Earth Systems and Environment
https://doi.org/10.1007/s41748-025-00858-x

ORIGINAL ARTICLE

Modeling Daily River Discharge Using Machine Learning Ensembles in the Context of Climate Change: Application To the zhaiyk-caspian basin, Kazakhstan

Sayat Alimkulov[1] · Lyazzat Makhmudova[1] · Bekzat Satenova[1,2] · Aisulu Tursunova[1] · Lyazzat Birimbayeva[1,2] · Elmira Talipova[1,2] · Darkhan Abdibekov[1] · Zhanibek Smagulov[1] · Oirat Alzhanov[1]

Received: 30 May 2025 / Revised: 18 September 2025 / Accepted: 24 September 2025
© The Author(s) 2025

Abstract

The study presents a comparative assessment of eight machine learning (ML) algorithms - Random Forest (RF), Lasso Regression (LASSO), AdaBoost (ADB), Gradient Boosting Regressor (GBR), Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), Light Gradient Boosting Machine (LGBM), and K-Nearest Neighbors (KNN) - for modeling daily river discharge at ten hydrological stations within the Zhaiyk - Caspian water management basin. Model performance was evaluated using mean absolute error (MAE), mean squared error (MSE), and symmetric mean absolute percentage error (SMAPE). The highest predictive accuracy (MAE≈0.3) was achieved by ensemble tree-based methods (Random Forest, CatBoost, Gradient Boosting, LightGBM, XGBoost), while LASSO and AdaBoost exhibited the weakest performance (MAE≈22). Identifying the most significant predictors enhanced both model interpretability and forecasting quality. The findings highlight the importance of tailoring ML approaches to the specific characteristics of river basins and suggest promising prospects for their integration with physically based hydrological models to improve river discharge forecasting and strengthen water resources management under climate change conditions.

Graphical Abstract

The Graphical Abstract Presents an Ensemble ML Framework for Modeling Daily River Discharge Under Climate Variability in the Zhayik–Caspian basin, Kazakhstan. Input Data Include Observed Daily runoff, Air temperature, and precipitation, Applied across Eight Machine Learning Algorithms: Random Forest, CatBoost, LightGBM, Gradient Boosting, KNN, LASSO, AdaBoost, and XGBoost.

The graphical summary illustrates: the study area, the structure of the modeling framework, key predictors influencing runoff, and a comparative analysis of observed and simulated discharge. The results highlight the potential of ensemble machine learning methods for improving runoff prediction in arid regions and their application in water resource planning, flood and drought risk management, and climate change adaptation strategies.

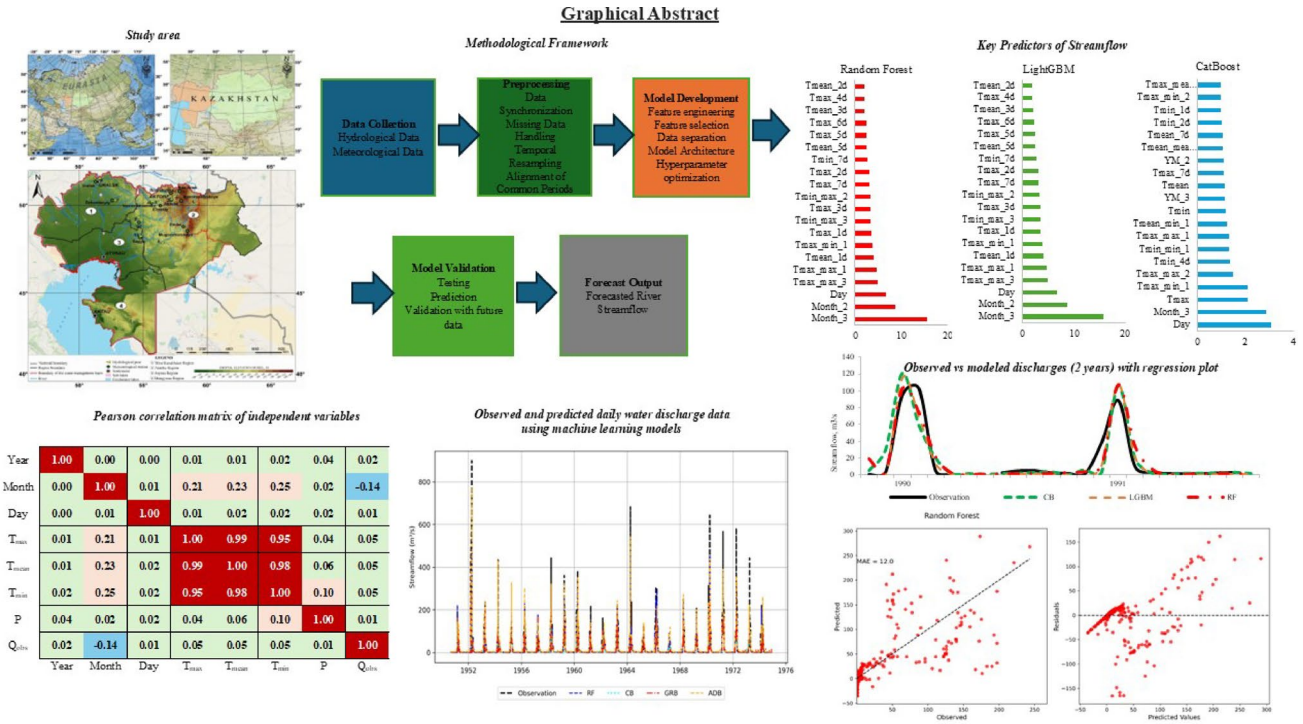
✉ Lyazzat Makhmudova
mlk2002@mail.ru

✉ Elmira Talipova
elmira_280386@mail.ru

[1] JSC Institute of Geography and Water Security, Almaty, Kazakhstan

[2] al-Farabi Kazakh National University, Almaty, Kazakhstan

Published online: 15 October 2025



Highlights

- This study evaluates eight machine learning models using daily discharge data from ten stations in the Zhaiyk–Caspian Basin.
- Tree-based ensemble methods, such as Random Forest and CatBoost, showed consistently strong performance.
- Simpler models like LASSO and AdaBoost struggled with the basin's nonlinear hydrological behavior.
- Feature selection techniques helped improve both model accuracy and interpretation of key predictors.
- The findings suggest that machine learning can enhance streamflow prediction under climate uncertainty.

Keywords Water resources · Climate · Hydrological regime · Modelling · Algorithm · Machine learning · Random forest · CatBoost · LightGBM · Gradient boosting · KNN · LASSO · AdaBoost · XGBoost

1 Introduction

In recent decades, climate change and its impacts on water resources have emerged as one of the most pressing global environmental challenges (IPCC 2022; Ciampittielloet al. 2024; Abbass et al. 2022). Water resources are a fundamental component of food security, ecosystem functioning, economic growth, and social well-being (Bangira 2019; Li and Wu 2024; Ringler et al. 2022). However, the increasing variability of climatic conditions, accompanied by more frequent extreme hydrometeorological events - including droughts, floods, and shifts in the seasonal distribution of precipitation - has led to substantial transformations in both the magnitude and dynamics of water resources (Matano et al. 2025; Birimbayeva et al. 2024; Alimkulov et al. 2024a, b, 2025; Nurbatsina et al. 2025; Davletgaliev et al. 2020).

In such conditions, there is a growing need for methods that ensure accurate and timely assessment and forecasting of river flow. Modern approaches to hydrological modelling include both physical and mathematical methods based on the description of runoff formation processes and ML algorithms, which demonstrate high efficiency in processing large amounts of data and identifying complex nonlinear relationships in hydroclimatic systems. At the same time, given the underdevelopment of the hydrometeorological monitoring system and existing limitations related to the availability of hydrometeorological data, the use of such methods becomes a particularly urgent task for ensuring accurate forecasting of water resources (Dong et al. 2025; Sheikh and Coulibaly 2024; Zhao et al. 2024; Singh et al. 2023; Rozos 2023; Lu et al. 2021).

Since the first studies using recurrent neural networks for hydrological problems in the 1990s, interest in the

application of ML methods in this field has grown significantly. Recent reviews emphasise that ML opens up new opportunities for analyzing complex hydrological processes, improving forecasting accuracy, and integrating heterogeneous data sources (Nearing et al. 2021; Xu and Liang 2021). This study used modern machine learning algorithms, including RF, LASSO, AdaBoost ADB, GBR, XGBoost, CatBoost, LGBM and KNN. These methods are highly capable of identifying complex nonlinear relationships and processing large amounts of data, making them promising for solving hydrological modelling problems in conditions of climate variability (Maity et al. 2024). The Random Forest algorithm has demonstrated high efficiency in predicting daily runoff under various hydrological regimes, outperforming traditional regression approaches (Pham et al. 2021). Gradient boosting methods, including XGBoost, have proven their ability to accurately model complex hydrological dependencies (Niazkar 2024). CatBoost, LightGBM, and KNN algorithms have also shown high performance in predicting river runoff, especially in basins with complex climatic and topographic conditions (Kumar et al. 2023). The growing interest in these models and their successful application in hydrology are analyzed in detail in a review article by 2024 ML methods allow not only to model current processes, but also to identify hidden dependencies, making it possible to more accurately predict changes in water resources under the influence of climatic factors.

Experience with machine learning in Central Asia and Kazakhstan confirms its high practical value for hydrological modelling. In Kyrgyzstan, as part of the SAPPHIRE project, modern deep learning models such as Temporal Fusion Transformer (TFT), Time-Series Mixer (TSMixer) and Time-series Dense Encoder (TiDE) were trained for short-term river flow forecasting based on data from 61 hydrological stations. The results showed high forecast accuracy, surpassing traditional methods (HydroSolutions.ch 2025).

