P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

EXTREME GRADIENT BOOSTING -DRIVEN HYDROLOGICAL FORECASTING STREAMFLOW UNDER CMIP6 SCENARİOS

Abdulhadi PALA[1]/ Prof.Dr., Aytaç GÜVEN [2]
[1]Gaziantep University, Civil Engineering, ahpala@gantep.edu.tr, 0000-0003-3090-2171
[2]Gaziantep University, Civil Engineering, aguven@gantep.edu.tr, 0000-0002-1935-1375

Abstract

Accurate streamflow forecasting is essential for sustainable water resource planning under increasing climate variability. This study investigates the performance of the Extreme Gradient Boosting (XGBoost) algorithm for monthly streamflow prediction in the Ceyhan River Basin, utilizing climate projections from General Circulation Models (GCMs) participating in CMIP6. The model was tested across four Shared Socioeconomic Pathways (SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5), using historical discharge data and climate predictors such as mean sea level pressure (mslp) and specific humidity (shum) as inputs. Evaluation metrics included RMSE, MAE, and the Nash–Sutcliffe Efficiency (NSE). The results show that the XGBoost model dynamically adapts to seasonal patterns and selects the best-performing configuration each month. Under high-emission scenarios, particularly SSP5-8.5, streamflow variability significantly increased. SHAP (SHapley Additive exPlanations) analysis enhanced model interpretability, identifying shum and mslp as key predictive features. These findings highlight the potential of XGBoost as a robust and explainable decision-support tool for climate-informed hydrological modeling.

Keywords: XGBoost, Streamflow forecasting, SHAP analysis, CMIP6, SSP Scenarios

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

1 Introduction

Accurate streamflow prediction is essential for sustainable water resource management, especially given the increasing impacts of climate variability and extreme events such as floods and droughts [1]. As the hydrological cycle is highly sensitive to changes in precipitation, temperature, and atmospheric circulation patterns, forecasting streamflow remains a major challenge in hydrological sciences [2].

To address these forecasting challenges, climate scientists rely on General Circulation Models (GCMs), which simulate the Earth's climate system and provide projections of future climatic variables such as temperature and precipitation. GCMs operate at a global scale and are essential for understanding the long-term impacts of greenhouse gas emissions on regional hydrological processes. However, due to their coarse spatial resolution, GCM outputs often require downscaling before they can be applied to basin-scale hydrological models [3].

To incorporate socioeconomic and emission-related uncertainties, climate projections are categorized under Shared Socioeconomic Pathways (SSPs). These pathways such as SSP1-2.6 (sustainable), SSP2-4.5 (middle of the road), SSP3-7.0 (regional rivalry), and SSP5-8.5 (fossil-fueled development) eflect different global development trajectories. They are used to assess how future social, technological, and economic developments may influence greenhouse gas emissions and consequently hydrological dynamics [4].

CMIP6 (Coupled Model Intercomparison Project Phase 6) provides a coordinated framework for climate model experiments based on these SSPs. It includes a wide array of GCM simulations under varying assumptions and has become a standard reference for hydrological studies evaluating future water availability and extremes. CMIP6 allows researchers to evaluate the effects of climate change under multiple scenarios and timeframes, enhancing decision-making for sustainable water and energy management [5].

In response to these challenges, machine learning (ML) methods have gained increasing attention for their capacity to capture complex, non-linear relationships among hydrological variables. Recent research has applied various ML algorithms—including Support Vector Machines (SVM), Random Forest (RF), CatBoost,

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

LightGBM, and XGBoost—for discharge prediction in different river basins [6]. Among these, XGBoost (Extreme Gradient Boosting) stands out due to its high predictive accuracy, scalability, and strong regularization capabilities that help prevent overfitting [7].

Studies conducted in the Ceyhan River Basin have primarily explored hybrid modeling approaches or comparisons among different algorithms using CMIP6 climate projections and Shared Socioeconomic Pathways (SSPs) [8]. However, there is a noticeable lack of research focusing solely on the standalone performance of XGBoost in combination with SHAP (SHapley Additive exPlanations) analysis for interpretable and climate-informed streamflow prediction.

This study addresses this research gap by applying XGBoost to forecast monthly streamflow in the Ceyhan River Basin using historical discharge data and climate variables extracted from four SSP scenarios (SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5). Model performance is evaluated through RMSE, MAE, and NSE metrics, while SHAP analysis is employed to identify the key drivers of prediction. The results aim to support more robust and transparent hydrological planning under future climate conditions [9].

2 METHODOLOGY

2.1 Study Area and Dataset

Located in the eastern Mediterranean region of Turkey, the Ceyhan River Basin spans approximately 20,000 square kilometers. This basin is a vital water resource for agricultural and hydroelectric activities in the region [10]. According to the General Directorate of State Hydraulic Works (DSI), the basin experiences a Mediterranean

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

climate characterized by wet winters and dry summers. The primary hydrological station utilized in this study is positioned at the Menzelet Dam, offering reliable and long-term discharge data. Monthly streamflow observations from 1984 to 2014 were sourced from DSI records [11].

In addition to streamflow data, climate variables including precipitation (prcp), mean sea level pressure (mslp), specific humidity (shum), temperature (temp), and surface wind speed (sfcWind) were collected from CMIP6 climate models, specifically under four SSP scenarios: SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5. These variables were downloaded from the Canadian Climate Data and Scenarios (CCDS) platform using the NorESM2-MM model, which offers a balanced representation of global climate systems. [12].

The selection of the NorESM2-MM model was based on its demonstrated accuracy in reproducing historical climate conditions over the Eastern Mediterranean region, particularly in terms of seasonal temperature and precipitation variability. This model effectively captures coupled interactions among land, ocean, cryosphere, and atmosphere, and has shown improved skill in representing monsoonal and mid-latitude circulations when compared to earlier CMIP5 models [13]. Furthermore, its moderate climate sensitivity and consistent bias correction performance across SSP scenarios make it a suitable choice for hydrological impact assessments.

To assess future hydrological changes, projections were analyzed under four CMIP6 Shared Socioeconomic Pathway (SSP) scenarios: SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5. These scenarios represent different global development pathways and greenhouse gas emission levels:

- SSP1-2.6: A sustainable scenario with low emissions and strong climate policies, targeting a temperature rise below 2°C.

- SSP2-4.5: A middle-of-the-road scenario with moderate challenges to mitigation and adaptation.

- SSP3-7.0: A regional rivalry scenario with high challenges to mitigation and adaptation, and relatively high emissions.

