1 Forecasting the Discharge for the River of Bosna

2 Computational Framework

2.1 Compared Models

In this study, three data-driven modeling approaches were evaluated for predicting the monthly mean river discharge (*proticaj*) of the Bosna River: *Random Forest* (RF), *Extreme Gradient Boosting* (XGBoost), and *Long Short-Term Memory* neural networks (LSTM).

Random Forest is an ensemble learning method based on bootstrap aggregation of decision trees. Due to its robustness to noise, ability to capture nonlinear relationships, and limited sensitivity to multicollinearity, RF is frequently applied in hydrological modeling and is considered a strong baseline approach.

XGBoost is a gradient boosting framework that sequentially constructs decision trees, where each tree aims to correct the errors of its predecessors. Compared to RF, XGBoost offers enhanced control over model complexity through regularization and learning-rate parameters, and has shown strong performance in hydro-climatic prediction tasks.

LSTM represents a class of recurrent neural networks specifically designed to capture temporal dependencies in sequential data. It is included in this study to assess whether deep learning models can extract additional predictive skill from long-term hydro-climatic time series compared to tree-based ensemble methods.

Together, these models represent complementary modeling paradigms: bagging-based ensembles (RF), boosting-based ensembles (XGBoost), and sequence-based deep learning (LSTM).

2.2 Dataset Description

The dataset consists of monthly aggregated observations spanning the period from 1961 to 2020. Each record corresponds to a single month.

Meteorological predictors include the mean monthly precipitation (P) and mean monthly air temperature (T), observed at five meteorological stations located within the Bosna River basin: Bjelašnica, Sarajevo, Zenica, Tuzla, and Doboj. The target variable for prediction is the monthly mean river discharge (Q), measured at the basin outlet. All variables were aggregated to monthly resolution to ensure temporal consistency and long-term coverage.

We noticed a few missing data for the Zenica station within a gap of 6 month in 2017. These are properly filled in by applying the Climatological Monthly Mean Filling across all historical data for that station.

2.3 Feature Engineering and Lagged Variables

To account for hydrological memory effects and delayed basin response, lagged predictor variables were introduced. Let P_t, T_t, and Q_t denote precipitation, temperature, and discharge at time step t, respectively.

The following lag structure was applied:

- Precipitation lags: P_{t-1}, P_{t-2}, P_{t-3}
- Temperature lags: T_{t-1}, T_{t-2}
- Autoregressive discharge term: Q_{t-1}

The autoregressive discharge term represents short-term persistence in river flow and is commonly used in hydrological forecasting to improve predictive performance. Lagged features were generated independently for each station and variable. Observations containing missing values caused by lagging were removed prior to model training.

2.4 Performance Metrics

Model performance was evaluated using hydrologically relevant efficiency metrics. Besides Minimum Absolute Error (MAE) and the Root Mean Square Error (RMSE), Nash–Sutcliffe measure, and Kling–Gupta measure are determined.

The Nash–Sutcliffe Efficiency (NSE) measures the predictive skill of a model relative to the mean of observed discharge values and is defined as:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}, \tag{1}$$

where Q_i and $\hat{Q}_i$ denote observed and predicted discharge values, respectively, and $\bar{Q}$ is the mean of observed discharge.

The Kling–Gupta Efficiency (KGE) decomposes model performance into correlation, bias, and variability components:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}, \tag{2}$$

where r is the Pearson correlation coefficient, β is the bias ratio, and γ is the variability ratio between predicted and observed discharge.

NSE was used as the primary optimization criterion during model tuning, while KGE was employed for complementary evaluation.

2.5 Training, Validation, and Temporal Splitting

To respect the temporal structure of the data and avoid information leakage, a time-based splitting strategy was adopted. The dataset was divided into:

- Training period: 1961–2010
- Testing period: 2011–2020

Within the training period, *rolling-origin cross-validation* was implemented using a time-series split with five folds. In this setup, validation sets always occur chronologically after the corresponding training subsets.

2.6 Hyperparameter Tuning of XGBoost, RF, and LSTM

To tune the parameters of all three models, the basic grid search technique combined with time-series cross-validation is utilized on the train data. For example, for XGBoost, the following parameters were considered:

- Number of trees ($n_{\text{estimators}}$)
- Maximum tree depth (max_depth)
- Learning rate (η)
- Subsampling ratio
- Column subsampling ratio

For each parameter combination, initially pre-defined, model performance was evaluated using the mean NSE across the five validation folds. The parameter set yielding the highest NSE was selected as optimal.

