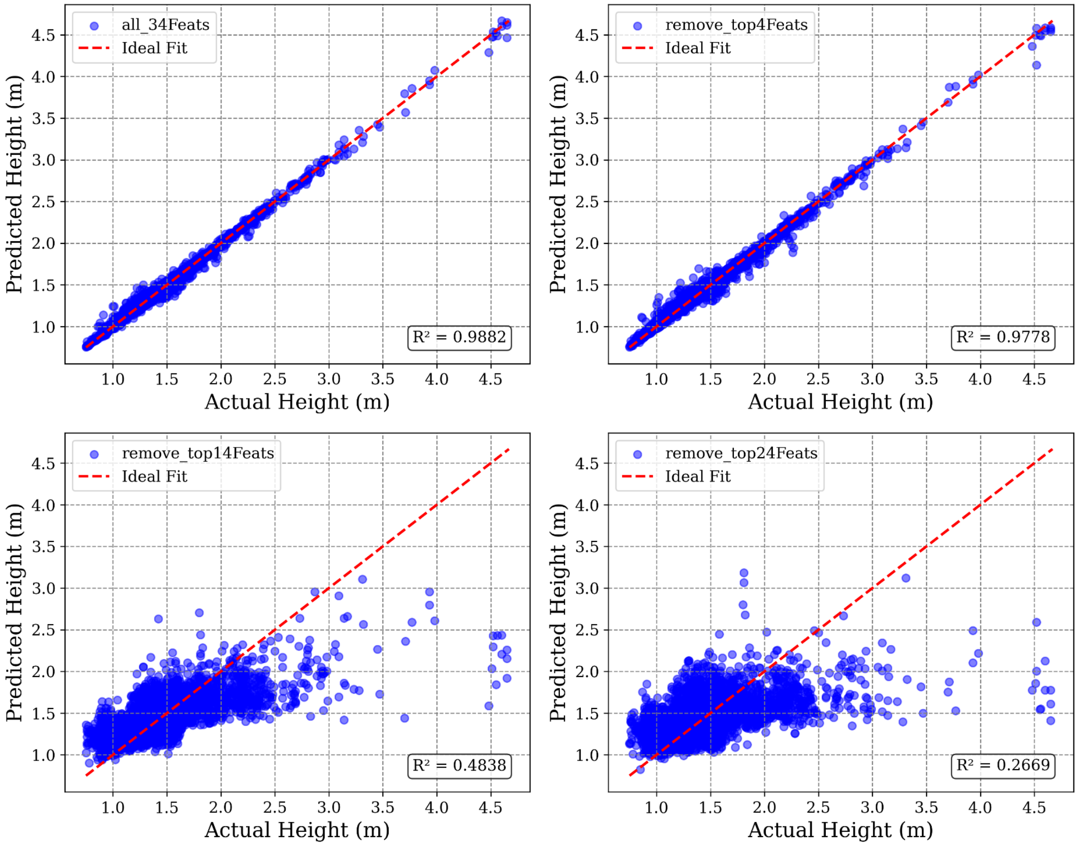

Fig. 5: Comparison of predicted and actual river stage heights under feature exclusion scenarios.

4. Conclusion

This study demonstrates the effectiveness of GIS-enhanced data and machine learning techniques in predicting river stage in ungauged basins, with a specific focus on the Upper Ping River Basin in northern Thailand. By leveraging five years of hourly hydrometeorological data from the ERA5-Land reanalysis product, along with observed river stage measurements from a hydrological station, we evaluated the performance of four machine learning models: Random Forest Regression (RFR), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP). The models were optimized using Bayesian hyperparameter tuning and validated through five-fold cross-validation. Among the models tested, XGB consistently delivered the highest predictive accuracy, with an R^2 score of 0.9882, followed closely by RFR at 0.9792. SVR and MLP exhibited lower performance and greater sensitivity to data partitioning and outlier, with R^2 scores of 0.8464 and 0.7575, respectively. These findings demonstrate the superior capability of ensemble tree-based methods in capturing the complex, nonlinear interactions inherent in hydrological systems, albeit with increased computational demands.

To further interpret model behaviour, SHAP analysis was employed to assess the contribution of each input feature. The results revealed that variables related to runoff and soil moisture were the most influential, while wind components, temperature, and solar radiation were comparatively less impactful. Interestingly, precipitation showed a limited effect on model performance, contrary to conventional hydrological assumptions. A feature exclusion experiment was conducted to evaluate model robustness under varying input conditions. As expected, the removal of top-ranking features resulted in a significant decline in prediction accuracy. While the full-feature model achieved an R^2 of 0.9882, the models excluding the top 4, 14, and 24 features yielded R² scores of 0.9778, 0.4838, and 0.2669, respectively, highlighting the critical role of the top 14 features in maintaining model reliability.