- SSP5-8.5: A fossil-fueled development scenario with very high greenhouse gas emissions and high radiative forcing. [14].

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

2.2 XGBoost Algorithm and Pipeline

Extreme Gradient Boosting (XGBoost) is an efficient and scalable tree-based ensemble machine learning algorithm that iteratively minimizes error using the gradient boosting framework. It has gained popularity due to its scalability, regularization techniques, and high performance in both classification and regression tasks [7].

In this study, a predictive pipeline was developed by integrating climate variables with streamflow observations to estimate monthly discharge. The XGBoost model was trained using a range of hyperparameters, including learning rate, maximum tree depth, and the number of estimators, which were optimized through grid search in conjunction with k-fold cross-validation [15].

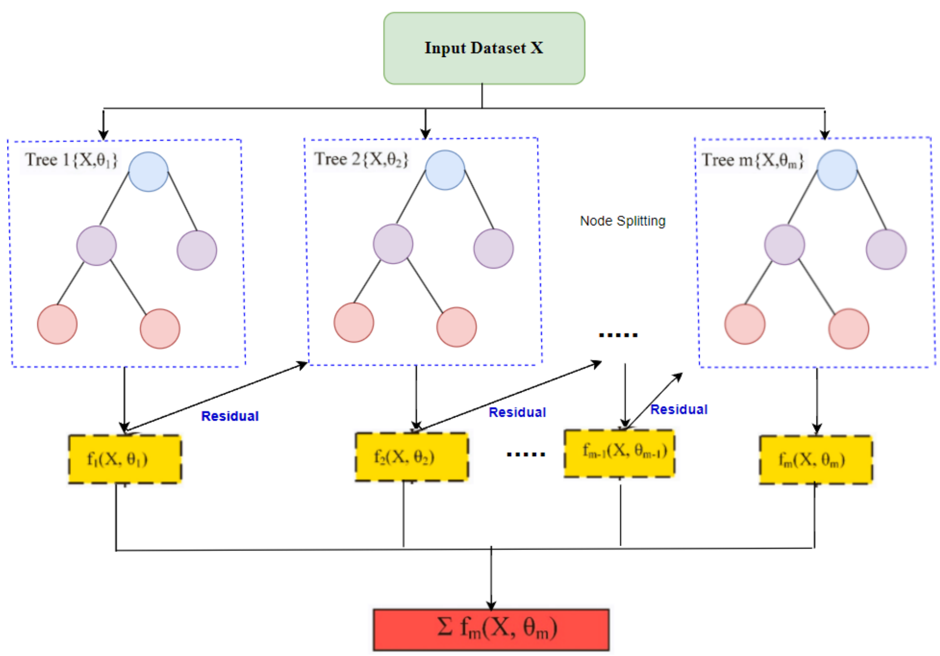

Figure 1. Structure of XGBooost algorithm

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

Figure 1 presents a schematic view of the XGBoost algorithm demonstrating how discharge and climate data were processed through successive boosting rounds. In each round, residual errors were iteratively minimized to improve prediction accuracy. Following model training, SHAP (SHapley Additive exPlanations) analysis was performed to interpret the contribution of individual predictors. This approach provided transparency in model interpretation and highlighted the relative importance of each input variable in streamflow estimation [16].

2.3 Xgboost Model

The available dataset was split into training (70%) and testing (30%) sets. The training phase involved fitting the XGBoost model using the selected hyperparameters. Model development was conducted independently for each of the four SSP scenarios to evaluate discharge behavior under varying climate conditions [14].

After training, the model was applied to each SSP scenario's climate data to estimate streamflow values. These estimates were compared with observed records to assess model accuracy and potential applicability for climate-resilient hydrological planning [9].

To ensure the robustness and generalizability of the developed XGBoost model, a 10-fold cross-validation technique was employed during the training phase. This approach aimed to reduce the risk of overfitting and enhance the reliability of model predictions across different subsets of the dataset [6].

To evaluate the model's predictive performance, several error metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Nash–Sutcliffe Efficiency (NSE) were calculated for each SSP scenario. These metrics provide insight into the accuracy and reliability of simulated streamflow values compared to observed data [6].

The model was trained and tested separately for each SSP scenario (SSP1–2.6, SSP2–4.5, SSP3–7.0, and SSP5–8.5) to analyze the sensitivity of future streamflow estimations under varying climate change projections.

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

This approach enables a comprehensive understanding of potential hydrological shifts and supports adaptive water management strategies under uncertainty [9].

Table 1. The optimal hyperparameters for XGBoost model.

Methods	Hyperparameters	Selected value
XGBoost	Verbosity	0
	N estimator	23
	Colsample-by tree	0.4
	Max depth	2
	Subsample	0.3
	Min chid weight	3
	Reg lambda	1.075
	Reg alpha	0.05
	Max bin	260
	Gamma	10

Table 1 presents the optimal hyperparameters selected for the XGBoost model. These hyperparameters were determined through extensive trial-and-error processes, supported by prior literature and optimization strategies such as grid search [7]. This process ensured a balanced trade-off between model bias and variance.

Each hyperparameter influences the model's behavior uniquely. For example, 'max depth' controls the complexity of decision trees, while 'gamma' acts as a regularization term to prevent overfitting. Parameters such as 'subsample' and 'colsample_bytree' randomly select portions of the training data and input features, respectively, enhancing model robustness and generalization [16].

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

2.4 SHAP Interpretation

The Shapley Additive exPlanations (SHAP) method was utilized to evaluate the contribution of individual input variables to the XGBoost model's predictions [16]. By computing the marginal effect of each feature across all possible feature subsets, SHAP enables a robust assessment of variable importance. This analysis highlighted the dominant influence of variables such as precipitation, humidity, and temperature under different climate scenarios.

2.5 Evaluation Metrics

The model's performance was assessed using three widely adopted metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash–Sutcliffe Efficiency (NSE). These metrics were calculated separately for each SSP scenario to enable a comparative evaluation of the XGBoost model's predictive performance under varying climate conditions [17].

3 RESULTS AND DISCUSSION

3.1 Model Performance Evaluation

This section evaluates the performance of the XGBoost model in simulating monthly streamflow for the Ceyhan River Basin under various SSP scenarios. Figures 2 to 6 present both scatter and line plots that illustrate the correlation between observed and predicted discharge during training and testing periods. The model's performance was assessed using RMSE, MAE, and NSE metrics [17]. Additionally, SHAP analysis was applied to interpret the model outputs [16].

