SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Multi-Model River Water Level Forecasting Framework Under Low, Medium and High-Water Level Regimes

Raunak Kumar[1], Abhishek Kumar[2*], Premlata Vidyarthy[3], Pankaj Kumar[4], Deepak Kumar[5], Bhavesh Chandra[6]

[1]Department of Civil Engineering, Government Engineering College West Champaran, Bihar India. E-mail: raunaks1991@gmail.com, Orcid Id: 0009-0008-2438-1385
[2*]Department of Civil Engineering, Government Engineering College West Champaran, Bihar India. E-mail: getaktiwari@gmail.com, Orcid Id: 0000-0003-3527-790X
[3]Department of Civil Engineering, Government Engineering College Sheikhpura, Bihar India. E-mail: rajkiprem@gmail.com, Orcid Id: 0009-0004-8813-6689
[4]Department of Civil Engineering, Government Engineering College Sheohar, Bihar India. E-mail: pankaj.kumar19979@gmail.com, Orcid Id: 0009-0006-9146-9091
[5]Department of Civil Engineering, Government Engineering College Sheohar, Bihar India. E-mail: deepakkumar@gecsheohar.ac.in Orcid Id: 0009-0007-2256-3145
[6]Department of Civil Engineering, Government Engineering College Gopalganj, Bihar India. E-mail: bhaveshc.ph23.ce@nitp.ac.in, Orcid Id: 0009-0007-5817-9348

ABSTRACT

This study develops and evaluates a multi-step forecasting framework for the Punpun River using Artificial Neural Network (ANN), Bidirectional Long Short-Term Memory (BiLSTM), XGBoost, and a Dynamic Hybrid model. Water-level forecasts were generated at 1-, 3-, 5-, 7-, and 10-day lead times using historical hydrological time-series data. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The forecasts were further examined under low-, medium-, and high-water-level regimes to assess model behaviour across different hydrological conditions. The results showed a progressive reduction in forecasting skill with increasing lead time. Overall R^2 values ranged from 0.95–0.97 at the 1-day lead time and declined to 0.69–0.70 at the 10-day lead time. XGBoost produced the highest R^2 values at the 1-, 3-, 5-, and 7-day lead times, whereas ANN produced the highest R^2 at 10 days. XGBoost also yielded relatively lower RMSE at short and intermediate lead times, while the Dynamic Hybrid model showed comparatively lower MAE from 3 to 10 days. Regime-wise assessment indicated that the Dynamic Hybrid model maintained positive R^2 values with comparatively lower errors under low-water-level conditions across the evaluated lead times. XGBoost showed relatively better representation of high-water-level conditions at short and intermediate horizons. Overall, incorporating multi-step forecasting with regime-wise evaluation provides a more comprehensive assessment of model behaviour and offers a useful basis for developing river water-level forecasting systems for flood early warning and water-resources management.

Keywords: Punpun River, water-level forecasting, multi-step forecasting, ANN, BiLSTM, XGBoost, Dynamic Hybrid model, hydrological regimes.

Introduction

Accurate prediction of river water levels is essential for flood early-warning systems, reservoir management, and sustainable water-resources planning, as timely water-level information supports the assessment and management of changing hydrological conditions (Hipni et al., 2013; Sapitang et al., 2020). Increasing hydrological variability and the occurrence of extreme events have further increased the need for reliable forecasting systems capable of providing predictions over different lead times (Chaubey et al., 2021). Hydrological forecasting approaches are broadly represented by physically based and data-driven methods; although physically based models provide a process-oriented representation of hydrological behaviour, their application generally requires extensive catchment information, parameter calibration, and computational effort (Ranjan & Roshni, 2023; Tripura et al., 2021). In contrast, machine-learning approaches have gained increasing attention in water-level and streamflow forecasting because they can represent complex nonlinear relationships directly from observed data without requiring explicit formulation of all underlying physical processes (Ahmed et al., 2022; Zakaria et al., 2021). Previous studies have demonstrated the applicability of neural-network and machine-learning techniques for river and reservoir water-level prediction, while also showing that model performance can vary according to the forecasting horizon and the characteristics of the available hydrological data (Ahmed et al., 2022; Zakaria et al., 2021; Shortridge et al., 2016). Therefore, evaluating multiple forecasting models across short and extended lead times is important for understanding their predictive behaviour under different hydrological conditions.

Among data-driven approaches, Artificial Neural Networks (ANNs) have been widely applied in water-resources modelling because of their ability to represent nonlinear relationships between hydrological inputs and outputs (Maier & Dandy, 2000). Long Short-Term Memory (LSTM) networks have subsequently received considerable attention in hydrological time-series forecasting because their memory-cell and gating mechanisms allow temporal information to be retained over extended sequences (Hochreiter & Schmidhuber, 1997; Kratzert et al., 2018). The applicability of LSTM-based models has been demonstrated for rainfall–runoff simulation and streamflow forecasting, including applications involving multi-day prediction horizons (Kratzert et al., 2018; Hunt et al., 2022; Sabzipour et al., 2023). However, the predictive performance of LSTM models may decrease during rapidly changing or extreme hydrological conditions, particularly as the forecasting horizon increases (Cho & Kim, 2022; Lees et al., 2021). In parallel, tree-based ensemble methods such as Random Forest and XGBoost have demonstrated strong capability in representing nonlinear relationships and complex interactions within hydrological datasets (Breiman, 2001; Chen & Guestrin, 2016). XGBoost has also been applied successfully to water-level and streamflow forecasting, while its computational and regularization characteristics make it suitable for nonlinear regression problems (Chen & Guestrin, 2016; Zakaria et al., 2021). Nevertheless, tree-based models do not possess the intrinsic sequential-memory structure of recurrent neural networks, creating scope for combining complementary modelling characteristics within a unified forecasting framework (Konapala et al., 2020; Yang et al., 2020).

The complementary characteristics of neural networks and ensemble tree-based methods have motivated the development of hybrid hydrological forecasting frameworks that combine temporal sequence learning with nonlinear error correction or complementary decision processes (Konapala et al., 2020; Yang et al., 2020). Previous studies have shown that hybrid modelling can improve the representation of complex hydrological behaviour by integrating different learning mechanisms within a single forecasting framework (Cho & Kim, 2022; Konapala et al., 2020). At the same time, comparisons of machine-learning models for water-level forecasting have indicated that the relative performance of different algorithms can change with forecasting horizon and dataset characteristics, demonstrating that no single modelling architecture necessarily provides identical performance under all prediction conditions (Zakaria et al., 2021). This indicates a need to evaluate forecasting models not only using aggregate performance measures but also under different water-level conditions and across multiple lead times. In this context, the present study develops a Dynamic Hybrid framework that combines ANN, BiLSTM, and XGBoost models within a flow-regime-based forecasting strategy and evaluates their performance for 1-, 3-, 5-, 7-, and 10-day water-level prediction in the Punpun River. The analysis further distinguishes low-, medium-, and high-water-level regimes to examine how predictive performance varies under different hydrological conditions, providing a more detailed assessment of multi-step forecasting behaviour than a single overall performance measure.

The Punpun River Basin provides an important setting for evaluating multi-step water-level forecasting because of its pronounced seasonal hydrological variability. The river originates in the highlands of Palamu district in Jharkhand and ultimately joins the Ganga River at Fatuha, near Patna, while its flow regime is characterized by very low or negligible flow during the non-monsoon period and substantially increased flow during the monsoon season (Prakash & Tripura, 2026). A large proportion of the annual rainfall occurs during the monsoon season, producing substantial temporal variation in river conditions and increasing the importance of timely water-level forecasting for flood-related applications (Prakash & Tripura, 2026). Previous investigations in the basin have demonstrated the applicability of machine-learning and hybrid approaches for water-level prediction, but reported performance varies with the selected model and forecasting horizon (Prakash & Tripura, 2026). More broadly, previous studies have also shown that forecasting accuracy commonly decreases as the prediction horizon increases, emphasizing the need to evaluate models systematically across multiple lead times rather than relying only on short-term forecasts (Kao et al., 2020; Zakaria et al., 2021).