In conclusion, this study demonstrates that integrating high-resolution GIS reanalysis data with interpretable machine learning approaches provides a powerful framework for river stage prediction, particularly in data-scarce regions. The findings support the application of such methods in real-time flood forecasting and water resource planning in ungauged or poorly instrumented basins. Future work may explore a variety of catchment areas to evaluate geological and spatial effects, as well as the integration of additional GIS-derived features and advanced deep learning architectures to further enhance prediction accuracy and spatial generalizability.

Acknowledgements

This research was supported by the Erawan High Performance Computing (HPC) Project, Information Technology Service Center (ITSC), Chiang Mai University, Chiang Mai, Thailand. The authors gratefully acknowledge the Royal Irrigation Department, Ministry of Agriculture and Cooperatives, for providing the river stage data used in this study.

References

[1] T. Davie and N. W. Quinn, *Fundamentals of hydrology*, 3rd ed. Abingdon, Oxon, UK: Routledge, 2019.

[2] Y. Zhao, L. Jiang, X. Zhang, and J. Liu, "Tracking River's Pulse From Space: A Global Analysis of River Stage Fluctuations," *Geophysical Research Letters,* vol. 50, no. 23, 2023.

[3] C. Sinsabvarodom, W. Chai, B. J. Leira, K. V. Høyland, and A. Naess, "Uncertainty assessments of structural loading due to first year ice based on the ISO standard by using Monte-Carlo simulation," *Ocean Engineering,* vol. 198, 2020.

[4] C. Sinsabvarodom, B. J. Leira, K. V. Høyland, A. Næss, I. Samardžija, W. Chai, S. Komonjinda, C. Chaichana, and S. Xu, "On Statistical Features of Ice Loads on Fixed and Floating Offshore Structures," *Journal of Marine Science and Engineering,* vol. 12, no. 8, 2024.

[5] C. Sinsabvarodom, A. Næss, B. J. Leira, and W. Chai, "Extreme Value Estimation of Beaufort Sea Ice Dynamics Driven by Global Wind Effects," *China Ocean Engineering,* vol. 36, no. 4, pp. 532-541, 2022.

[6] C. Hu, Q. Wu, H. Li, S. Jian, N. Li, and Z. Lou, "Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation," *Water,* vol. 10, no. 11, 2018.

[7] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks," *Hydrology and Earth System Sciences,* vol. 22, no. 11, pp. 6005-6022, 2018.

[8] M. Zounemat-Kermani, A. Mahdavi-Meymand, M. Alizamir, S. Adarsh, and Z. M. Yaseen, "On the complexities of sediment load modeling using integrative machine learning: Application of the great river of Loíza in Puerto Rico," *Journal of Hydrology,* vol. 585, 2020.

[9] Y. Ding, Y. Zhu, J. Feng, P. Zhang, and Z. Cheng, "Interpretable spatio-temporal attention LSTM model for flood forecasting," *Neurocomputing,* vol. 403, pp. 348-359, 2020.

[10] P. Panyadee and P. Champrasert, "Spatiotemporal Flood Hazard Map Prediction Using Machine Learning for a Flood Early Warning Case Study: Chiang Mai Province, Thailand," *Sustainability,* vol. 16, no. 11, 2024.

[11] K. Ullah, Y. Wang, Z. Fang, L. Wang, and M. Rahman, "Multi-hazard susceptibility mapping based on Convolutional Neural Networks," *Geoscience Frontiers,* vol. 13, no. 5, 2022.

[12] F. Boodoo, R. Hostache, N. Skifa, J. Guerin, and C. Delenne, "Are LSTM and conceptual rainfall-runoff models able to cope with limited training datasets under diverse hydrometeorological conditions?," *Modeling Earth Systems and Environment,* vol. 11, no. 2, 2025.

[13] A. Muhebwa, C. J. Gleason, D. Feng, and J. Taneja, "Improving Discharge Predictions in Ungauged Basins: Harnessing the Power of Disaggregated Data Modeling and Machine Learning," *Water Resources Research,* vol. 60, no. 9, 2024.

[14] Upper Northern Region Irrigation Hydrology Center. Hydrological Data Services – Upper Northern Thailand [Online] Available: https://hydro-1.net/

[15] T. Chaipimonplin, "Investigation internal parameters of neural network model for Flood Forecasting at Upper river Ping, Chiang Mai," *KSCE Journal of Civil Engineering,* vol. 20, no. 1, pp. 478-484, 2016.

[16] J. M. Sabater. *ERA5-Land hourly data from 1950 to present*, Copernicus Climate Change Service (C3S) Climate Data Store (CDS). [Online]. Available: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download

[17] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001/10/01 2001.

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," presented at the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.

[20] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 4765-4774.