P-ISSN: 3050-9572 E-ISSN: 3050-9580

EUROPEAN DATA SCIENCE JOURNAL

Volume: 03 Issue: 03

RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

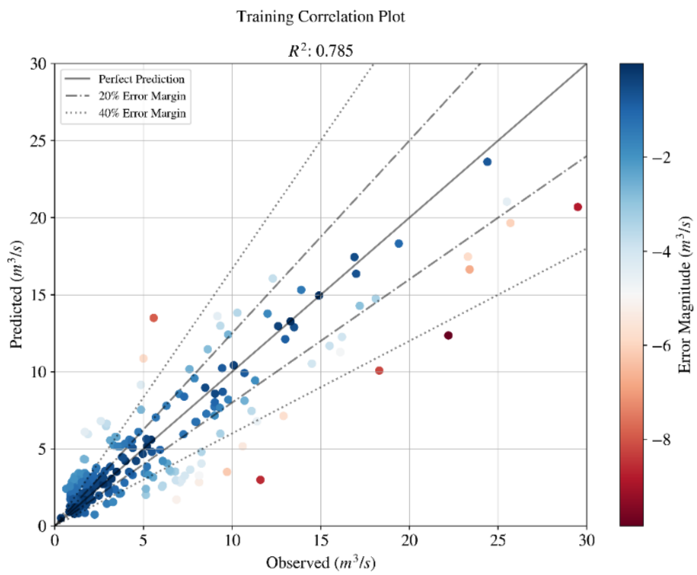

Figure 2. Scatterplot of observed and predicted Q for training period with XGBoost

Figure 2 shows a strong correlation between observed and predicted discharge values during the training period (1984–2004). The dense alignment of data points along the 1:1 reference line confirms the model's high predictive accuracy and validates the XGBoost algorithm's ability to capture hydrological patterns effectively.

P-ISSN: 3050-9572 E-ISSN: 3050-9580

EUROPEAN DATA SCIENCE JOURNAL

Volume: 03 Issue: 03

RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

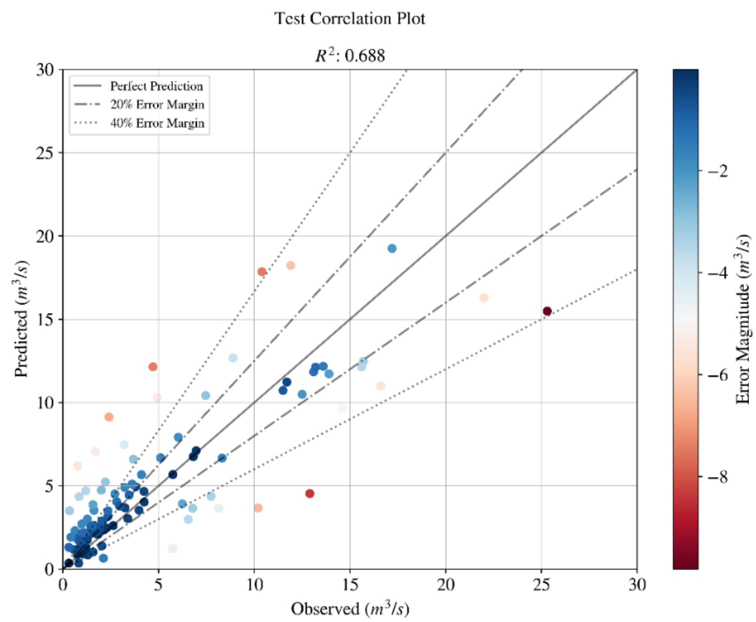

Figure 3. Scatterplot of observed and predicted Q for testing period with XGBoost

Figure 3 shows the scatter plot for the testing period (2005–2014), which reveals a consistently strong correlation between predicted and observed discharge values. Although slightly more dispersion is observed compared to the training set, the model maintains robust performance, indicating effective generalization.

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

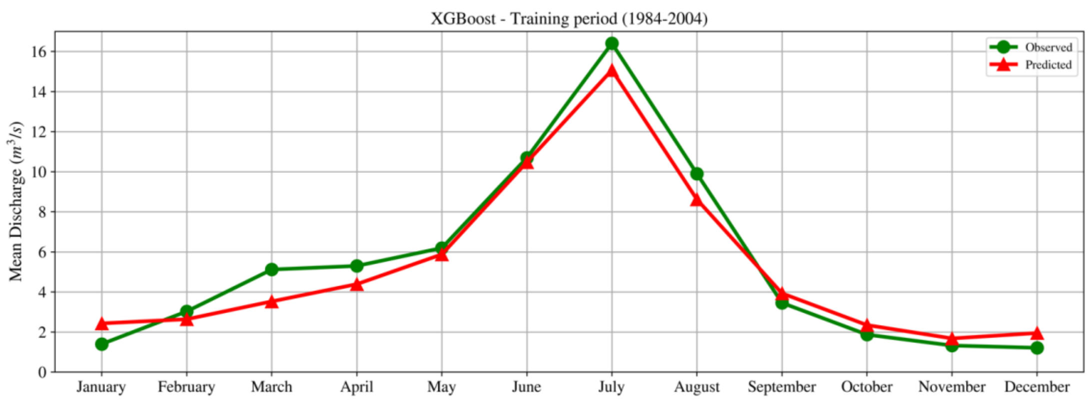

Figure 2. Comparison of the result of observed and predicted discharges for training period in XGBoost method

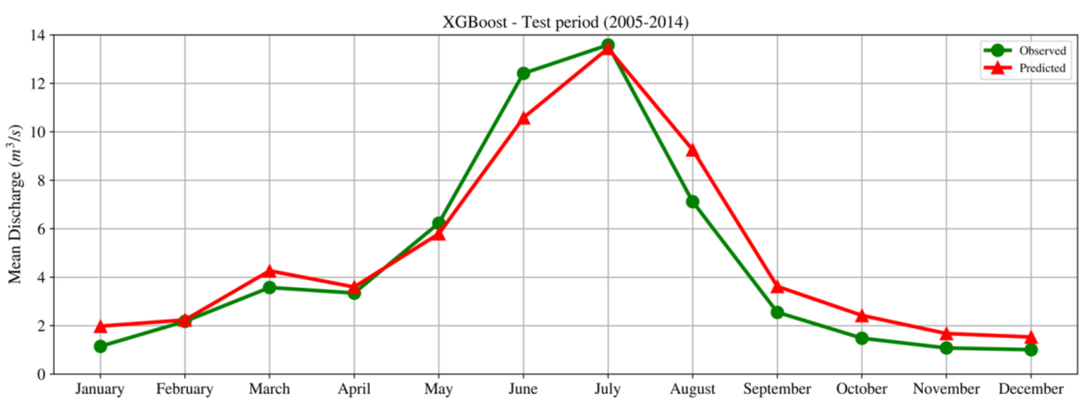

Figure 5. Comparison of the result of observed and predicted discharges for testing period in XGBoost method