In Kazakhstan, the application of ML and deep learning (DL) in hydrometeorology is actively expanding to include drought and flood forecasting. Thus, Sadrtdinova et al. (2024) showed that the Gated Recurrent Unit (GRU), Random Forest and XGBoost models provide SPI/SPEI index forecasts for up to six months with an accuracy of 99% (Sadrtdinova et al. 2024). In a study by Ziyatbekova et al. (2025) developed an intelligent flood forecasting system that combines Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (LST) satellite indices and ground data using ensemble and neural models, achieving an accuracy of up to 94% (Ziyatbekova et al. 2025). Kabdrakhova et al. (2024) demonstrated the effectiveness of ML and DL with oversampling techniques in forecasting floods on the Malaya Almatinka River (Kabdrakhova et al. 2024). Experience with ML in Kazakhstan confirms their high practical value for hydrological modelling and the development of early warning systems.

Despite significant achievements in the application of ML methods for forecasting droughts and floods in Kazakhstan, research focused at their use in the region under consideration is still insufficient, which emphasizes the need for further development and adaptation of modern approaches, taking into account the specific hydrometeorological and climatic characteristics of the region. The hydrological regime of rivers in the West Kazakhstan region is characterized by significant fluctuations caused by both natural climatic factors and anthropogenic influences. The interplay of precipitation, air temperature, other hydrometeorological parameters, and land-use changes creates complex nonlinear relationships that challenge the accuracy of streamflow prediction using traditional methods. These methods and approaches allow for the integration of diverse data sources, taking into account spatial and temporal variability and identifying hidden patterns in river flow dynamics. Their use significantly improves the accuracy of forecasts, providing a more reliable basis for water resource management, flood and drought prevention, and the development of effective strategies for adapting to changing climatic conditions in the region.

The novelty of this study lies in the comprehensive assessment of the effectiveness of various ML algorithms for modelling river flow using daily hydrometeorological data, which has not been done before for the river basins of Western Kazakhstan. The relevance of the work is determined by the need to improve the accuracy of water resource assessment in the context of increasing climate variability and limited observation networks. The main focus is on analysing seasonal changes in river flow, the impact of changes in air temperature and precipitation, and their relationship with the risk of extreme events (droughts, floods). The application of machine learning algorithms opens up new opportunities for improving the accuracy of long-term water resource forecasts and developing adaptation strategies in conditions of climate uncertainty. The results obtained will provide an understanding of the relationship between climate and water resources in order to create a scientific basis for sustainable water resource management strategies.

2 Study Area and Data Used

2.1 Study Area

In the scientific study, the object of research is the Zhaiyk-Caspian water management basin, which occupies most of

the territory of Western Kazakhstan. The basin covers the territory of four regions of Kazakhstan: West Kazakhstan, Atyrau, Mangystau regions and most of the territory of Aktobe region (Fig. 1).

Extending from the eastern edge of the Volga River delta in the west to the Turan Lowland in the southeast and from the southern foothills of the Ural Mountains and the Obschiy Syrt to the Ustirt Plateau, the region covers an area of 736,241 km², stretching about 1,200 km from north to south and 1,300 km from west to east. The Mugalzhar Ridge, rising 200–300 m above the Zhaiyk-Zhem Plateau, is the watershed between Western and Central Kazakhstan and is the source of the region's main waterways: Oiyl, Zhem, Elek, Or, etc. The relief is characterized by rolling plains, dunes, salt marshes and drainless depressions. The Caspian Lowland, stretching along the northern coast of the Caspian Sea, is of particular geographical importance, with its characteristic flat surface, numerous lakes and salt marshes (Tokbergenova et al. 2025; Sarsekov et al. 2017; Akiyanova et al. 2014).

There are more than 1,500 rivers and temporary watercourses over 10 km long in the region, with a total length of more than 45,000 km. The main surface water sources in the basin are the north-eastern and eastern coasts of the Caspian Sea, the rivers Zhaiyk, Oiyl, Zhem, Sagyz, Elek, Or, Kobda, Shagan, Karaozen, Saryozen, the delta arms of the Volga-Kigach, Sharanovka and other small rivers.

The climate of the western regions of Kazakhstan (West Kazakhstan, Aktobe region, Mangistau and Atyrau) is characterized by a pronounced continental climate, manifested in sharp seasonal and daily temperature fluctuations, a rapid transition from winter to summer, general aridity and a lack of precipitation. The region is characterised by cold and relatively short winters, as well as long and hot summers.



Fig. 1 Overview map of the study area and location of hydrological posts and meteorological stations of the Zhaiyk-Caspian water management basin

The eastern regions have harsher winter conditions, while the western territories have a slightly milder climate due to the influence of the Mugodzhar Mountains and the influx of warm air masses from Central Asia. In the Mangistau and Atyrau regions, the climate is influenced by Arctic, Turanian and Iranian air masses, which make it extremely arid and sharply continental. The influence of the Caspian and Aral Seas is limited to a narrow coastal zone, where there is a slight increase in air humidity and a smoothing of temperature amplitudes, both seasonally and daily (Kazhydromet 2025).

The area under consideration is not only the largest in terms of territory, but also one of the most important regions in economic terms, thanks to its powerful oil industry. Industry in the region is an important factor in economic development and is characterized by high demand for high-quality water resources. Since the early 2000s, the region has seen intensive growth in oil and gas production, which has increased pressure on the water balance. At the same time, the industrial sector is also vulnerable to water shortages, although its vulnerability is somewhat lower than that of irrigated agriculture (Birimbayeva et al. 2025).

2.2 Data Used

This study uses daily data of instrumental observations of river runoff (water discharge) at 9 hydrological posts (HP), maximum, mean and minimum air temperature and the sum of daily precipitation at 12 meteorological stations (MS) within the study area. The information series cover the whole period of observations, at the operating hydrological posts up to 2022 inclusive. Data were obtained from the official resource of RSE 'Kazhydromet', presented on the websites: https://www.kazhydromet.kz/ru/gidrologiya/ezhegodnye-dannye-o-rezhime-i-resursah-poverhnostnyh-vod-sushi-eds and http://ecodata.kz:3838/dm_climat_ru/.

Table 1 List of hydrological posts and meteorological stations of the Zhaiyk-Caspian water management basin and their calculation periods

№	Hydrological post	Meteorological station	Calculation period
1	Or – Bogetsay	Novorossiyskoye	1958–2022
2	Elek – Aktobe	Aktobe Rodnikovka	1946–1974
3	Kargaly – Kargaly	Rodnikovka Kos-Istek	1957–1974
4	Ulken Kobda – Kobda	Il'insky Novoalekseevka	1962–2022
5	Shagan – Kamenny	Uralsk	1952–2010
6	Olenty – Zhympity	Zhambeyty	1964–2022
7	Zhem – Zharkamys	Mugodzharskaya Emba	1950–1991
8	Oiyl – Taltogay	Uil	1936–1980
9	Sagyz – Sagyz	Sagyz	1951–1992

The list of hydrological posts and meteorological stations whose data were used in the study and the calculation period are presented in Table 1.

3 Methodology

3.1 Input Datasets

The study utilized long-term hydro-meteorological data collected from a number of hydrological posts and nearby meteorological stations within the river basin. The hydrological dataset consisted of daily river discharge values (Q), while the meteorological dataset included air temperature indicators (Tmax, Tmean, Tmin) and precipitation (P). The observation period varied depending on data availability at each station and spanned from the mid-1930s to 2022. For each hydrological post, the corresponding meteorological stations were selected based on geographical proximity and expert knowledge of the basin's hydrological characteristics.

To ensure the reliability of model evaluation, the original dataset was divided into training and independent test sets. In this study, the last two years of observations were reserved for testing, while the remaining period was used for model training and cross-validation (Fig. 2).

3.2 Preprocessing

MissingDataHandling: Meteorological and hydrological datasets contained missing observations across several stations. Gaps were imputed using the RF algorithm, which accounts for nonlinear dependencies and interactions among predictors. This approach was chosen for its robustness against outliers and its proven accuracy in reconstructing missing climate records.

To capture temporal and seasonal dynamics of river flow, the following features were generated:

- Aggregated variables: monthly and seasonal means, minima, and maxima of temperature (Tmax, Tmean, Tmin) and precipitation.
- Lagged variables: monthly lags up to three months and daily lags up to seven days to incorporate short-term memory effects.
- Seasonality indicator: a categorical feature representing four climatological seasons (winter, spring, summer, autumn).
- This preprocessing ensured the dataset represented both short-term variability and long-term climatic cycles.