Despite the progress in machine-learning-based hydrological forecasting, further investigation is required into how different model architectures behave across extended forecasting horizons and contrasting water-level regimes. Previous studies have generally emphasized overall predictive accuracy, while comparatively less attention has been given to evaluating model behaviour separately under low-, medium-, and high-water-level conditions (Zhu et al., 2020; Zounemat-Kermani et al., 2020). In addition, comparative studies have demonstrated that different algorithms may exhibit varying strengths depending on the forecasting horizon and characteristics of the hydrological dataset, making direct multi-model assessment necessary (Zakaria et al., 2021). Therefore, the present study develops a Dynamic Hybrid forecasting framework incorporating ANN, BiLSTM, and XGBoost and evaluates it against the corresponding standalone models for 1-, 3-, 5-, 7-, and 10-day water-level forecasting in the Punpun River. The study further classifies the observed water levels into low-, medium-, and high-level regimes and evaluates model performance using R^2, RMSE, and MAE to examine how predictive behaviour changes with both forecasting horizon and hydrological condition. The analysis is intended to provide a comprehensive assessment of multi-step water-level forecasting performance and to identify model-specific characteristics that may not be evident from aggregated performance statistics alone.

STUDY AREA

Figure 1 presents the location and extent of the Punpun River Basin, which forms the study area of this research. The Punpun River originates in the upland region of Palamu district, Jharkhand, at an elevation of approximately 300 m above mean sea level, near 24°11′ N latitude and 84°09′ E longitude. From its source, the river flows towards the Gangetic plain and ultimately drains into the Ganga River near Fatuha, approximately 25 km downstream of Patna. The river exhibits strong seasonal variability in flow. During the non-monsoon period, streamflow is generally very low and several reaches may remain dry. In contrast, monsoon rainfall generates substantial flows, which can cause flooding and inundation in the adjoining low-lying areas. This pronounced seasonal contrast between high monsoon discharge and limited dry-season flow presents important challenges for water-resource management within the basin.

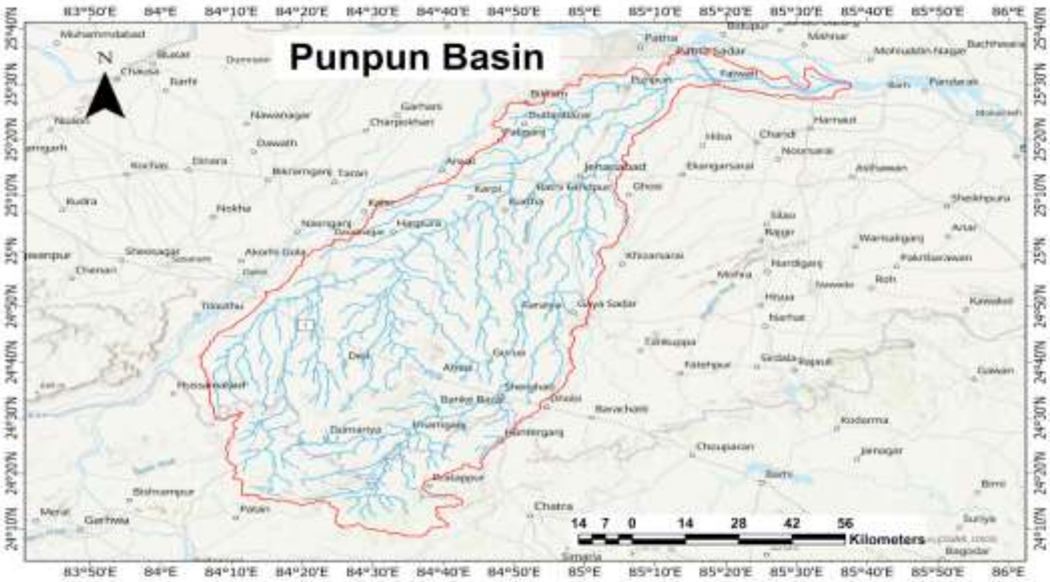


Figure 1. Representation of study area used for study

Rainfall is concentrated mainly during the monsoon period, with substantial precipitation occurring from August to October. Mean annual rainfall across the basin ranges from approximately 99 to 134 cm, with the upper catchment, particularly the Palamu region, generally receiving higher rainfall than the downstream areas. Around 80–87% of the annual precipitation occurs during the monsoon season. In the lower part of the basin, rainfall is comparatively uniform spatially, resulting in relatively small variations across the region. The Punpun River Basin covers an area of approximately 7055 km². Its geological setting is diverse, reflecting the transition from the upland terrain to the Gangetic plains

Methodology

The methodology developed in this study comprises a dynamic flow-regime-based framework for multi-step river water-level forecasting as shown in Figure 2. The approach integrates three standalone data-driven models, namely Artificial Neural Network (ANN), Bidirectional Long Short-Term Memory (BiLSTM), and Extreme Gradient Boosting (XGBoost), with a dynamic hybrid framework that accounts for different water-level regimes. The modelling procedure involves preparation of the time-series data, formulation of predictor variables and future water-level targets, development of the forecasting models, and evaluation of their predictive performance. Water-level forecasts were generated for 1-, 3-, 5-, 7-, and 10-day lead times to examine the variation in predictive capability with increasing forecasting horizon. The developed framework was further assessed under low-, medium-, and high-water-level conditions to examine model behaviour across different hydrological regimes.

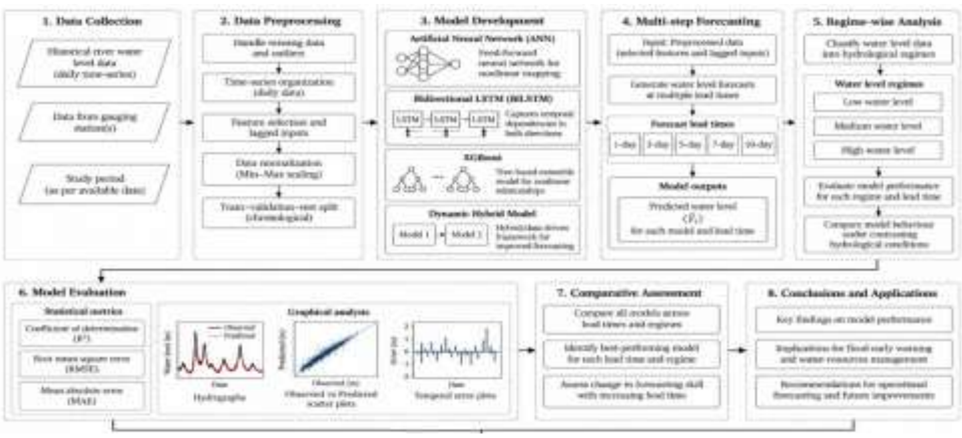


Figure 2. Methodological framework adopted for study

DATA PREPARATION AND TARGET FORMULATION

The observed hydrological time series were first arranged chronologically according to the observation date to preserve the temporal structure of the dataset. The date variable was converted into a standard datetime format, and the day and month components were extracted as temporal predictors. Missing values in the runoff series were treated using linear interpolation, while observations remaining incomplete after the construction of lagged variables were removed from the modelling dataset. The resulting data were divided chronologically into training and testing subsets using an 80:20 ratio, with the earlier observations used for model development and the later observations reserved for independent evaluation. Prior to model training, the predictor variables were normalized using Min–Max scaling.