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

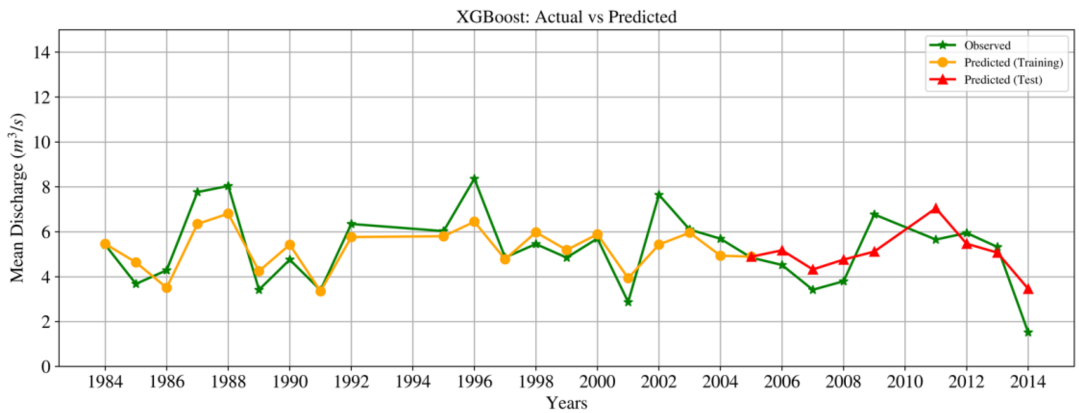

Figure 6. Comparison of the result of observed and predicted discharges for testing and training period in XGBoost method

Figures 4 and 5 compare the observed and predicted monthly streamflow values for the training (1984–2004) and testing (2005–2014) periods. The model accurately captures seasonal variations, particularly the peak discharges observed in summer months such as June and July. The slight dispersion observed during the testing period indicates acceptable generalization.

Figure 6 illustrates the temporal trend of observed and predicted streamflow over the entire period. The XGBoost model successfully replicates long-term streamflow variability and captures key seasonal dynamics.

3.2 Feature Contribution and SHAP Interpretation

The SHAP (SHapley Additive exPlanations) analysis was utilized to interpret the contribution of each input variable to the final prediction of streamflow. According to the results, variables such as mean sea level pressure (mslp), specific humidity (shum), and precipitation emerged as the most influential features impacting model

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

output. These findings are consistent with previous research by Lundberg and Lee (2017) [16], who highlighted SHAP's effectiveness in improving model interpretability and transparency.

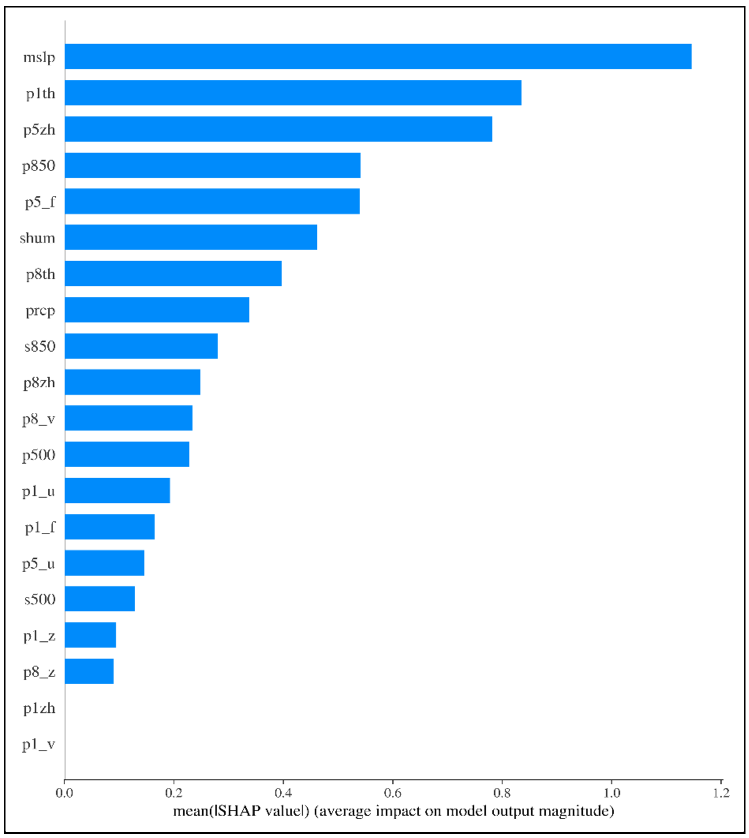

Figure 7. SHAP values avaerage impact on model output magnitude in XGBoost

Figure 7 displays the mean SHAP values, revealing the average impact of each feature on the XGBoost model output. The higher the SHAP value, the more significant the contribution of that variable to streamflow predictions.