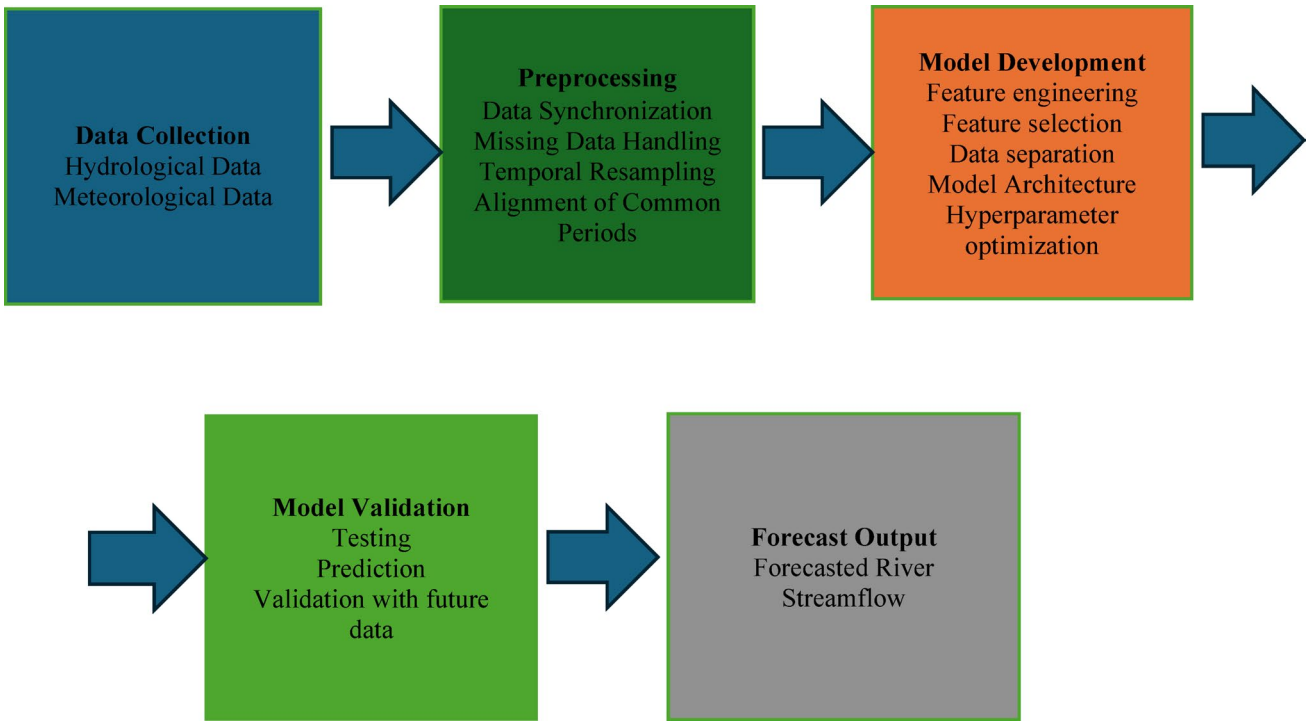


Fig. 2 Methodological framework

3.3 Machine Learning Models (the Section Was shortened)

In this study, we evaluated eight ML algorithms representing different methodological families that are widely applied in hydrological forecasting: RF, LASSO, ADB, GBR, XGBoost, CatBoost, LGBM and KNN (Table 2). The selection was motivated by the proven efficiency of ensemble decision-tree approaches in capturing nonlinear hydrological processes, as well as the interpretability of linear models and the flexibility of instance-based methods.

3.4 Hyperparameter Optimization

Hyperparameter optimization for each model was performed using the Bayesian Optimization Algorithm (BOA) (Pelikan 2005; Wu et al. 2019) combined with 5-fold cross-validation for automated parameter tuning.

Bayesian optimization is based on the Bayesian theorem, where the acquisition function is maximized to select the next most promising evaluation point x_i Then, the value of the target function y_i, is computed, and the new pair (x_i, y_i) is added to the set of known evaluations. Through iterative updating of the probabilistic representation model allows the optinal solution to be found efficiently. Due to its ability to efficiently leverage historical information, BOA significantly outperforms other optimization methods in terms of efficiency.

The base machine learning models and their corresponding hyperparameter grids are presented in Table 3.

3.5 Model Evaluation

Eight machine learning models were tested: Random Forest, Lasso, AdaBoost, Gradient Boosting, XGBoost, CatBoost, LightGBM, and KNN. Model accuracy was evaluated on the test set using the following metrics:

- Mean Absolute Error (MAE).
- Mean Squared Error (MSE).
- Symmetric Mean Absolute Percentage Error (SMAPE).
- Coefficient of determination (R^2).

In this study, the performance of the models was evaluated using the MAE, MSE, and SMAPE. The performance metrics were calculated using the following formulas:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{1}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 \tag{2}$$

$$\mathrm{SMAPE} = \frac{100\%}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{|\mathrm{y_i} - \widehat{\mathrm{y}}_\mathrm{i}|}{(|\mathrm{y_i}| + |\widehat{\mathrm{y}}_\mathrm{i}|)/2} \tag{3}$$

Table 2 List of machine learning models applied in the scientific study

Model	Category	Reported Strengths	Reported Limitations	Key References
Random Forest (RF)	Ensemble (decision trees)	Handles nonlinear relationships; robust to noise; less prone to overfitting; effective with large feature sets	Computationally expensive with many trees; limited interpretability	Breiman 2001
K-Nearest Neighbors (KNN)	Instance-based	Simple; no training required; effective on small–medium datasets	High computational cost at prediction; sensitive to dimensionality	Cover and Hart 1967
CatBoost	Gradient boosting	Native handling of categorical variables; strong generalization; resistant to overfitting	High memory demand; sensitive to hyperparameters	Roth 2004
XGBoost	Gradient boosting	High predictive accuracy; efficient on large-scale data; handles missing values	Requires careful parameter tuning; risk of overfitting	Chen and Guestrin 2016
Lasso Regression (LASSO)	Linear regression	Performs feature selection; interpretable; effective in high-dimensional data	Struggles with nonlinearities; sensitive to outliers	Szczepanek 2022
LightGBM (LGBM)	Gradient boosting	Very fast and memory efficient; scalable to large datasets	Sensitive to parameter settings; potential instability	Ke et al. 2017
Gradient Boosting Regressor (GBR)	Gradient boosting	Strong predictive performance; accommodates various data types	Computationally expensive; limited interpretability	Xu et al. 2023
AdaBoost (ADB)	Boosting	Improves weak learners; relatively simple to implement; resistant to overfitting on small datasets	Sensitive to noisy data and outliers	Freund and Schapire 1997

where y_i – is the actual value; $\widehat{y}_i$ – is the predicted value; N – is the number of observations.

3.6 Model Interpretation

The Pearson correlation analysis (Fig. 3) demonstrates the relationships among the independent variables and highlights the climatic factors that contribute most significantly to the formation of water resources. In particular, air temperature and precipitation show the strongest associations with river flow, confirming their central role in shaping the region's water balance. The highest positive correlation with flow was observed for daily temperature data ($r=0.05$), while the lowest correlation values were recorded for aggregated monthly indicators ($r=-0.14$). Although the correlation coefficients are relatively low, they still reflect the hydrological relevance of these variables. It should be noted that Pearson's coefficient captures only linear dependencies and does not fully represent more complex, nonlinear relationships that may influence river discharge.

To account for such complexities and to improve feature selection, several machine learning methods were applied, including CatBoost, XGBoost, GBR, LightGBM, and RF. These algorithms made it possible to identify the most significant variables by evaluating their contribution to the model's predictions. This ensured consideration of both linear and non-linear dependencies, thereby increasing the accuracy of river flow forecasting.

Feature Importance (FI) in decision trees is estimated based on Impurity Reduction (IR), which allows for quantifying the contribution of each variable to the final model. One of the most common measures is the Gini Impurity (Menze et al. 2009), the reduction of which during node splitting indicates a high information content of the feature. Formally, the feature importance for X_i is calculated using the following formula:

$$\mathrm{FI}\left(\mathrm{X_i}\right) = \sum {}_{\epsilon(\mathrm{X_i})} \Delta\, \mathrm{I}\left(\mathrm{t}\right) \mathrm{P}\left(\mathrm{t}\right) \tag{4}$$

where: t –is the tree node; $\Delta I\,(t)$ – is the reduction in impurity at the node; $P\,(t)$ – is the fraction of objects trapped in this node.

Figure 4 presents the most significant variables for the Elek–Aktobe hydrological station. This analysis allows for the identification of key predictors and optimization of the input dataset by excluding irrelevant or redundant features, ultimately improving both forecast accuracy and the interpretability of the models.

3.7 Software

All experiments were conducted in Python 3.X using the following libraries: scikit-learn, RF, GBR, KNN, LASSO,

Table 3 Hyperparameter grid for each machine learning model

Name of machine learning model	Parameters	Optimal values of hyperparameters								
		Or-Bogetsay	Elek-Aktobe	Kargaly-Kargaly	Ulken Kobda-Kobda	Shagan-Kamenny	Olenty-Zhympity	Zhem-Zharka-mys	Oiyl-Tal-togay	Sagyz-Sagyz
Random Forest	N estimators	250	150	250	200	900	550	500	750	500
	Max depth	20	11	13	13	13	28	13	13	17
	Min samples split	12	16	5	18	12	15	12	12	9
KNN	N neighbores	39	40	40	10	19	10	50	3	10
	Weights	Uniform	Uniform	Uniform	Uniform	Uniform	Uniform	Uni-form	Uni-form	Uni-form
CatBoost	Iterations	1700	1400	2300	1300	1400	1000	1500	2300	1000
	Max depth	11	8	8	4	6	10	10	8	11
	Learning rate	0.17	0.02	0.09	0.01	0.6	0.01	0.05	0.09	0.03
XGBoost	N estimators	12	800	800	700	500	800	800	750	800
	Max depth	0.25	7	7	6	4	14	7	6	14
	Learning rate	4	0.02	0.02	0.014	0.03	0.02	0.02	0.06	0.02
	Min child weight	5	7	7	4	9	1	7	9	1
Lasso	Iterations	4300	1400	1400	2500	3500	2400	2900	1900	800
	Alpha	2.8	2.45	2.45	1.85	2.62	0.5	7.18	1.81	1.8
	Tol	0.11	0.08	0.008	0.05	0.07	0.07	0.03	0.01	0.07
Light GBM	N estimators	100	3000	3000	300	100	800	700	1250	700
	Max depth	9	6	6	9	9	7	9	8	13
	Learning rate	0.02	0.05	0.05	0.03	0.02	0.02	0.4	0.07	0.4
	Num leaves	92	50	50	269	92	110	197	189	289
Gradient Boosting	N estimators	260	210	210	900	800	240	700	210	690
	Max depth	4	10	10	3	5	6	4	10	5
	Learning rate	0.03	0.01	0.009	0.05	0.006	0.01	0.02	0.009	0.04
AdaBoost	Learning rate	0.008	0.04	0.04	0.007	0.02	0.007	0.002	0.01	0.03
	Max depth	9	9	9	9	5	7	7	9	10
	Min samples leaf	3	6	6	2	4	10	9	7	7
	Min samples split	8	5	5	10	11	20	10	18	4

ADB, XGBoost, LightGBM, CatBoost, SHAP (for interpretability), NumPy, and Pandas. Hyperparameter tuning was performed using Bayesian optimization via Optuna, which allowed to efficiently explore the search space and select the best set of parameters for each model.

4 Results

This section presents the performance of the applied machine learning models in river flow forecasting across the studied hydrological posts. Model accuracy was evaluated using the MAE, which allows a direct comparison of predictive capabilities under different basin conditions.



Fig. 3 Pearson correlation matrix of independent variables

Analysis of Table 4 showed that the MAE values for CB, LGBM, RF, and GRB were higher for the Ulken Kobda and Oiyl rivers. The best MAE was 8.6 for Elek - Aktobe HP RF model, followed by GRB with a score of 9.7, while the worst result was shown by the LASSO model with a score of 22.7.