To formulate the multi-step forecasting problem, the observed water-level series was shifted according to the selected forecasting horizon. Thus, the target variable represented the water level occurring Ldays after the predictor observation, where Lcorresponds to 1, 3, 5, 7, or 10 days. The forecasting relationship can therefore be expressed as $\widehat{WL}_{t+L} = f(X_t)$ where $\widehat{WL}_{t+L}$is the predicted water level at lead time L, $X_t$represents the predictor vector available at time t, and $f(\cdot)$denotes the forecasting model. This procedure was repeated independently for each lead time to examine the effect of increasing forecasting horizon on model performance.

The predictor set was constructed to represent both the prevailing and antecedent hydrological conditions influencing future water-level fluctuations. The inputs consisted of current rainfall, current runoff, day of the month, month of the year, and lagged observations of rainfall, runoff, and water level for the preceding 10 days. In total, 34 predictor variables were used for model development. The inclusion of lagged water-level, rainfall, and runoff variables was intended to represent the temporal persistence and delayed hydrological response of the Punpun River system.

Model Development

Three standalone data-driven models, namely Artificial Neural Network (ANN), Bidirectional Long Short-Term Memory (BiLSTM), and Extreme Gradient Boosting (XGBoost), were developed for river water-level prediction. In addition, a dynamic hybrid framework was formulated by incorporating water-level regime classification into the prediction process. The three standalone models provide independent benchmark predictions, whereas the dynamic framework selects a regime-specific prediction model according to the predicted future water-level condition. Such comparative and hybrid modelling approaches are increasingly used in hydrological forecasting because different learning architectures can represent complementary aspects of nonlinear and temporal hydrological behaviour (Maier and Dandy, 2000; Kratzert et al., 2018; Chen and Guestrin, 2016; Konapala et al., 2020).

Artificial Neural Network

The Artificial Neural Network (ANN) was employed as a feed-forward nonlinear regression model for water-level prediction. Neural-network-based methods have been widely applied to water-resources forecasting because of their ability to establish nonlinear relationships between predictor variables and hydrological responses (Maier and Dandy, 2000). The ANN used in the present study consisted of two hidden layers with 64 and 32 neurons, respectively. The rectified linear unit (ReLU) activation function and Adam optimization algorithm were used for model training. Within the dynamic framework, ANN was assigned to the low-water-level regime, while a separate ANN was also developed as a standalone benchmark model.

Bidirectional Long Short-Term Memory

The Bidirectional Long Short-Term Memory (BiLSTM) network was employed to represent temporal dependencies in the water-level time series. LSTM was originally introduced by Hochreiter and Schmidhuber (1997) to overcome the limitations of conventional recurrent neural networks in learning long-term dependencies, and its application has subsequently been demonstrated extensively in hydrological forecasting (Kratzert et al., 2018). In the present study, the BiLSTM architecture comprised two bidirectional LSTM layers containing 64 and 32 neurons, respectively, with dropout layers of 0.2 used between the recurrent layers. A single dense neuron provided the final water-level output. The network was trained using the Adam optimizer and mean squared error loss, with early stopping employed to reduce overfitting. BiLSTM was used both as a standalone benchmark and as the regime-specific predictor for the medium-water-level condition.

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) was used as a tree-based nonlinear regression model. XGBoost is an efficient implementation of gradient-boosted decision trees that incorporates regularization and ensemble learning to improve predictive performance and control model complexity (Chen and Guestrin, 2016). In this study, XGBoost was configured with 500 trees, a maximum tree depth of 6, a learning rate of 0.03, a subsample ratio of 0.8, and a column-sampling ratio of 0.8. The model was developed both as a standalone benchmark and as the expert predictor associated with the high-water-level regime.

Dynamic Hybrid Framework

The proposed dynamic hybrid framework was developed to account for variations in prediction behaviour across different water-level conditions. The future target water level was classified into three regimes using percentile-based thresholds: low, medium, and high. The lower regime corresponds to values below the 33rd percentile, the medium regime to values between the 33rd and 66th percentiles, and the high regime to values above the 66th percentile. A Random Forest classifier was used to identify the expected future regime from the predictor variables. Random Forest is an ensemble learning approach based on multiple decision trees and has been widely used for nonlinear hydrological prediction and classification (Breiman, 2001).

Following regime identification, the corresponding expert model was activated: ANN for low water levels, BiLSTM for medium water levels, and XGBoost for high water levels. The prediction generated by the selected expert model was taken as the final Dynamic Hybrid prediction. Thus, the framework combines a classification stage with regime-specific regression models, allowing different model structures to be applied according to the expected hydrological condition. This type of hybrid formulation follows the broader principle that combining complementary learning approaches can improve representation of complex hydrological behaviour (Konapala et al., 2020).

Multi-Step Forecasting and Regime-Wise Comparison

The developed models were evaluated for five forecasting horizons of 1, 3, 5, 7, and 10 days. For each lead time, the target water level was shifted by the corresponding number of days, and the models were trained and tested independently using the same predictor formulation. This multi-lead-time strategy enables the variation in predictive capability and error magnitude with increasing forecasting horizon to be assessed systematically, as adopted in previous water-level and streamflow forecasting studies (Prakash and Tripura, 2026; Zakaria et al., 2023).

In addition to the overall comparison, the forecasting performance was evaluated under different water-level regimes to examine whether model behaviour varied with hydrological conditions. The observed target water levels in the testing period were classified into three regimes, namely low, medium, and high flow, using the 33rd and 66th percentile thresholds. The same regime classification was applied to the predictions of ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model so that their performances could be compared under equivalent hydrological conditions. Thus, the regime-wise analysis was considered as an additional assessment of model behaviour rather than a separate forecasting experiment. The use of percentile-based flow-regime classification provides a simple way of distinguishing contrasting hydrological conditions and is consistent with the regime-based framework described for the present study.

For each lead time and flow regime, the predicted water levels were compared with the corresponding observed values using the selected performance indicators. This enabled the models to be assessed not only for their overall predictive capability but also for their ability to reproduce low-, medium-, and high-water-level conditions as the forecasting horizon increased. Such stratified evaluation is particularly useful for identifying differences in model behaviour that may not be apparent from aggregate performance measures alone.

Performance Evaluation

The predictive performance of the developed models was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), which are commonly used to evaluate hydrological forecasting models (Moriasi et al., 2007; Prakash and Tripura, 2026). The evaluation was performed primarily on the independent testing period to assess the ability of the models to generalize to unseen observations. The same performance measures were calculated for each forecasting lead time and for the low-, medium-, and high-water-level regimes, allowing both overall and regime-specific model behaviour to be examined.

The coefficient of determination was used to quantify the agreement between observed and predicted water levels and is expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (WL_{obs,i} - WL_{pred,i})^2}{\sum_{i=1}^{N} (WL_{obs,i} - \bar{WL}_{obs})^2} \qquad (1)$$

where $WL_{obs,i}$ and $WL_{pred,i}$ denote the observed and predicted water levels, respectively, $\bar{WL}_{obs}$ is the mean observed water level, and N is the number of observations. Higher R^2 values indicate greater agreement between the predicted and observed series.