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

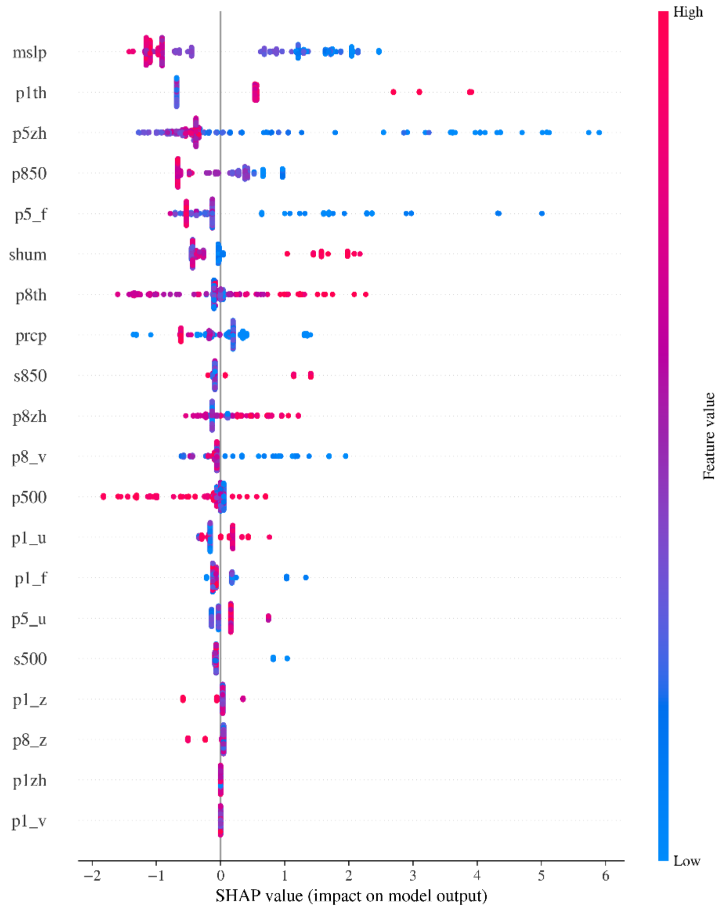

Figure 8. Impact results with SHAP value for XGBoost

Figure 8 illustrates the direction and magnitude of SHAP values for key features. Positive values (rightward direction) indicate a positive contribution to discharge prediction, while negative values (leftward direction) suggest a suppressing effect. The color gradient represents the magnitude of each feature's value, providing deeper insight into their influence.

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

3.3 Future Projections under SSP Scenarios

Using CMIP6 climate data, the XGBoost model projects streamflow under four Shared Socioeconomic Pathway (SSP) scenarios: SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5.The results (figures 9-17) show declining trends in discharge with higher emissions scenarios. These projections reveal significant seasonal and interannual variations, indicating the need for adaptive water resource management strategies under climate uncertainty.

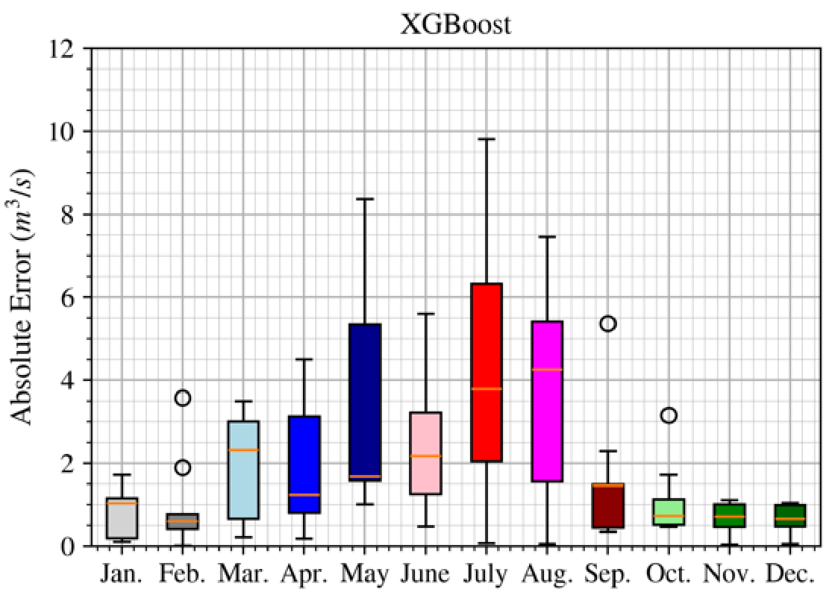

Figure 9. Absolute (RMSE) and Mean Absolute Error Results for test period with XGBoost

The performance of XGBoost was evaluated using standard metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash-Sutcliffe Efficiency (NSE). The results demonstrate that the XGBoost algorithm outperformed traditional methods in predictive accuracy across all SSP scenarios, especially under SSP5-8.5. This finding aligns with previous studies that emphasize the robust predictive capabilities of gradient boosting algorithms [17].

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

Monthly error analysis reveals that XGBoost outperformed other models in the 2nd, 6th, 8th, 9th, and 10th months, as shown by lower RMSE and MAE values. This demonstrates XGBoost's superior reliability in simulating seasonal streamflow variability [17].

Overall, the results confirm that XGBoost is an effective tool for climate-based streamflow prediction, providing high accuracy and explainability. These results are in line with existing literature, confirming the strong performance of machine learning models in hydrological forecasting under varying climate scenarios [14].

3.4 Future Prediction Results of XGBoost

The monthly mean discharge projections obtained from the XGBoost model under the SSP1-2.6 scenario are presented in Figure 10. The model accurately reflects the seasonal discharge dynamics, showing increased values during summer months particularly in July and decreased values in December. This seasonal pattern remains consistent across all projection periods from 2015 to 2100.

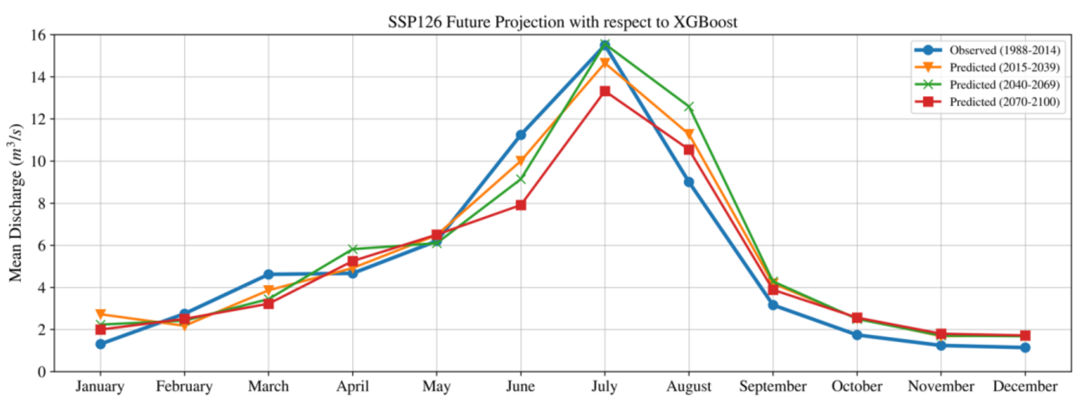

Figure 10. Line graph of observed and predicted mean discharge with default model parameters under SSP1-2.6 scenario with XGBoost