In Ulken Kobda River basin, the best MAE was 0.58 for the GRB, followed by CB with a MAE of 0.6, and the worst result was demonstrated by the LASSO model with a result of 4.6.

For Shagan – Kamenny HP, the best MAE was 3.4 for the LGBM model, followed by the KNN model with a score of 4, while the ADB model showed the worst result, which was 4.6.

For Oiyl – Taltogay HP, the RF model demonstrated the best performance, with a MAE result of 10.3, followed by CB with a score of 10.4, and the worst result was 13 for the ADB model.

For Olenty – Zhympity HP, the performance of the KNN and GRB models (0.3) was the best, while ADB was 1.64. For Sagyz River, all models showed an average of 0.5.

All machine learning models synchronously tracked the dynamics of observed values in the annual runoff series based on historical instrumental data at hydrological stations, which confirms the selection and validity of the ML models. The simulated runoff hydrographs of the rivers under consideration, generated using eight machine learning models, are clearly presented in Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 for a more detailed and intuitive understanding of the forecasting effectiveness for each machine learning model. The figures show a comparison of the simulated and predicted daily runoff for the entire period using machine learning models for the rivers of the Zhaiyk-Caspian WMB.

Comparison of monthly average observed river discharge values for each river for the two-year test period with modeling results obtained using the most accurate machine learning algorithms is presented in Figs. 14, 15, 16, 17, 18, 19, 20, 21 and 22. Scatter plots showing the correspondence between observed and modeled data for the algorithms that showed the best results.

A comprehensive evaluation of the most effective machine learning models involves analyzing the predicted values in comparison with the observed values using scatter plots, where ideally the points should be located along a 45-degree line. The proximity of the points to the diagonal line of perfect agreement indicates a high degree of consistency between the actual and modeled values.

Visualization of the simulation results allows for an assessment of the selected models' ability to reproduce the seasonal dynamics of river flow, encompassing both high and low water regimes. During low flow periods, all models adequately capture the low discharge values, generally following the observed data's overall dynamics. However, more significant discrepancies between models are observed during flood periods.

For the Elek – Aktobe and Kargaly rivers, the RF model most accurately reproduces the peak flow values. Conversely, the GRB and CB models exhibit a time lag in reaching peak values for the Elek – Aktobe River, while for the Kargaly River, the GRB model noticeably overestimates values during peak flow periods.

In the second test year for the Elek – Aktobe River, the CB model significantly overestimates the peak discharge, whereas RF and GRB yield results closer to observed data, albeit with certain deviations. For the Sagyz, Zhem, and Oiyl rivers, the CB, GRB, RF, and LGBM models effectively capture flood peaks, particularly during rapid increases in water levels. The RF model often tends to overestimate peaks,

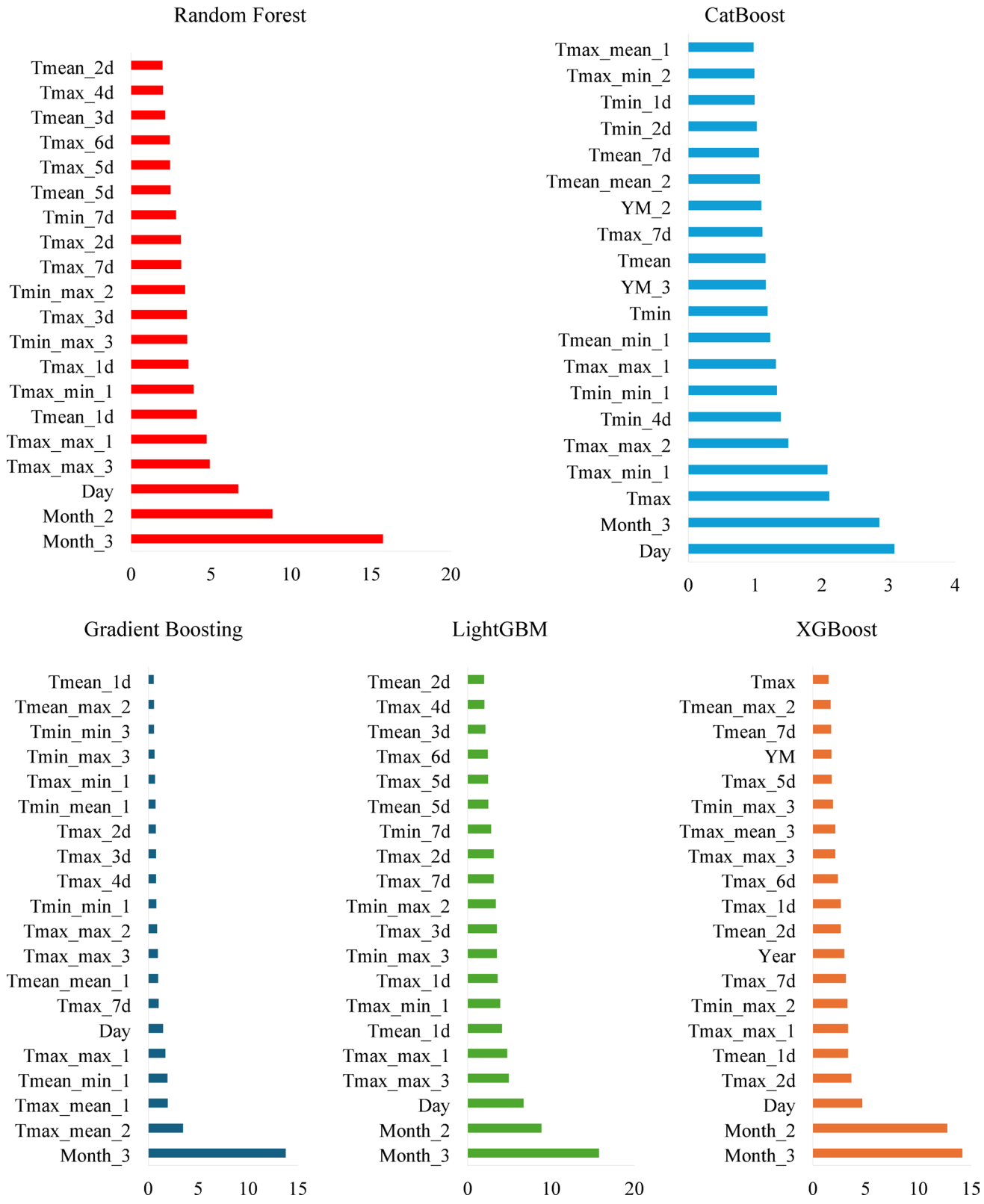


Fig. 4 Feature Importance plot for the hydrological post Elek – Aktobe

Table 4 ML metrics performance for nine basins

	RF	LASSO	CB	GRB	LGBM	XGB	KNN	ADB
Or – Bogetsay								
MAE	0.87	6.9	4.2	0.79	2.45	1.97	6.9	4.3
MSE	16	115	125	4	40	34	412	123
SMAPE	84	169	148	77	119	89	153	75
R^2	0.72	0.15	0.40	0.85	0.55	0.62	-0.25	0,38
Elek – Aktobe								
MAE	8.6	18	11.9	9.7	11.4	13.1	16	10
MSE	1893	3527	1822	1920	1889	2016	2446	1779
SMAPE	54	129	93	80	61	64	56	59
R^2	0.68	0.29	0.61	0.55	0.68	0.55	0.42	0.58
Kargaly – Kargaly								
MAE	7.67	16.6	15	8.86	11.4	12.3	12.5	10.8
MSE	1255	2021	1983	2215	1317	1567	1900	1348
SMAPE	48	137	103	74	61	70	51	63
R^2	0.58	-0.20	0.05	-0.10	0.62	0.50	0.15	0.40
Ulken Kobda – Kobda								
MAE	3.4	4.6	0.6	0.58	4.4	4.5	0.8	4.2
MSE	81	45	1.68	1.51	147	215	4.2	110
SMAPE	43	128	39	37	102	118	35	95
R^2	0,65	0.20	0.95	0.96	0.45	0.30	0.72	0.55
Shagan – Kamenny								
MAE	4.1	6.6	4.1	4.1	3.4	5.9	4	4.6
MSE	258	250	151	196	255	246	312	204
SMAPE	63	128	65	63	51	75	61	69
R^2	0.55	0.05	0.70	0.68	0.50	0.30	0.45	0.55
Olenty – Zhympity								
MAE	0.6	0.94	0.55	0.3	0.65	0.75	0.44	1.64
MSE	11	14	12.4	10	12	14	12	21
SMAPE	182	187	183	178	183	185	152	76
R^2	0.68	0.05	0.45	0.70	0.65	0.40	0.60	0.36
Zhem – Zharkamys								
MAE	12	18	13	13	14	15	15	16
MSE	2145	2956	2459	2458	2554	2667	2689	2759
SMAPE	79	143	79	89	73	74	82	74
R^2	0.68	0.10	0.45	0.51	0.48	0.45	0.25	0.32
Oiyl – Taltogay								
MAE	10.3	13.3	10.4	10.8	11	12	10.5	13.6
MSE	1495	2458	1610	1522	1390	1955	1887	1904
SMAPE	52	112	87.4	63	77	57	79	51
R^2	0.70	0.10	0.68	0.67	0.62	0.55	0.22	0.36
Sagyz – Sagyz								
MAE	0.47	0.67	0.49	0.46	0.45	0.67	0.45	0.63
MSE	1.38	1.14	1.3	0.98	1	2	0.73	2
SMAPE	172	163	161	169	167	167	128	169
R^2	0.88	0.64	0.85	0.95	0.72	0.55	0.85	0.60

especially on the Oiyl and Zhem rivers. CB and GRB more consistently follow the recession limb, particularly during moderate floods for the Sagyz River. LGBM demonstrates high accuracy during sharp fluctuations on the Zhem River.