RMSE was used to represent the magnitude of prediction errors while giving greater weight to larger deviations:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (WL_{pred,i} - WL_{obs,i})^2} \qquad (2)$$

MAE was used to quantify the mean absolute deviation between predicted and observed water levels:

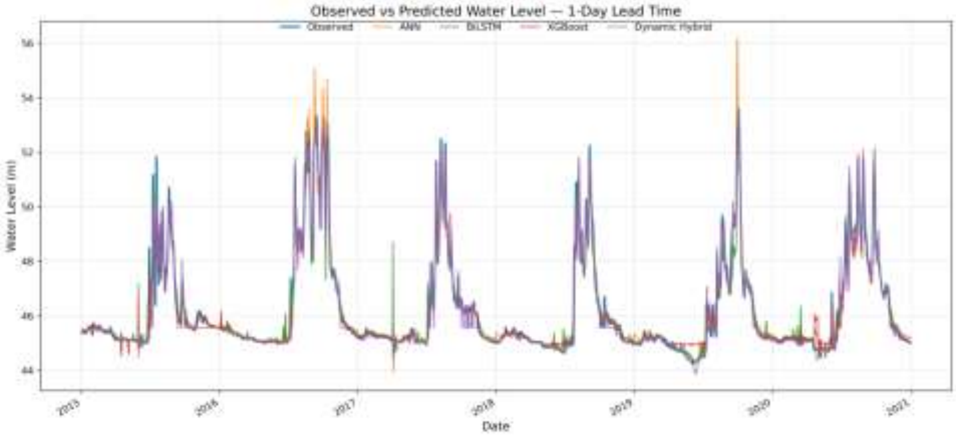
$$MAE = \frac{1}{N}\sum_{i=1}^{N} |WL_{pred,i} - WL_{obs,i}| \qquad (3)$$

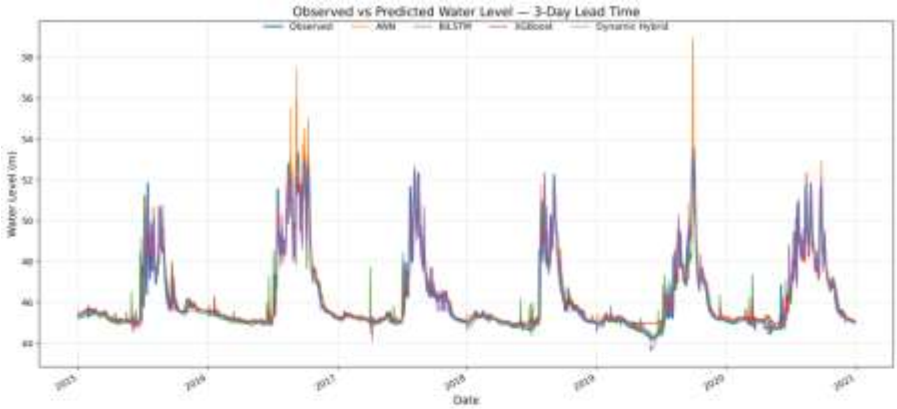
Lower RMSE and MAE values indicate smaller prediction errors. The combination of R^2, RMSE, and MAE provides complementary information on model agreement and error magnitude and has been used in previous multi-lead-time hydrological forecasting studies.

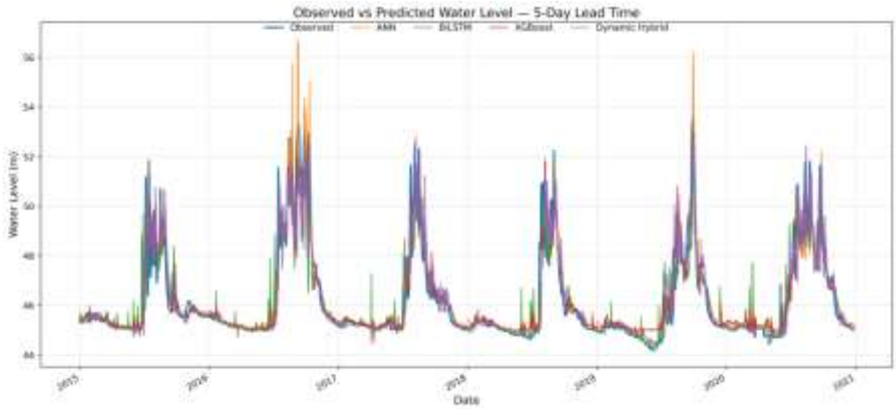
For graphical assessment, observed and predicted water-level series were further examined using hydrographs, observed-versus-predicted scatter plots, and temporal error plots. The results were analyzed across the 1-, 3-, 5-, 7-, and 10-day forecasting horizons to determine the change in predictive performance with increasing lead time. In addition, the regime-wise evaluation was used to examine how the ANN, BiLSTM, XGBoost, and Dynamic Hybrid approaches represented low-, medium-, and high-water-level conditions. This combination of statistical and graphical assessment provides a comprehensive evaluation of both general forecasting capability and behaviour under contrasting hydrological conditions.

Results And Discussion

The developed forecasting framework was evaluated for 1-, 3-, 5-, 7-, and 10-day lead times using the independent testing dataset. The performance of ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model was examined using R^2, RMSE, and MAE, together with graphical comparisons of observed and predicted water levels. In addition to the overall assessment, the predictions were further examined under low-, medium-, and high-water-level conditions to identify differences in model behaviour across contrasting hydrological regimes. Such multi-lead-time comparison provides a systematic basis for examining the effect of increasing forecast horizon on predictive performance, as demonstrated in previous hydrological forecasting studies.







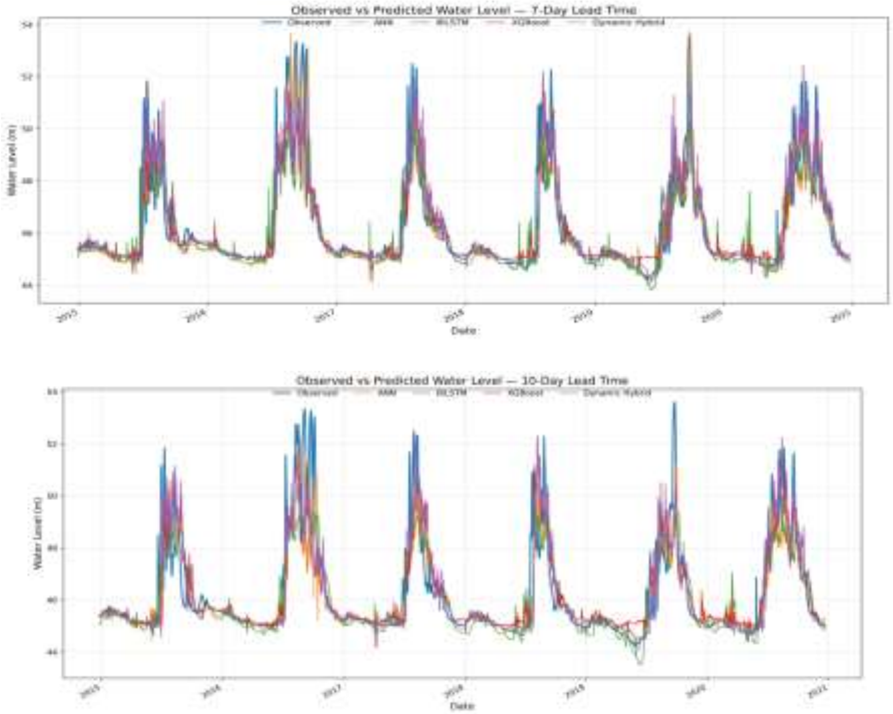
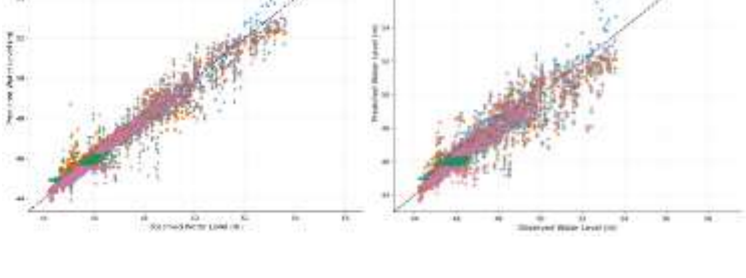