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

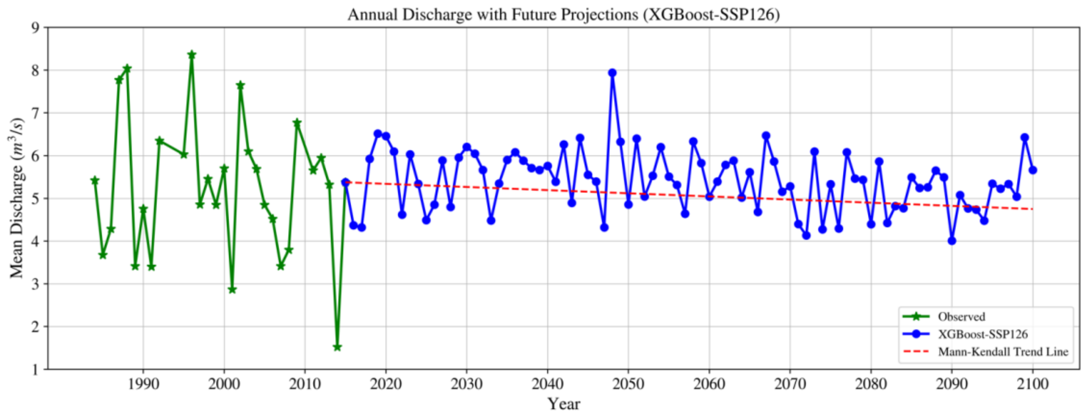

Figure 11. The projection of annual discharge (Q) values under the SSP1-2.6 scenario by XGBoost model and observed data between 1984-2014 in normal scale with a trendline

In Figure 11, the annual discharge projections are compared with the observed data from 1984 to 2014. The Mann-Kendall trend analysis reveals a weak and statistically insignificant trend throughout the projection period, suggesting minimal long-term changes in streamflow under the low-emission scenario of SSP1-2.6.

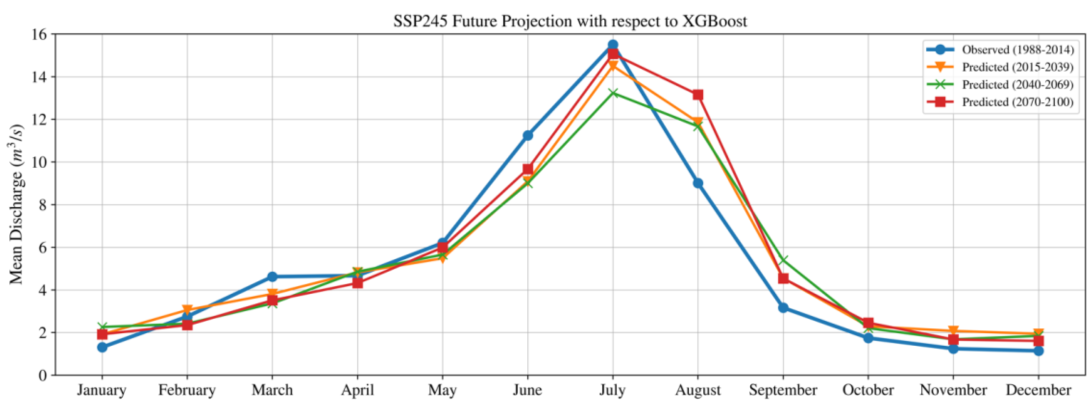

Figure 12. Line graph of observed and predicted mean discharge with default model parameters under SSP2-4.5 scenario with XGBoost

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

Figure 12 illustrates the comparison between observed and predicted mean monthly discharge values under the SSP2-4.5 scenario using the XGBoost model. The projections clearly exhibit seasonal discharge behavior, with peak values occurring in July and significant decreases observed in December. This seasonal pattern remains consistent across all future periods , indicating that the model effectively captures intra-annual variability under moderate emission scenarios.

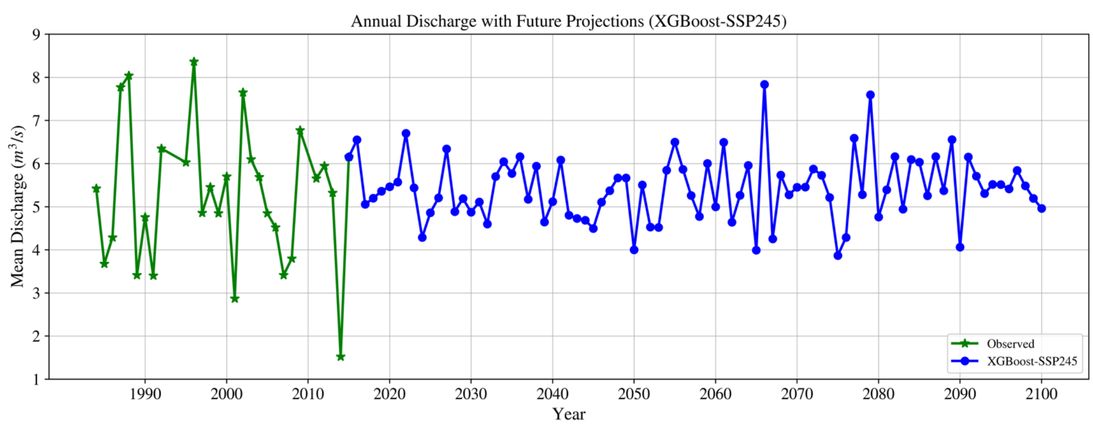

Figure 13. The projection of annual discharge (Q) values under the SSP2-4.5 scenario by XGBoost model and observed data between 1984-2014 in normal scale with a trendline

Figure 13 presents the annual discharge projections from 2015 to 2100 under the SSP2-4.5 scenario, alongside the historical observations from 1984 to 2014. While short-term fluctuations are visible, the Mann-Kendall trend analysis indicates no statistically significant long-term trend in discharge. This suggests a relatively stable annual discharge regime under SSP2-4.5, despite interannual variability.

Figure 14 illustrates the projected monthly mean discharge under the SSP3-7.0 scenario using the XGBoost model. A consistent seasonal pattern is observed, with peak discharge values in July and the lowest

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

values in December. However, compared to other scenarios, the monthly discharge shows higher interannual variability, especially during the summer months.