Our analysis of runoff forecasting models for the Bolshaya Kobda River during the test periods showed that both Gradient Boosting (GRB) and CatBoost (CB) models satisfactorily reproduce both flood and low-water periods. The GRB model more accurately tracked observed data during spring floods,especially in 2021, while CB tended to slightly overestimate maximum discharges. During low-water

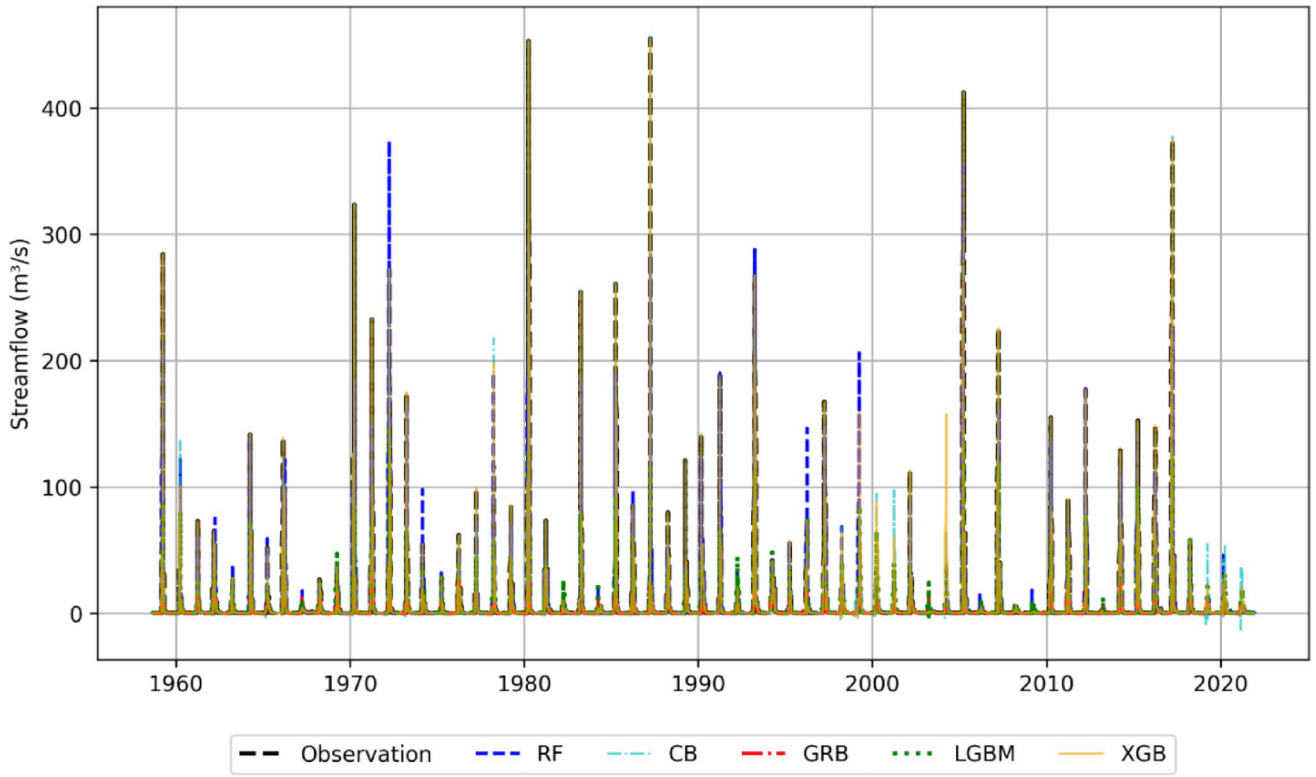


Fig. 5 Observed and predicted daily water discharge data for Or – Bogetsay using machine learning models

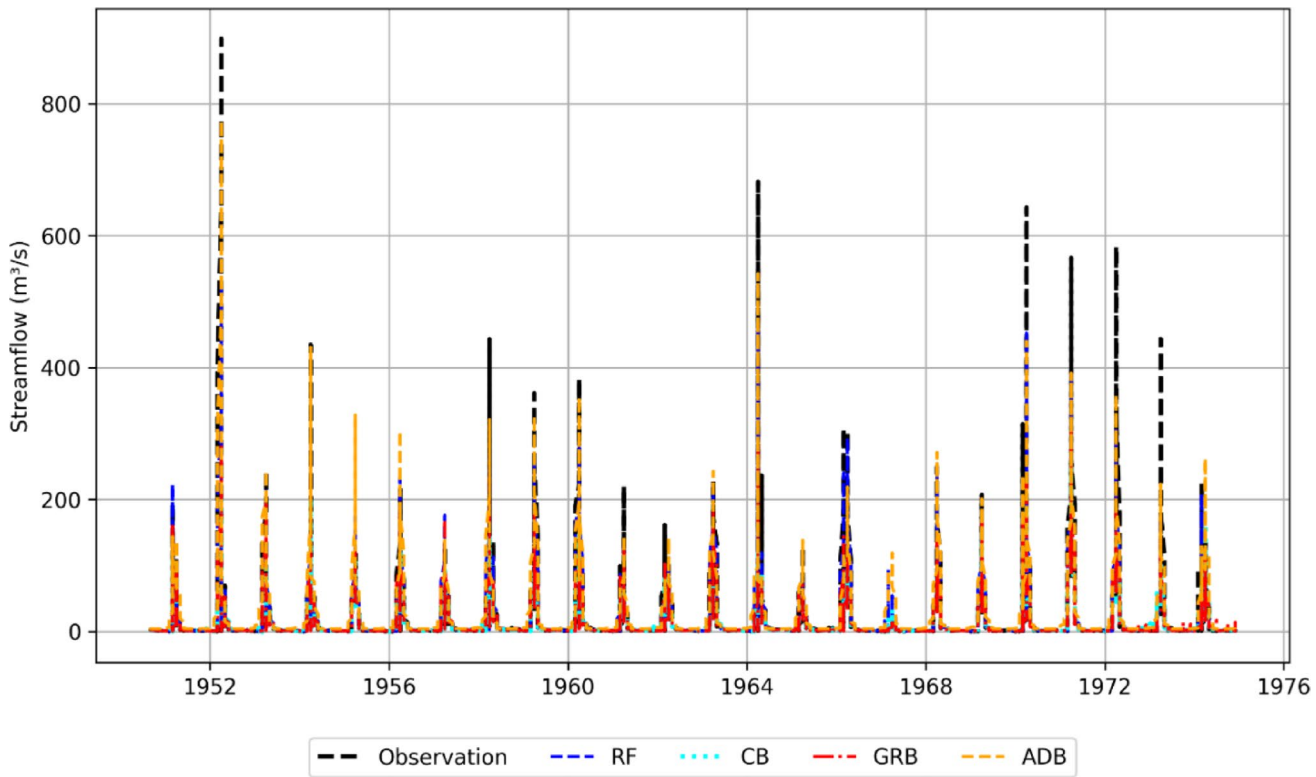


Fig. 6 Observed and predicted daily water discharge data for Elek – Aktobe using machine learning models

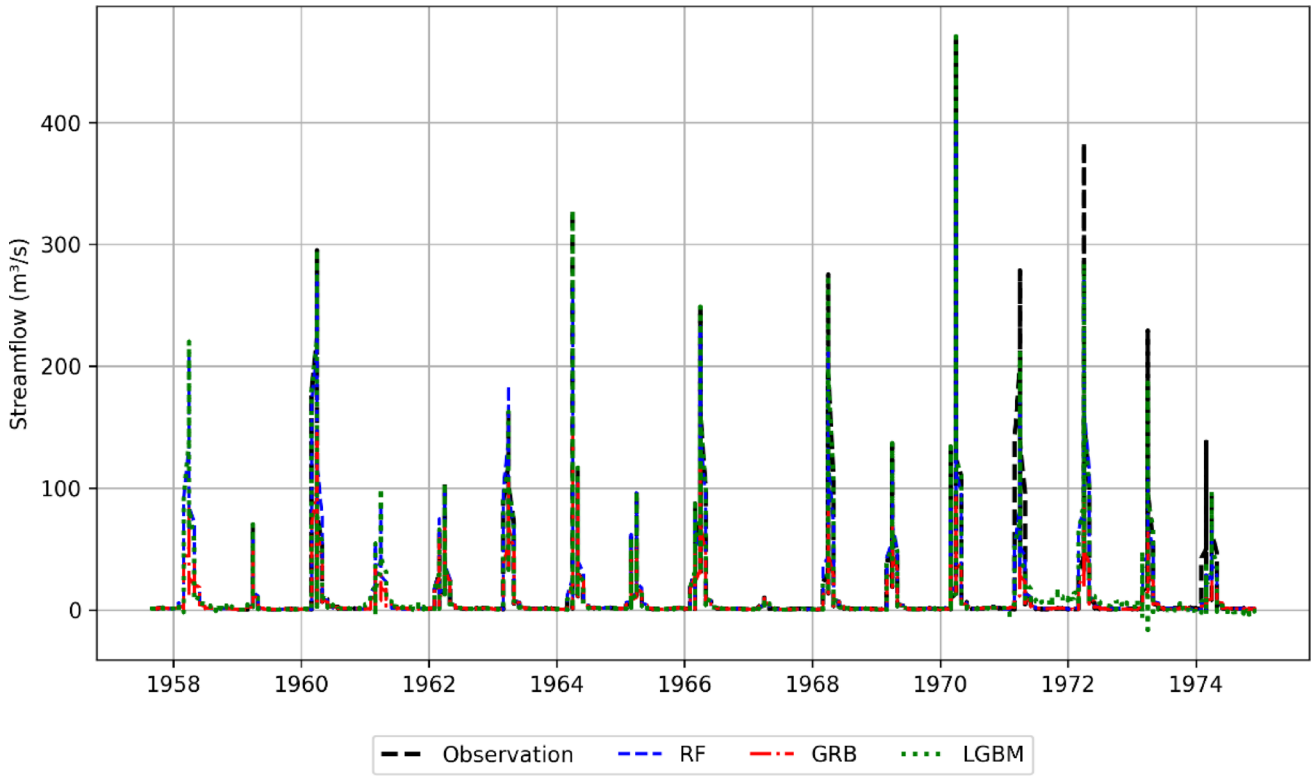


Fig. 7 Observed and predicted daily water discharge data for Kargaly – Kargaly using machine learning models