Figure 3: Observed and predicted water-level time series for ANN, BiLSTM, XGBoost, and Dynamic Hybrid models at multiple lead times

Figure 3 presents the observed and predicted water-level series obtained from ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model for the 1-, 3-, 5-, 7-, and 10-day lead times, respectively. At the 1-day forecasting horizon, all four models closely reproduced the overall temporal pattern of the observed water-level series, particularly during the relatively stable low-water periods and the rising and recession limbs of the major events. Although the models generally followed the observed hydrograph, some differences were evident during rapid increases and peak water-level events, where individual models showed occasional over- or underestimation. With increasing lead time, the divergence between observed and predicted series became more noticeable, particularly during high-magnitude events, while the broader seasonal and inter-event patterns remained reasonably well represented. The 3-day forecasts continued to reproduce the principal fluctuations, although deviations around several peaks became more pronounced. At the 5-day horizon, the spread among the model predictions increased, with greater differences in the magnitude and timing of the higher water-level events. The 7- and 10-day forecasts showed the largest deviations from the observed hydrograph, especially during sharp peak events, indicating increasing difficulty in reproducing rapid hydrological changes at extended forecasting horizons. Nevertheless, all models retained the general temporal structure of the water-level series, demonstrating that the developed forecasting approaches were capable of representing the major variations in river water level over the testing period. The increasing discrepancy between observed and predicted peaks with lead time is consistent with the progressively increasing uncertainty associated with multi-step hydrological forecasting.

Figure 4 shows the scatter relationships between observed and predicted water levels for ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model at 1-, 3-, 5-, 7-, and 10-day lead times, respectively. At the 1-day horizon, the prediction points from all four models are concentrated relatively close to the 1:1 reference line, indicating good agreement between observed and predicted water levels, with the main scatter occurring around the higher water-level observations. The ANN predictions show several noticeable departures from the reference relationship during the highest observed events, whereas BiLSTM, XGBoost, and Dynamic Hybrid generally remain more concentrated around the central trend.



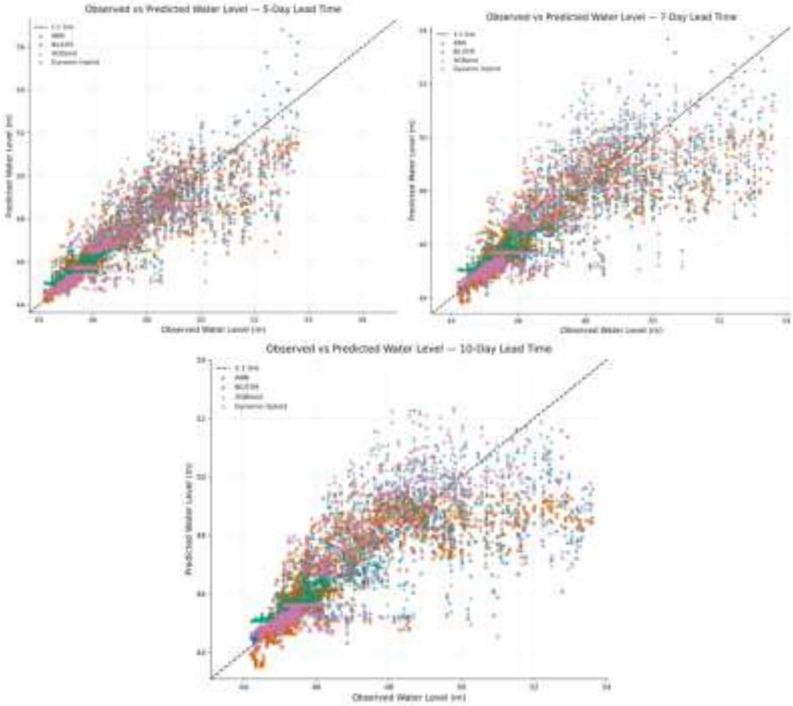
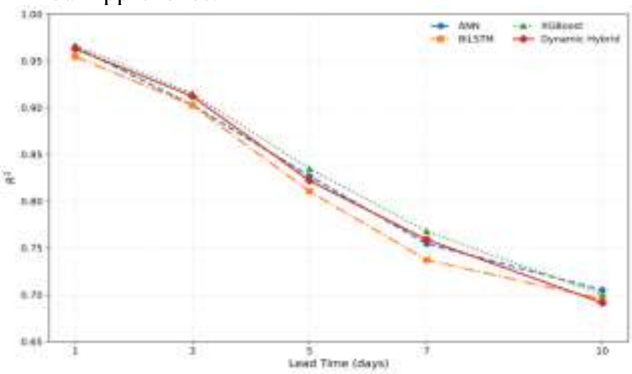


Figure 4: Scatter plots of observed versus predicted water levels for ANN, BiLSTM, XGBoost, and Dynamic Hybrid models at multiple lead times

At the 3-day horizon, the overall linear association remains evident, but the spread of the prediction points increases, particularly for observations above approximately 48–50 m, indicating a gradual reduction in predictive precision as the lead time increases. This dispersion becomes more pronounced at the 5-day horizon, where all four models show a wider spread around the 1:1 line and increasing differences in the representation of higher water levels. At 7 days, the scatter becomes substantially broader, with a noticeable tendency of the predictions from several models to cluster within a narrower range than the observed high water levels, indicating difficulty in reproducing the full magnitude of rapid water-level fluctuations. The 10-day results show the greatest dispersion, particularly for the higher observed water levels, although the general positive relationship between observed and predicted values is still retained. The Dynamic Hybrid predictions remain concentrated around the dominant data structure across the lead times, while differences among the individual models become increasingly evident as the forecasting horizon extends. Overall, the scatter analysis indicates that the models reproduce the general relationship between observed and predicted water levels effectively at shorter lead times, whereas prediction dispersion increases progressively at longer horizons, particularly for high-water-level events.

Figure 5 presents the variation in R^2, RMSE, and MAE for ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model across the 1-, 3-, 5-, 7-, and 10-day forecasting horizons. A clear deterioration in predictive performance is observed with increasing lead time for all four approaches.