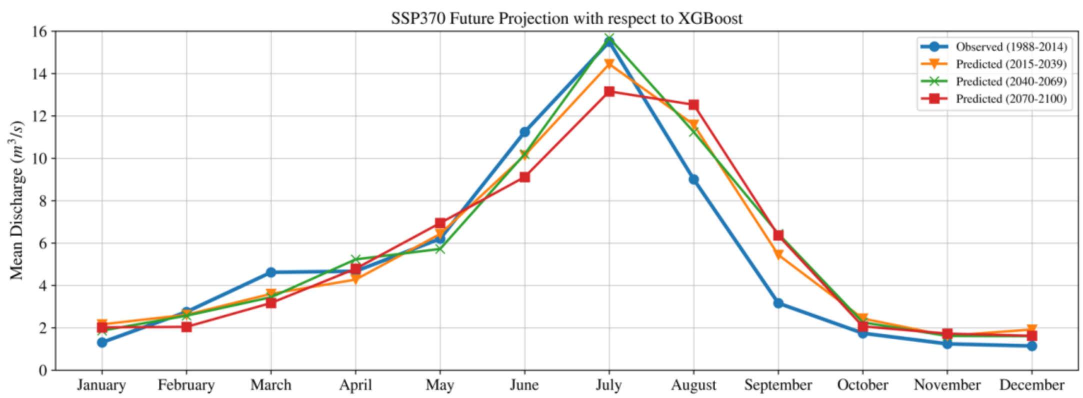

Figure 14. Line graph of observed and predicted mean discharge with default model parameters under SSP3-7.0 scenario with XGBoost

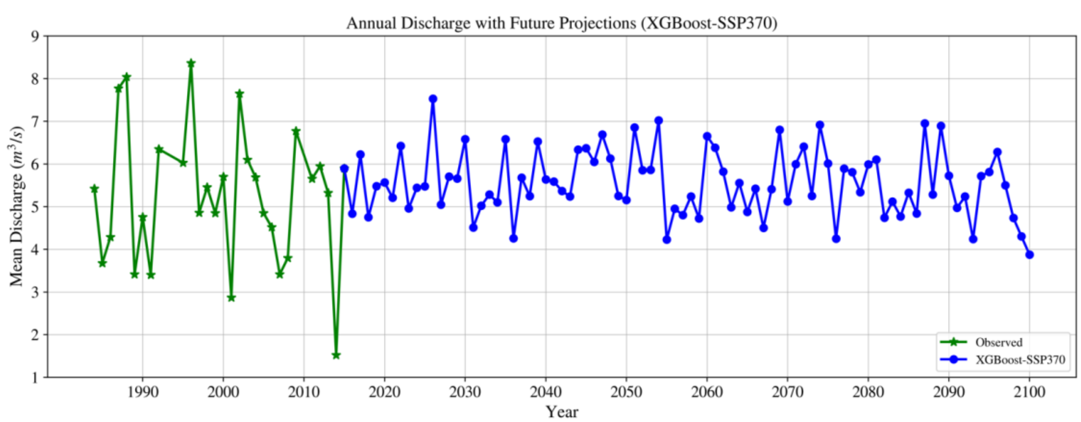

Figure 15. The projection of annual discharge (Q) values under the SSP3-7.0 scenario by XGBoost model and observed data between 1984-2014 in normal scale with a trendline

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

Figure 15 presents the annual discharge projections from 2015 to 2100, along with the observed values from 1984 to 2014. Despite visible year-to-year fluctuations, the Mann-Kendall trend analysis reveals no statistically significant long-term trend under the SSP3-7.0 scenario, indicating limited directional change in annual streamflow over time.

Figure 16 illustrates the projected monthly mean discharge under the SSP5-8.5 scenario using the XGBoost model. Compared to other scenarios, more pronounced variations are observed, especially during the summer period. July continues to exhibit the highest discharge rates, indicating intensified runoff events. Despite these fluctuations, the seasonal discharge pattern remains visible across future projection periods.

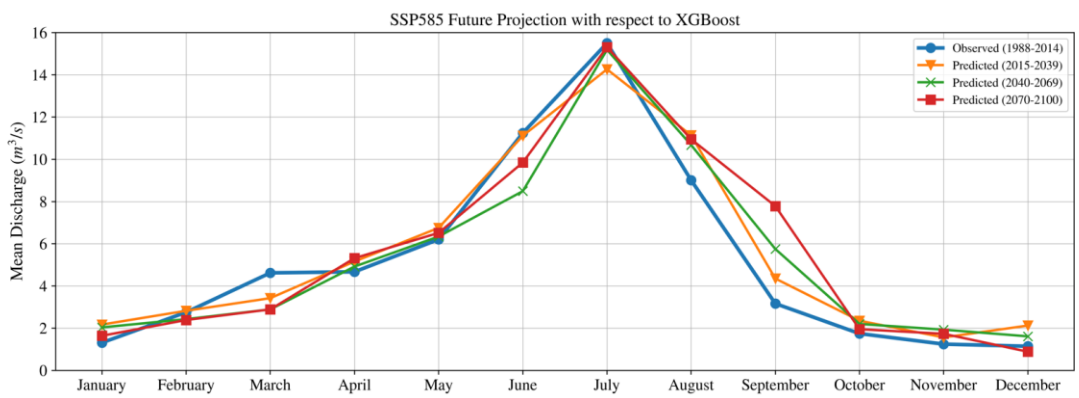

Figure 16. Line graph of observed and predicted mean discharge with default model parameters under SSP5-8.5 scenario with XGBoost

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

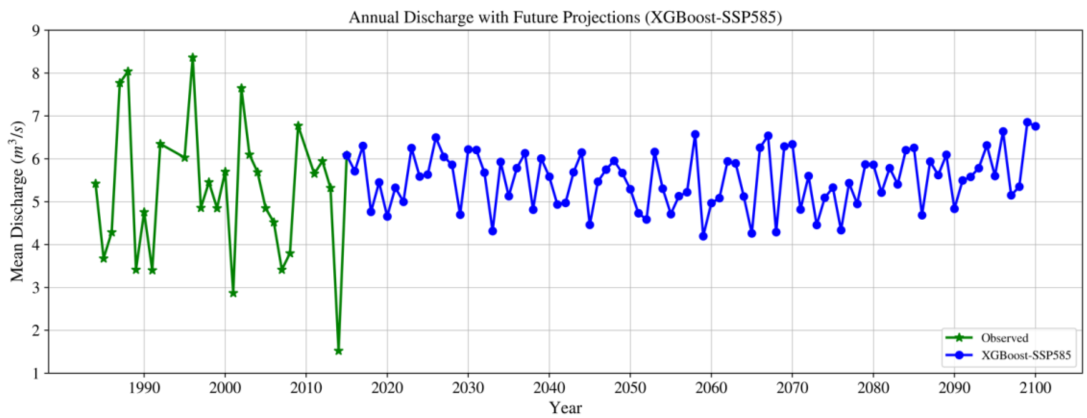

Figure 17. The projection of annual discharge (Q) values under the SSP5-8.5 scenario by XGBoost model and observed data between 1984-2014 in normal scale with a trendline

Figure 17 presents the annual discharge (Q) projections from 2015 to 2100 under SSP5-8.5, along with observed discharge values from 1984 to 2014. While the projected values show visible variability, the Mann-Kendall trend test indicates no statistically significant long-term trend, suggesting that the annual discharge levels are not expected to change in a directional manner under this high-emission scenario.