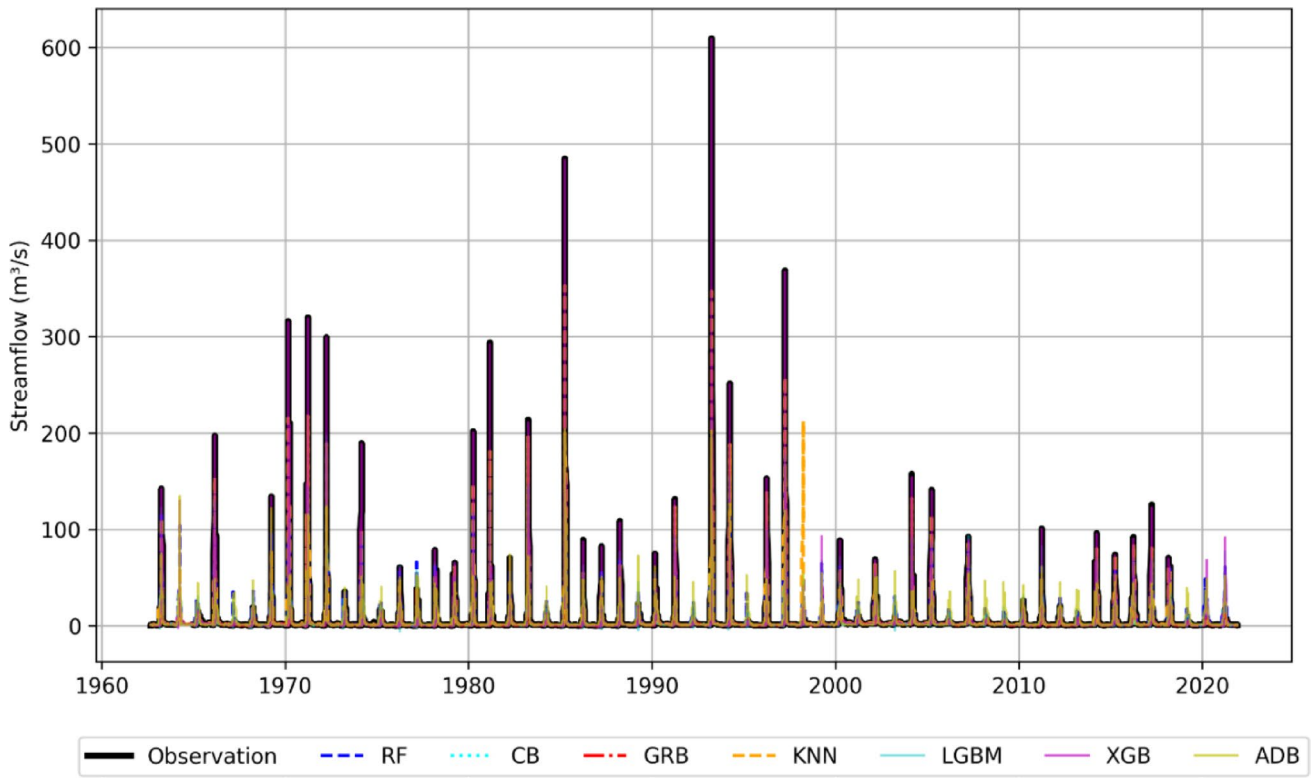


Fig. 8 Observed and predicted daily water discharge data for Ulken Kobda – Kobda using machine learning models

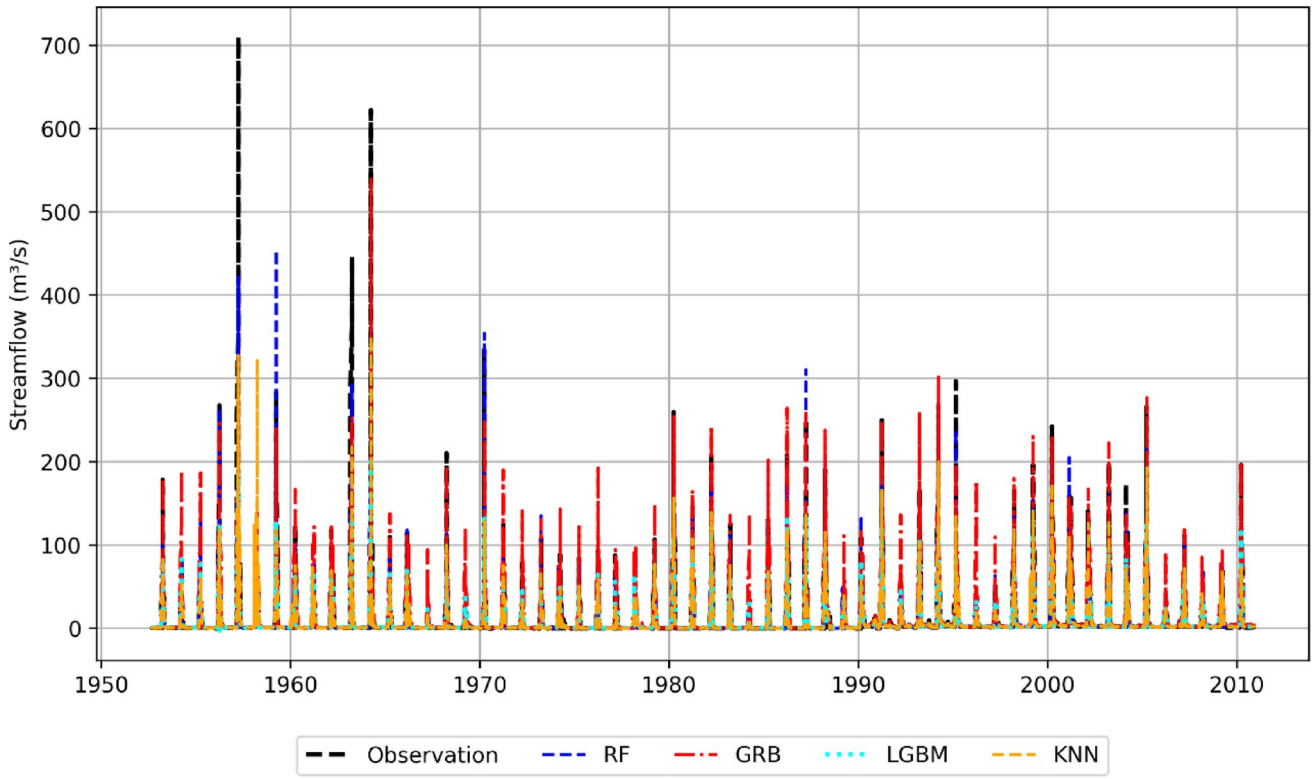


Fig. 9 Observed and predicted daily water discharge data for Shagan – Kamenny using machine learning models

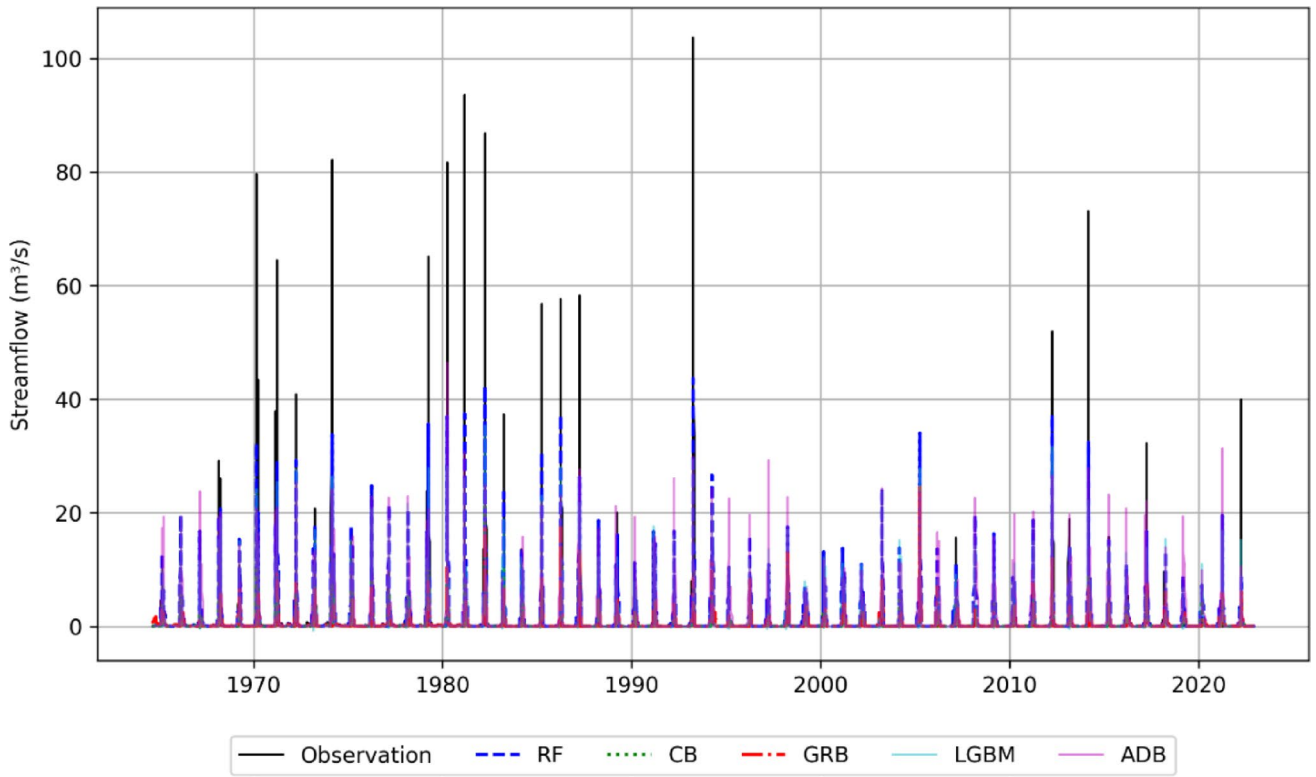


Fig. 10 Observed and predicted daily water discharge data for the Olenty – Zhympity using machine learning models