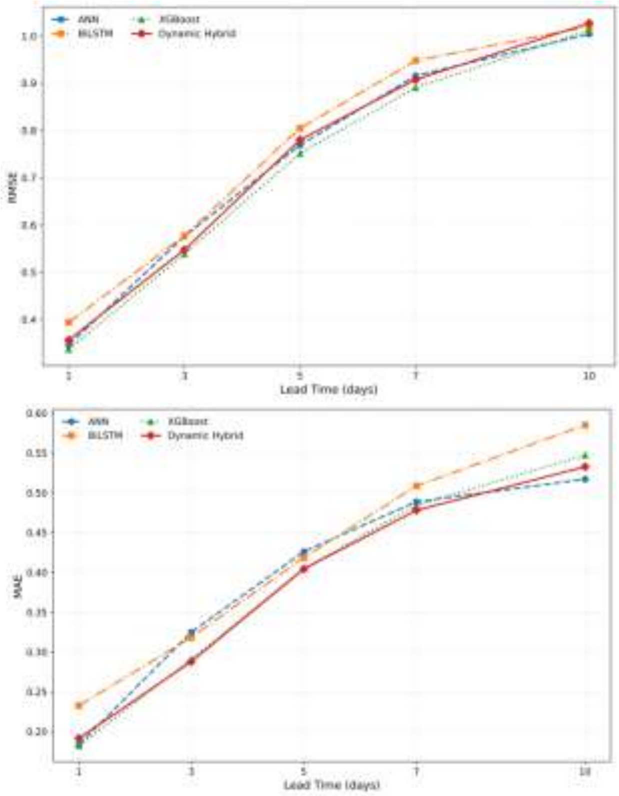
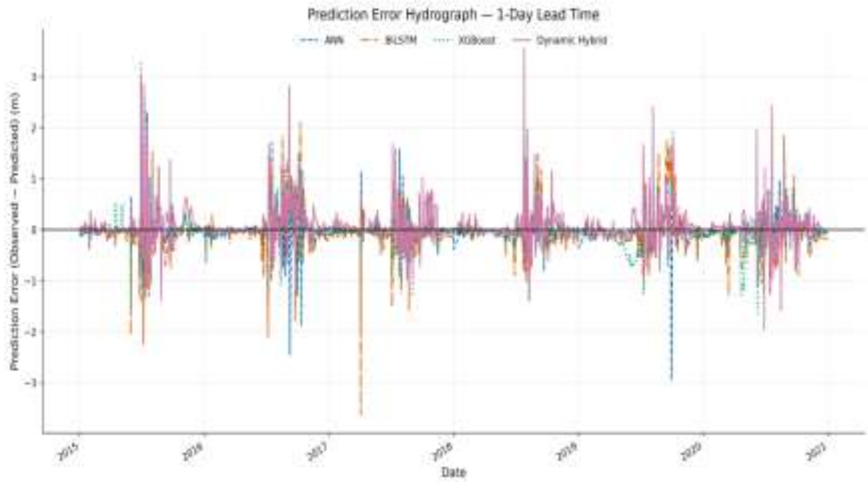
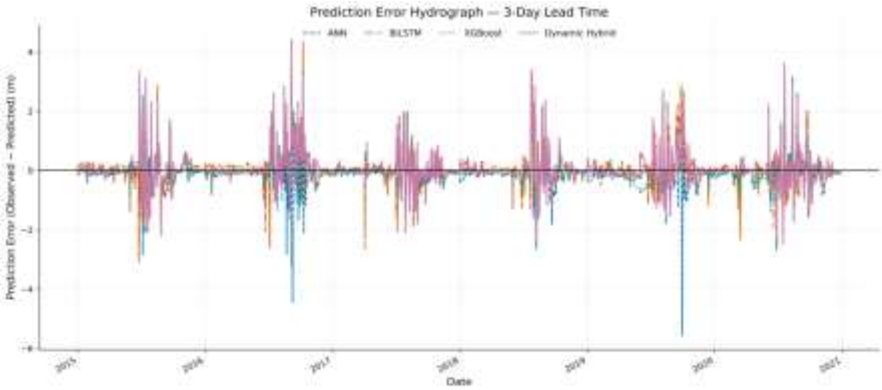


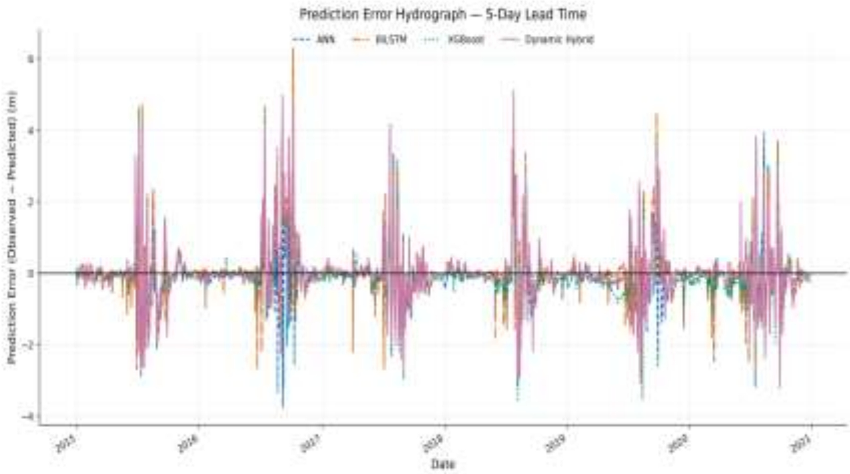
Figure 5: Variation of R^2, RMSE, and MAE for ANN, BiLSTM, XGBoost, and Dynamic Hybrid models across the 1-, 3-, 5-, 7-, and 10-day forecasting horizons.

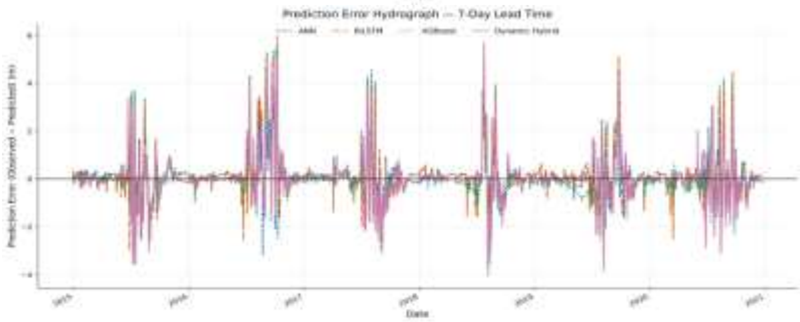
The R^2 values decrease progressively from the 1-day horizon to the 10-day horizon, indicating a reduction in the ability of the models to reproduce the observed variability as the forecast horizon becomes longer. At the 1-day lead time, the R^2 values were 0.965, 0.955, 0.967, and 0.963 for ANN, BiLSTM, XGBoost, and Dynamic Hybrid, respectively. At 3 days, these values decreased to 0.903, 0.903, 0.915, and 0.912, while at 5 days they further declined to 0.827, 0.811, 0.835, and 0.822. A similar reduction was observed at 7 days, with R^2 values of 0.755, 0.737, 0.768, and 0.759, and at 10 days the corresponding values were 0.705, 0.696, 0.700, and 0.691. The error-based metrics show the opposite trend, with RMSE increasing from 0.346, 0.394, 0.337, and 0.357 m at 1 day to 1.004, 1.020, 1.012, and 1.027 m at 10 days for ANN, BiLSTM, XGBoost, and Dynamic Hybrid, respectively. MAE likewise increased with lead time, although the Dynamic Hybrid maintained comparatively lower MAE values at the longer horizons, decreasing the absolute error relative to the other approaches from 3 days onward. Specifically, the Dynamic Hybrid produced MAE values of 0.288, 0.404, 0.478, and 0.533 m at the 3-, 5-, 7-, and 10-day horizons, respectively, compared with 0.325, 0.426, 0.489, and 0.517 m for ANN; 0.318, 0.419, 0.509, and 0.585 m for BiLSTM; and 0.291, 0.405, 0.484, and 0.547 m for XGBoost. The results therefore indicate that the relative behaviour of the models depends on the performance measure considered: XGBoost provides the highest R^2 at the 1-, 3-, 5-, and 7-day horizons and the lowest RMSE over these horizons, whereas ANN records the highest R^2 at 10 days and the lowest RMSE at that horizon, while Dynamic Hybrid provides the lowest MAE from 3 to 10 days. Overall, the progressive decline in R^2 and increase in RMSE and MAE demonstrate the increasing difficulty of multi-step water-level forecasting as the prediction horizon extends, while the differences among the models become more apparent at the longer lead times. These results are consistent with the broader observation in hydrological forecasting that prediction uncertainty tends to increase as the forecasting horizon increases.

Figure 6 presents the temporal error hydrographs for ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model at 1-, 3-, 5-, 7-, and 10-day lead times, respectively, where the prediction error is expressed as the difference between the observed and predicted water levels. At the 1-day lead time, the errors for all models remain generally close to the zero-error line during the majority of the testing period, with larger positive and negative deviations occurring mainly around periods of rapid water-level rise and peak events.