Overall, the results emphasize the effectiveness of the XGBoost algorithm in modeling streamflow under climate change scenarios, particularly in capturing seasonal variations. Moreover, the integration of SHAP analysis further improves model interpretability by identifying the key contributing variables to discharge predictions. These findings highlight the model's potential applicability in adaptive water resource management under future uncertainties [7].

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

4 CONCLUSION

This study comprehensively evaluated the performance of the XGBoost algorithm in forecasting streamflow within the Ceyhan River Basin under various CMIP6-based climate change scenarios. By integrating observed discharge data with projected meteorological variables, the model demonstrated strong predictive capability, especially in capturing seasonal hydrological dynamics.

Among all scenarios, the SSP5–8.5 pathway exhibited the most pronounced variability, indicating the model's sensitivity to extreme emission trajectories. July consistently displayed elevated discharge values, while December maintained a decreasing pattern across projections. Key climate drivers—such as mean sea level pressure (MSLP), precipitation (PRCP), and temperature (TEMP)—were identified as significant influencers, consistent with hydrological theory and prior findings [1,6].

Furthermore, the inclusion of SHAP analysis provided explainable insights by identifying dominant variables influencing discharge fluctuations. This not only enhanced the interpretability of the results but also strengthened the model's applicability in adaptive water resource management under climate uncertainty [2].

XGBoost's scalability and computational efficiency, coupled with its ability to model nonlinear relationships, highlight its suitability for long-term hydrological forecasting. The model's flexibility makes it particularly valuable for data-limited basins and evolving climate contexts. Future work should incorporate additional hydroclimatic variables—such as soil moisture and evapotranspiration—and explore ensemble-based comparisons to generalize findings across diverse basin types [7].

In conclusion, the combination of XGBoost and explainable AI methods offers a powerful, transparent, and adaptable solution for climate-resilient streamflow prediction. This framework holds significant promise for

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

guiding sustainable water resource planning and informing decision-making under evolving environmental conditions.

REFERENCES

1. Klemeš, V. (1986). Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1), 13–24.
2. Seneviratne, S. I., et al. (2012). Changes in climate extremes and their impacts on the natural physical environment. IPCC Special Report on Extremes.
3. Guven, A., & Şebcioğlu, Ş. (2019). Climate change impact on hydropower: A review. Journal of Hydrology Research and Development, vol. 35, no. 2, pp. 123-135.
4. Nagabhyru, K. C. (2022). Bridging Traditional ETL Pipelines with AI Enhanced Data Workflows: Foundations of Intelligent Automation in Data Engineering. Available at SSRN 5505199.
5. O'Neill, B. C., et al., "The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6," Geoscientific Model Development, vol. 9, pp. 3461–3482, 2016.
6. Eyring, V., et al., "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," Geoscientific Model Development, vol. 9, pp. 1937–1958, 2016.

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

7. Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning. Hydrology and Earth System Sciences, 23(12), 5089–5110.

8. Kolla, S. K. (2024). Clinical Knowledge Intelligence through Deep Learning and Natural Language Understanding in Healthcare Platforms. International Journal of Advanced Engineering Science and Information Technology (IJAESIT), 7(3), 14088.

9. Chen, T., and Guestrin, C., "XGBoost: A scalable tree boosting system," Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.

10. Feng, D., Wang, Y., Hao, Z., & Hao, F. (2020). Deep learning and traditional machine learning models for groundwater level prediction: A comparative study. Journal of Hydrology, 582, 124541.

11. Feng, D., Xu, Z., & Duan, Z. (2021). Application of machine learning techniques in hydrological modeling under CMIP6 scenarios: A review. Environmental Research Letters, 16(12), 124001.

12. DSI (State Hydraulic Works). (2014). Hydrological Data of Turkish River Basins. General Directorate of State Hydraulic Works, Ankara, Turkey.

13. Reddy Segireddy, A. (2024). Federated Cloud Approaches for Multi-Regional Payment Messaging Systems. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(2), 442-450.

14. DSI (State Hydraulic Works). (2015). Menzelet Dam Streamflow Records. General Directorate of State Hydraulic Works, Ankara, Turkey.

15. Canadian Climate Data and Scenarios (CCDS), Government of Canada. Available at: https://climate-scenarios.canada.ca/?page=pred-cmip6#cmip6-predictors

16. Meinshausen, M., et al. (2020). The SSP greenhouse gas concentrations and their extensions to 2500. Geoscientific Model Development.

P-ISSN: 3050-9572 E-ISSN: 3050-9580

EUROPEAN DATA SCIENCE JOURNAL

Volume: 03 Issue: 03

RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

17. Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937–1958.

18. Reddy, V. A. R. (2024). Generative Intelligence for Healthcare Claims Processing and Personalized Benefits Management. International Journal of Advanced Engineering Science and Information Technology (IJAESIT), 7(3), 14099.

19. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.

20. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777).

21. Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models: Part I—A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.

22. Hussain, A. Ansari, M. Danish, M and Azamathulla, H.M. (2025). Experimental and Machine Learning Approaches for Discharge Estimation in Arched Labyrinth Weirs, European Advanced Journal for Emerging Technologies (EAJET)-p-ISSN 3050-9734.

23. Gottimukkala, V. R. R. (2021). Digital Signal Processing Challenges in Financial Messaging Systems: Case Studies in High-Volume SWIFT Flows.

24. U Hansamali, RK Makumbura, U Rathnayake, HM Azamathulla, N Muttil

25. (2025).Leaky Dams as Nature-Based Solutions in Flood Management Part I: Introduction and Comparative Efficacy with Conventional Flood Control Infrastructure, Hydrology 12 (4), 95

26. U Hansamali, RK Makumbura, U Rathnayake, HM Azamathulla, N Muttil

27. Garapati, R. S. (2022). Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning. Current Research in Public Health, 2, 1346.

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 03
RECEIVED: MAY 25 REVISED: JUNE 28 ACCEPTED: JULY 08 PUBLISHED: JULY 26

28. Leaky Dams as Nature-Based Solutions in Flood Management Part II: Mechanisms, Effectiveness, Environmental Impacts, Technical Challenges, and Emerging Trends. Hydrology (2306-5338) 12 (4).

29. B Chaplot (2021). Prediction of rainfall time series using soft computing techniques

30. Environmental Monitoring and Assessment 193 (11), 721