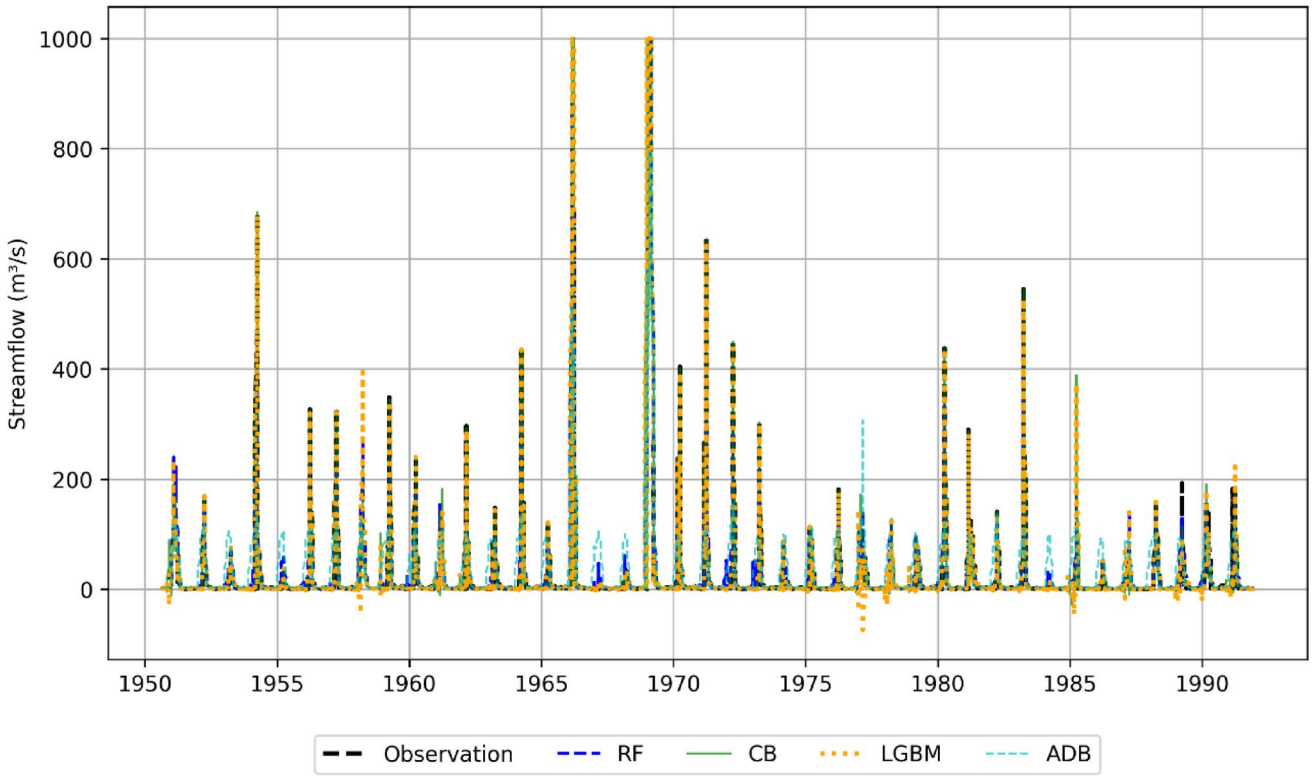


Fig. 11 Observed and predicted daily water discharge data for Zhem – Zharkamys using machine learning models

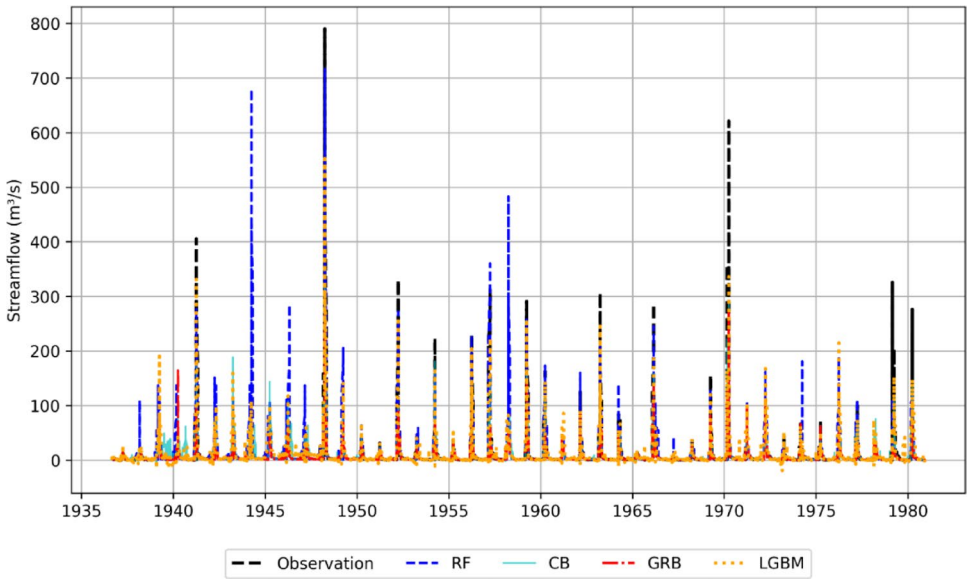


Fig. 12 Observed and predicted daily water discharge data for the Oiyl – Taltogay using machine learning models

conditions, differences between the models were minimal, with both adequately reproducing near-baseflow values.

For the Shagan River, model behavior varied with the hydrological regime. The GRB model systematically overestimated maximum water discharges during flood periods, nearly doubling the spring 2010 peak compared to observed values. In contrast, K-Nearest Neighbors (KNN) and LightGBM (LGBM) models provided a more balanced hydrograph reproduction, showing good agreement with observed values during both high-flow and low-water periods. The Random Forest (RF) model, however, tended to underestimate peaks but demonstrated stability during periods of low water. Thus, based on this comparison, KNN and LGBM proved to be the most robust models under both high and low flow conditions. Despite its accuracy during low flow, the GRB model is prone to overestimating peak values

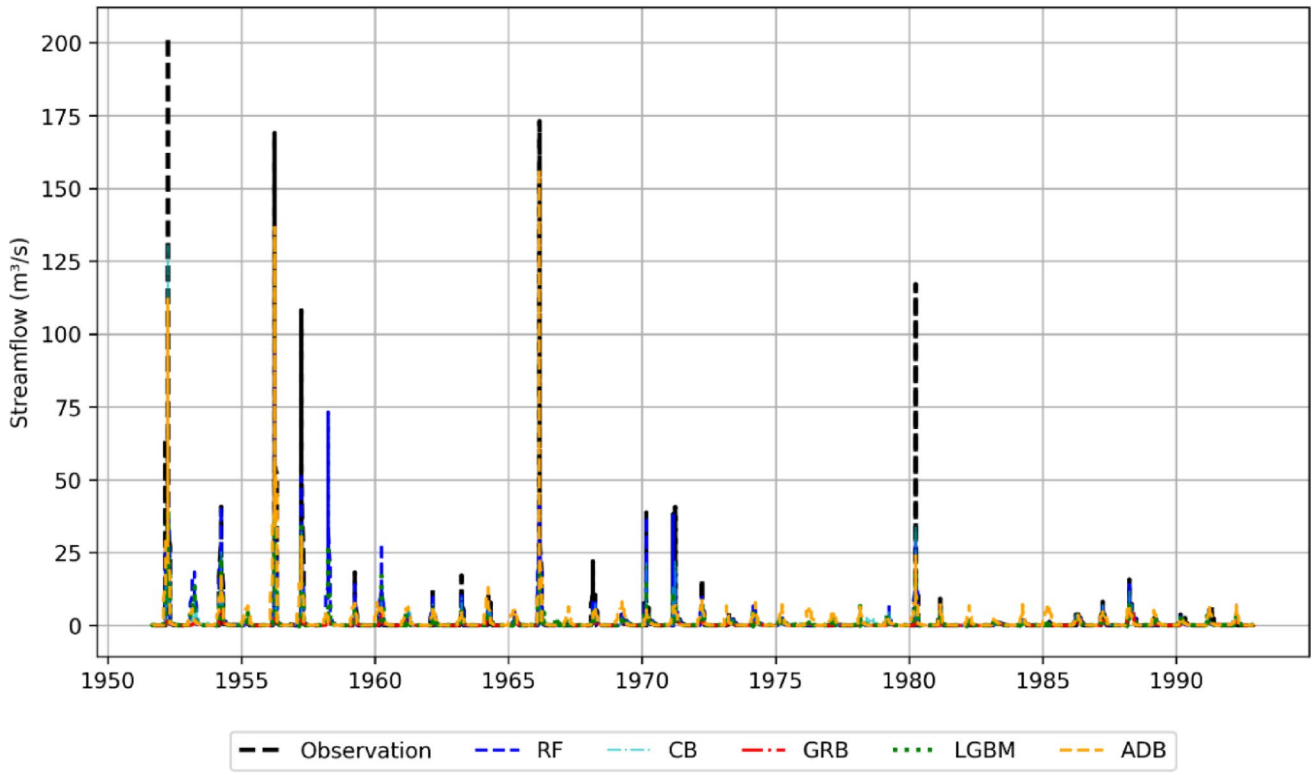


Fig. 13 Observed and predicted daily water discharge data for Sagyz – Sagyz using machine learning models

and requires additional tuning for use in flood-prone areas. CB and RF adequately reproduce trends but may be less accurate in extreme conditions.

An analysis of runoff simulation results for the Olenti River revealed differences in the models' ability to reproduce hydrological extremes. The most significant discrepancy with observed data was seen in the GRB model, which substantially overestimated discharge values during spring floods, particularly in 2010, where the simulated peak was almost twice the observed value. Conversely, LGBM and KNN models demonstrated good agreement with actual data during both flood and low-water periods. The RF model consistently underestimated peak discharges but showed stable accuracy at low water levels. Therefore, for the Olenti River, LGBM and KNN achieved the best results, demonstrating balanced behavior and a high degree of agreement with observed values. For the Or River, GRB and XGBoost models tend to overestimate peak flow values. However, GRB shows more stable and accurate results compared to XGBoost, especially in 2021. This may indicate GRB's superior ability to account for training sample characteristics and prevent overfitting on extreme values.