Prediction Error Hydrograph — 1-Day Lead Time
ANN
BiLSTM
XGBoost
Dynamic Hybrid
Prediction Error (Observed − Predicted) (m)
Date


Prediction Error Hydrograph — 3-Day Lead Time
ANN
BiLSTM
XGBoost
Dynamic Hybrid
Prediction Error (Observed − Predicted) (m)
Date


Prediction Error Hydrograph — 5-Day Lead Time
ANN
BiLSTM
XGBoost
Dynamic Hybrid
Prediction Error (Observed − Predicted) (m)
Date


Prediction Error Hydrograph — 7-Day Lead Time
ANN
BiLSTM
XGBoost
Dynamic Hybrid
Date

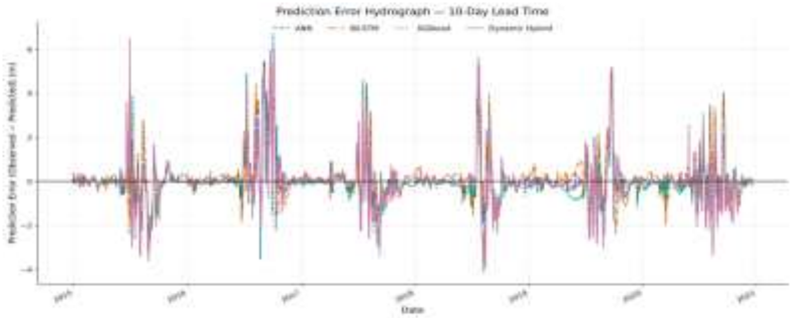


Figure 6: Temporal prediction error hydrographs for ANN, BiLSTM, XGBoost, and Dynamic Hybrid models at 1-, 3-, 5-, 7-, and 10-day lead times

This indicates that the models reproduce the relatively stable portions of the hydrograph more consistently than the rapidly changing portions. At the 3-day horizon, the magnitude and frequency of error excursions increase, particularly around the major high-water events, with several pronounced deviations occurring in the ANN and BiLSTM series. The 5-day results show a further widening of the error fluctuations, including larger departures during both rising and recession phases of major events, while the errors during relatively stable periods remain comparatively small. At 7 days, the error series exhibits greater variability and more frequent departures from zero, particularly around the high-water events distributed throughout the testing period. The largest error fluctuations are observed at the 10-day lead time, where all models exhibit greater positive and negative deviations during several major events, demonstrating the increasing difficulty of accurately reproducing rapid water-level changes at extended forecasting horizons. Across all lead times, the error series tend to remain comparatively small during lower and more stable water-level periods, whereas the largest deviations are concentrated around high-magnitude events. The Dynamic Hybrid model also shows substantial event-specific error fluctuations, indicating that its overall lower MAE at the longer lead times does not imply uniformly smaller errors at every individual observation. Overall, the temporal error analysis complements the statistical performance measures by showing how prediction errors evolve over time and where the models experience their greatest deviations from the observed water-level series.

The progressive increase in error variability with lead time is consistent with the reduction in $R^2$and increase in RMSE and MAE observed in the overall performance analysis, while the concentration of larger errors around high-water-level events emphasizes the importance of the subsequent low-, medium-, and high-regime comparison presented in Table 1. Similar error-based graphical assessments have been used in hydrological forecasting to identify temporal error patterns and the behaviour of models during rapidly varying hydrological conditions.

Table 1: Regime-wise testing performance of ANN, BiLSTM, XGBoost, and Dynamic Hybrid models in terms of R^2, RMSE, and MAE for low, medium, and high water-level conditions across different forecasting lead times.

1-Day				
Regime	**ANN (R^2/RMSE/MAE)**	**BiLSTM (R^2/RMSE/MAE)**	**XGBoost (R^2/RMSE/MAE)**	**Dynamic Hybrid (R^2/RMSE/MAE)**
Low	0.325/0.164/0.093	-0.570/0.250/0.139	-0.597/0.252/0.142	0.726/0.104/0.073
Medium	0.241/0.154/0.094	-0.770/0.235/0.134	0.598/0.112/0.062	0.646/0.105/0.079
High	0.920/0.554/0.360	0.910/0.588/0.422	0.931/0.514/0.344	0.907/0.597/0.419
3-Day				
Regime	**ANN (R^2/RMSE/MAE)**	**BiLSTM (R^2/RMSE/MAE)**	**XGBoost (R^2/RMSE/MAE)**	**Dynamic Hybrid (R^2/RMSE/MAE)**
Low	-0.169/0.215/0.157	-0.820/0.269/0.140	-0.550/0.248/0.164	0.444/0.149/0.109
Medium	-0.795/0.237/0.159	-1.481/0.279/0.166	0.259/0.152/0.094	0.465/0.129/0.082
High	0.769/0.940/0.654	0.780/0.918/0.643	0.796/0.883/0.613	0.777/0.923/0.668

5-Day				
Regime	**ANN (R^2/RMSE/MAE)**	**BiLSTM (R^2/RMSE/MAE)**	**XGBoost (R^2/RMSE/MAE)**	**Dynamic Hybrid (R^2/RMSE/MAE)**
Low	-0.953/0.279/0.206	-0.798/0.267/0.13	-0.912/0.276/0.199	0.645/0.119/0.090
Medium	-1.106/0.257/0.173	-1.934/0.303/0.170	-0.336/0.205/0.125	-1.075/0.255/0.148
High	0.578/1.271/0.894	0.540/1.327/0.939	0.592/1.250/0.886	0.549/1.314/0.964
7-Day				
Regime	**ANN (R^2/RMSE/MAE)**	**BiLSTM (R^2/RMSE/MAE)**	**XGBoost (R^2/RMSE/MAE)**	**Dynamic Hybrid (R^2/RMSE/MAE)**
Low	-0.494/0.244/0.174	-1.242/0.298/0.200	-1.109/0.289/0.211	0.487/0.143/0.107
Medium	-1.044/0.253/0.196	-2.496/0.331/0.216	-1.460/0.277/0.151	-1.938/0.303/0.151
High	0.381/1.539/1.086	0.355/1.572/1.101	0.426/1.483/1.084	0.391/1.527/1.164
10-Day				
Regime	**ANN (R^2/RMSE/MAE)**	**BiLSTM (R^2/RMSE/MAE)**	**XGBoost (R^2/RMSE/MAE)**	**Dynamic Hybrid (R^2/RMSE/MAE)**
Low	-0.809/0.268/0.183	-2.263/0.360/0.261	-1.575/0.320/0.241	0.314/0.165/0.131
Medium	-2.017/0.307/0.190	-5.353/0.446/0.284	-2.681/0.340/0.185	-3.407/0.372/0.156
High	0.261/1.682/1.168	0.278/1.662/1.200	0.262/1.681/1.208	0.226/1.722/1.298

To examine whether model performance varied under different hydrological conditions, the testing observations were further classified into low, medium, and high water-level regimes using the 33rd and 66th percentile thresholds of the observed target water level. The same observed-regime classification was then applied to the predictions from ANN, BiLSTM, XGBoost, and the Dynamic Hybrid model, allowing all four approaches to be compared under identical water-level conditions. The regime-wise results are summarized in Table 1 in terms of R^2, RMSE, and MAE for each forecasting horizon. A distinct pattern was observed across the three regimes. Under low-water-level conditions, the Dynamic Hybrid model consistently produced positive R^2 values at all lead times, ranging from 0.726 at 1 day to 0.314 at 10 days, while the standalone models showed substantially lower or negative R^2 values at most longer horizons. The Dynamic Hybrid also yielded the lowest RMSE and MAE for the low regime across all five forecasting horizons, indicating comparatively smaller prediction deviations during low-water-level conditions. For the medium regime, model behaviour was less consistent. At the 1-day horizon, XGBoost produced an R^2 of 0.598, compared with 0.646 for the Dynamic Hybrid, while at 3 days the Dynamic Hybrid reached 0.465 compared with 0.259 for XGBoost. At longer horizons, however, the R^2 values of all models became negative, indicating that the restricted variability within the medium-water-level range made this regime more difficult to predict accurately.