The following values of parameters are produced from the grid search:

- RF: max_depth: 30, max_features: sqrt, min_samples_leaf: 1 min_samples_split: 2, n_estimators: 2000
- XGBoost: colsample_bytree: 0.8 learning_rate: 0.05, max_depth: 4, n_estimators: 600, subsample: 0.7
- LSTM: units: 64, lr: 0.001, batch_size: 16, epochs: 200,

2.7 One-Step-Ahead Forecasting and Bias Correction

Using the optimized model, one-step-ahead forecasts were generated for the testing period. To reduce systematic bias, a linear scaling correction was applied. A linear regression model was fitted between predicted and observed discharge values on the training set:

$$Q_{\text{obs}} = a \cdot Q_{\text{pred}} + b, \tag{3}$$

where a and b are regression coefficients estimated from the training data.

The derived correction was subsequently applied to test-period predictions, preserving temporal dynamics while improving distributional consistency.

2.8 Summary of the Computational Workflow

The complete modeling workflow consists of the following steps:

1. Construction of lagged hydro-meteorological features
2. Time-based train–test splitting
3. Rolling-origin cross-validation for hyperparameter tuning

4. Model selection based on NSE
5. One-step-ahead discharge forecasting
6. Linear bias correction
7. Performance evaluation using NSE, KGE, RMSE, and MAE

2.9 Results & Discussion

Table 1: Performance comparison of the evaluated models on training and testing data. RMSE and MAE are expressed in discharge units, while NSE and KGE are dimensionless.

Model	Dataset	RMSE	MAE	NSE	KGE
XGBoost	Training	2.920	2.208	**0.999**	**0.993**
	Testing	64.680	48.029	0.589	0.589
Random Forest	Training	20.769	14.888	0.964	0.884
	Testing	64.500	47.169	**0.591**	**0.591**
LSTM	Training	81.581	57.432	0.443	0.382
	Testing	75.086	54.763	0.445	0.489

Discussion. The performance metrics presented in Table 1 reveal notable differences in both fitting capacity and generalization ability among the evaluated models.

XGBoost achieves near-perfect accuracy on the training dataset (NSE = 0.999, KGE = 0.993), indicating a very strong ability to capture nonlinear relationships between hydro-meteorological predictors and river discharge. However, this excellent training performance does not fully translate to the testing period, where NSE and KGE drop to approximately 0.59. The substantial gap between training and testing accuracy suggests that XGBoost may partially overfit the historical training data, despite the use of time-series cross-validation and regularization parameters.

Random Forest exhibits slightly lower training accuracy (NSE = 0.964, KGE = 0.884) compared to XGBoost, but achieves nearly identical performance on the testing dataset (NSE = 0.591, KGE = 0.591). This behavior indicates a more balanced bias–variance trade-off and suggests that Random Forest generalizes marginally better to unseen data. The similarity of test performance between Random Forest and XGBoost implies that, for the considered feature set and temporal resolution, additional model complexity does not yield substantial gains in predictive skill.

The LSTM model shows the weakest performance among the evaluated approaches. Training accuracy is moderate (NSE = 0.443), and testing performance remains limited (NSE = 0.445), indicating that the model struggles to effectively learn the discharge dynamics. Several factors may contribute to this behavior. First, the dataset consists of monthly aggregated observations, which

may not provide sufficient temporal resolution for LSTM models to exploit long-range dependencies. Second, the relatively limited number of training samples compared to the high parameterization of LSTM networks may hinder stable learning. Finally, the use of a single-step input structure with engineered lag features may reduce the advantage of sequence-based deep learning relative to tree-based methods.

Overall, the results demonstrate that tree-based ensemble models outperform deep learning approaches for monthly discharge prediction of the Bosna River under the given data availability and feature configuration. While XGBoost achieves the highest fitting accuracy, Random Forest provides comparable test performance with greater robustness. The observed performance gap between training and testing across all models also highlights the challenge of nonstationarity in long-term hydro-climatic time series, emphasizing the importance of careful validation strategies and model selection.

These findings suggest that, for long-term monthly discharge prediction in data-limited settings, well-regularized ensemble methods combined with physically motivated lag features can provide reliable predictive performance without the complexity of deep learning architectures.

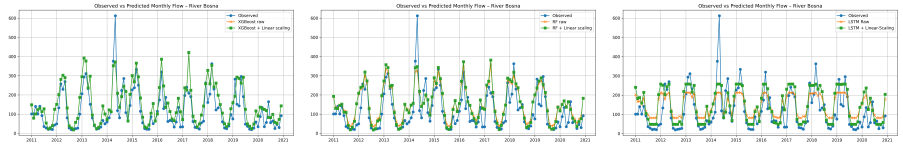

Figure 1: Observed versus predicted monthly river discharge during the testing period (2011–2020) for the evaluated models: (a) XGBoost, (b) Random Forest, and (c) LSTM. All predictions correspond to one-step-ahead forecasts after linear bias correction.

3 Feature Importance and Explainable Artificial Intelligence

While predictive accuracy is essential, understanding the contribution of individual hydro-meteorological predictors is equally important for decision support and scientific insight. To this end, an Explainable Artificial Intelligence (XAI) analysis was conducted to assess the relative importance of input features and their influence on river discharge prediction.

Given the heterogeneous nature of the evaluated models, different yet conceptually consistent feature importance approaches were employed. For the tree-based models (XGBoost and Random Forest), feature importance was derived from model-based importance measures, while for the LSTM model a permutation-based approach was adopted to ensure comparability across model classes.

3.1 Permutation-Based Feature Importance

Permutation importance quantifies the contribution of a feature by measuring the decrease in model performance when the feature's values are randomly permuted. Let f denote a trained model and $\mathcal{M}(\cdot)$ a performance metric, here the Nash–Sutcliffe Efficiency (NSE). The importance of feature x_j is computed as:

$$I_j = \mathcal{M}(f(\mathbf{X})) - \mathcal{M}(f(\mathbf{X}_{\pi(j)})), \tag{4}$$

where $\mathbf{X}_{\pi(j)}$ denotes the input matrix with feature x_j permuted across samples.

A larger decrease in NSE indicates a more influential feature. This model-agnostic approach enables consistent interpretation across both tree-based and neural network models and directly reflects the impact of each feature on predictive skill.

3.1.1 Dominant Predictors of River Discharge

The analysis focused on the five most influential features for each model, ranked by their contribution to NSE reduction. Results indicate that lagged precipitation variables consistently dominate model predictions, particularly precipitation from the preceding one to three months. This finding is hydrologically meaningful, reflecting basin-scale storage effects and delayed runoff generation.

The autoregressive discharge term (Q_{t-1}) also emerges as a key predictor, highlighting the strong persistence of monthly river flow and confirming the importance of short-term memory in discharge dynamics. Temperature-related features exhibit comparatively equally important influence, especially `T_bjelasnica`, likely capturing seasonal snowmelt and evapotranspiration effects, especially for upstream stations such as Bjelašnica.

Notably, feature importance rankings remain broadly consistent across XGBoost and Random Forest, especially the two most inflential features—`T_bjelasnica` and the auto-regressive `Q_lag1`—suggesting stable hydro-climatic controls on discharge. In contrast, the LSTM model displays weaker and less structured importance patterns, consistent with its overall lower predictive performance. The agreement between data-driven feature importance and established hydrological understanding provides confidence in the physical plausibility of the models. The dominance of lagged precipitation and discharge terms in both reasonably efficient models aligns with conceptual rainfall–runoff processes, while the secondary role of temperature in Bjelasnica reflects its indirect influence at monthly time scales.

From an operational perspective, these results demonstrate that explainable machine learning techniques can support transparent and trustworthy discharge forecasting. By identifying the most influential predictors, the proposed framework enables both improved model diagnostics and enhanced interpretability for water resource management and decision-making.

Overall, the XAI analysis confirms that the tree-based ensemble models not only achieve higher predictive skill but also produce feature importance patterns

that are consistent with hydrological theory, reinforcing their suitability for long-term monthly discharge prediction in the Bosna River basin.

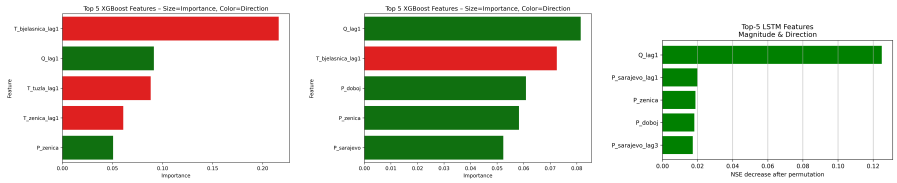

Figure 2: Direction and magnitude of the five most influential features identified through explainable artificial intelligence analysis: (a) XGBoost, (b) Random Forest, and (c) LSTM. Bars indicate the decrease in Nash–Sutcliffe Efficiency (NSE) after feature permutation, while color denotes the direction of influence (positive (green) or negative (red) contribution to predicted discharge).

Figure 2 illustrates the relative importance and directional influence of the dominant predictors for each model, highlighting the consistent dominance of lagged precipitation and autoregressive discharge terms across tree-based ensemble models.