The error measures showed that XGBoost and Dynamic Hybrid generally maintained comparatively smaller errors than BiLSTM, with the Dynamic Hybrid exhibiting the lowest MAE at 3 and 10 days. Under high-water-level conditions, the models retained substantially higher R^2 values than those obtained for the low and medium regimes. XGBoost produced the highest R^2 at 1, 3, 5, and 7 days, with values of 0.931, 0.796, 0.592, and 0.426, respectively, while BiLSTM recorded the highest R^2 at 10 days (0.278). XGBoost also generally exhibited lower RMSE during the high-water-level conditions up to the 7-day horizon. These results indicate that model behaviour depends strongly on the prevailing water-level regime: the Dynamic Hybrid showed comparatively stable performance for low-water-level conditions, whereas XGBoost provided relatively strong performance for high-water-level events. The medium regime exhibited greater uncertainty and larger relative deviations, particularly at the longer lead times. Thus, the regime-wise analysis demonstrates that overall performance metrics alone may not fully represent how individual models behave under different water-level conditions, and the inclusion of low-, medium-, and high-regime comparisons provides an additional perspective on the predictive characteristics of the developed forecasting approaches. The percentile-based regime formulation is consistent with the dynamic flow-regime framework in which different hydrological conditions are treated separately rather than assuming uniform model behaviour across the complete water-level range.

The combined analysis of the statistical and regime-wise results indicates that model performance varied with both forecasting horizon and water-level condition. At shorter lead times, all four models reproduced the observed water-level variability relatively well, as reflected by the high overall R^2values and comparatively low error magnitudes. With increasing lead time, the agreement between observed and predicted water levels gradually decreased, while RMSE and MAE generally increased, demonstrating the increasing uncertainty associated with extended forecasting horizons. The regime-wise analysis provided an additional perspective by showing that this degradation was not uniform across water-level conditions. The Dynamic Hybrid model maintained positive R^2values for the low-water-level regime at all five forecasting horizons and produced comparatively smaller RMSE and MAE values within this regime. In contrast, XGBoost showed relatively strong performance for the high-water-level regime, particularly at the shorter and intermediate lead times. The medium-water-level regime exhibited greater variability among the models, with negative R^2values appearing at several longer forecasting horizons. This behaviour indicates that aggregated model statistics can mask differences in predictive behaviour under specific hydrological conditions. The results therefore demonstrate the value of combining overall multi-step evaluation with regime-wise analysis, as the predictive characteristics of ANN, BiLSTM, XGBoost, and Dynamic Hybrid differed according to both forecast horizon and water-level condition.

Conclusions

The present study developed a multi-step river water-level forecasting framework for the Punpun River using Artificial Neural Network (ANN), Bidirectional Long Short-Term Memory (BiLSTM), XGBoost, and a Dynamic Hybrid approach, with forecasts generated for 1-, 3-, 5-, 7-, and 10-day lead times. The results demonstrated that forecasting performance was strongly dependent on the prediction horizon, with all models showing a gradual reduction in $R^2$and an increase in prediction errors as the lead time increased, reflecting the greater uncertainty associated with extended forecasting. At shorter horizons, particularly 1–7 days, XGBoost showed comparatively higher $R^2$and lower RMSE values, whereas ANN showed the highest $R^2$and lowest RMSE at the 10-day horizon. The Dynamic Hybrid model produced comparatively lower MAE values from the 3-day to the 10-day horizons, indicating its ability to reduce average absolute prediction error over the longer forecasting periods. The regime-wise assessment further demonstrated that model behaviour varied with water-level conditions: the Dynamic Hybrid maintained positive R^2values with comparatively lower errors for the low-water-level regime across all lead times, while XGBoost showed comparatively strong performance for high-water-level conditions at shorter and intermediate horizons; in contrast, the medium-water-level regime exhibited greater variability and reduced predictive agreement, particularly at longer lead times. Overall, the findings show that the predictive characteristics of ANN, BiLSTM, XGBoost, and Dynamic Hybrid were influenced by both forecasting horizon and hydrological regime, highlighting the importance of evaluating multi-step forecasts under different water-level conditions rather than relying only on aggregate performance measures. The increasing uncertainty observed at longer lead times also indicates scope for further improvement through the incorporation of additional hydro-meteorological and spatial information, improved representation of extreme events, and advanced hybrid or ensemble modelling approaches.

References

1. Ahmed, A. N., Yafouz, A., Birima, A. H., Kisi, O., Huang, Y. F., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2022). Water level prediction using various machine learning algorithms: A case study of Durian Tunggal River, Malaysia. Engineering Applications of Computational Fluid Mechanics, 16(1), 422–440.
2. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
3. Chaubey, P. K., Srivastava, P. K., Gupta, A., & Mall, R. K. (2021). Integrated assessment of extreme events and hydrological responses of Indo-Nepal Gandak River Basin. Environment, Development and Sustainability, 23(6), 8643–8668.
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
5. Cho, K., & Kim, Y. (2022). Improving streamflow prediction in the WRF-Hydro model with LSTM networks. Journal of Hydrology, 605, 127297.
6. Hipni, A., El-shafie, A., Najah, A., Karim, O. A., Hussain, A., & Mukhlisin, M. (2013). Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro-fuzzy inference system (ANFIS). Water Resources Management, 27(10), 3803–3823.
7. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

8. Hunt, K. M., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrology and Earth System Sciences, 26(21), 5449–5472.
9. Kao, I.-F., Zhou, Y., Chang, L.-C., & Chang, F.-J. (2020). Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. Journal of Hydrology, 583, 124631.
10. Konapala, G., Kao, S. C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters, 15(10), 104022.
11. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrology and Earth System Sciences, 22(11), 6005–6022.
12. Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environmental Modelling & Software.
13. Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885–900.
14. Prakash, R., & Tripura, J. (2026). Comparative insights into short-term streamflow forecasting in navigating urban flood risks. Proceedings of the Institution of Civil Engineers – Water Management.
15. Ranjan, A., & Roshni, T. (2023). Analysis of hydrological alteration and environmental flow in Sone River Basin. Acta Geophysica, 71(2), 949–960.
16. Sapitang, M., Wanie, M. R., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. (2020). Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. Sustainability, 12(15), 6121.
17. Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrology and Earth System Sciences, 20(7), 2611–2628.
18. Tripura, J., Roy, P., & Barbhuiya, A. K. (2021). Simultaneous streamflow forecasting based on hybridized neuro-fuzzy method for a river system. Neural Computing and Applications, 33(8), 3221–3233.
19. Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., & Zhao, B. (2020). A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. Journal of Hydrology, 590, 125206.
20. Zakaria, M. N. A., Malek, M. A., Zolkepli, M., & Ahmed, A. N. (2021). Application of artificial intelligence algorithms for hourly river level forecast: A case study of Muda River, Malaysia. Alexandria Engineering Journal, 60(4), 4015–4028.
21. Zhu, S., Lu, H., Ptak, M., Dai, J., & Ji, Q. (2020). Lake water-level fluctuation forecasting using machine learning models: A systematic review. Environmental Science and Pollution Research, 27(36), 44807–44819.
22. Zounemat-Kermani, M., et al. (2020). Neurocomputing in surface water hydrology and hydraulics: A review of two decades retrospective, current status and future prospects. Journal of Hydrology, 588, 125085.