On the graphs, RF model residuals are uniformly distributed without systematic bias, indicating stability and an absence of overfitting. Additionally, hydrographs obtained using RF and CatBoost are similar in shape to observed ones, including during flood and inter-flood periods. Therefore, model selection depends on the specific characteristics of the watercourse and the requirements for accuracy or computational efficiency: RF and GRB are most effective when dealing with extreme values, while LightGBM is more suitable for faster and more stable calculations.

Very good simulation results were obtained with the Random Forest method for the Elek, Kargaly, Zhem, Oiyl and Sagyz rivers. Its ability to accurately reproduce flood peaks, closely approximating observed values, is particularly noteworthy. On the graphs, the residuals of the RF model are evenly distributed without systematic bias, indicating stability and the absence of overfitting. Furthermore, the hydrographs obtained using RF and CB are similar in shape to the observed ones, including during flood and inter-flood periods. Thus, the choice of model depends on the specifics of the watercourse and the requirements for accuracy or computational efficiency: RF and GRB are most effective when working with extreme values, while LGBM is suitable for faster and more stable calculations.

A slight scattering of points is observed at high flow values, which is typical for models working with extreme events. Based on the performance evaluation results, all individual ML models demonstrated satisfactory results. Overall, approaches based on the use of machine learning have confirmed their suitability for modeling daily river flow.

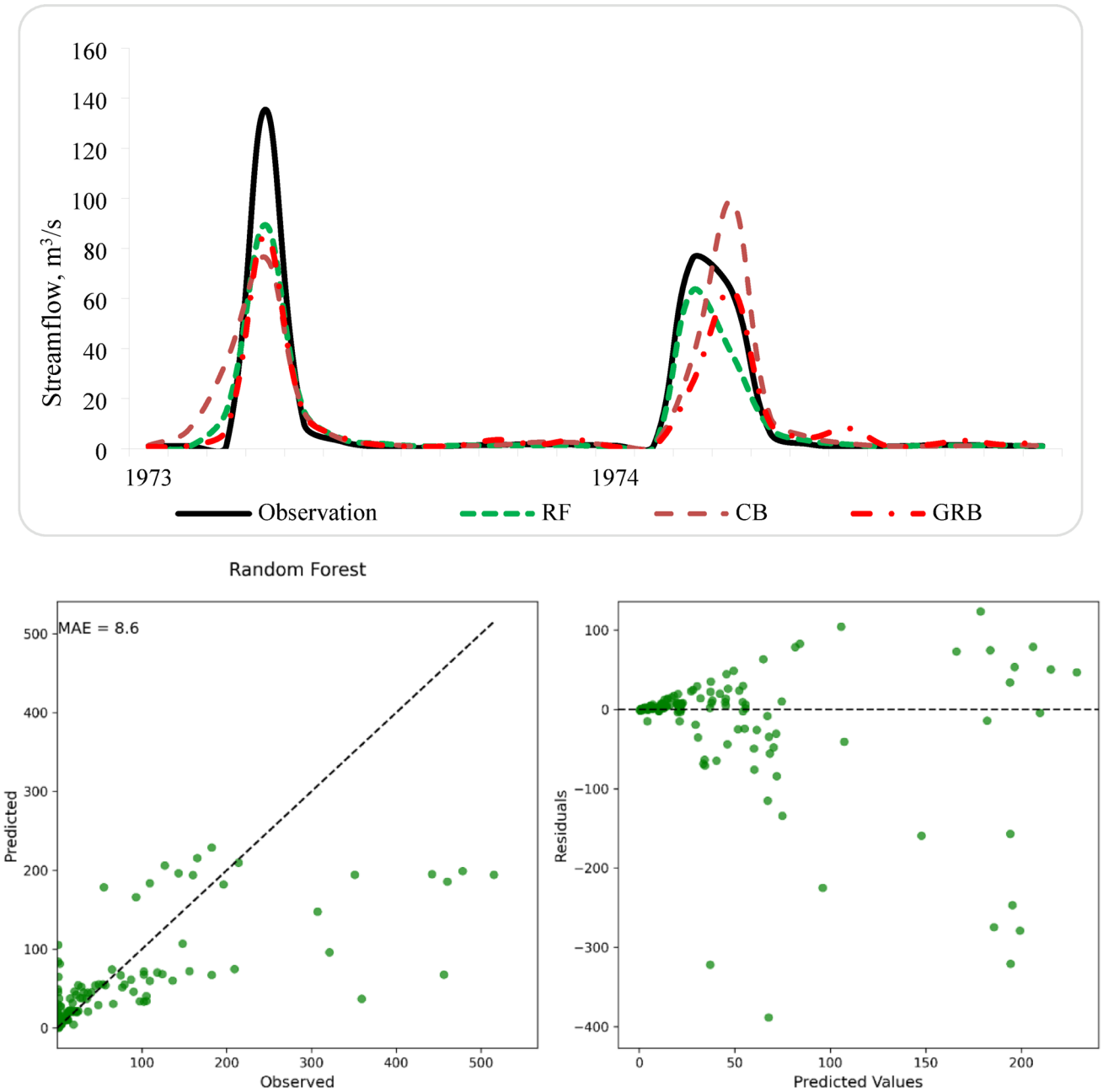


Fig. 14 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Elek – Aktobe, and a scatter plot of the linear regression for the best method

5 Discussion

Assessing and forecasting water resources in the transboundary Zhaiyk-Caspian water management basin, characterized by increasing climate variability and intensifying anthropogenic impact, is one of the key tasks for sustainable water resource management. This study aimed to address this challenge by applying ensemble ML algorithms for modeling daily river flow. A thorough comparative evaluation of five individual ML models and three ensemble approaches was conducted using rigorous cross-validation and hyperparameter optimization.

The effectiveness of various machine learning algorithms in modeling river flow varies depending on the hydrological regime phase. Particular attention was paid to data preprocessing,including the imputation of missing values using the RF algorithm, as well as the creation of new seasonal and temporal features, which ultimately allowed for a significant increase in prediction accuracy. This conclusion is consistent with a study (Hao and Bai 2023), in which

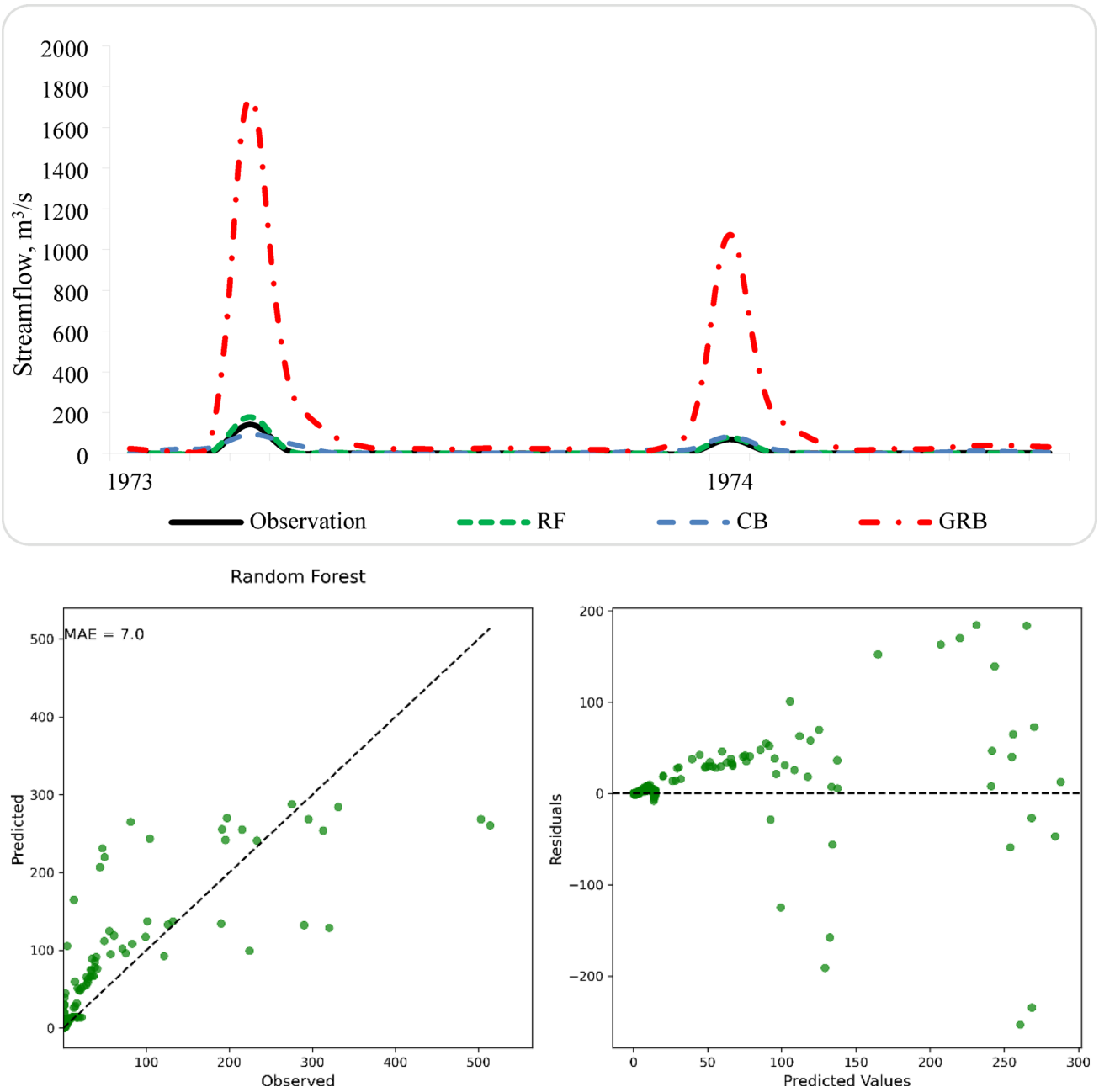


Fig. 15 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Kargaly – Kargaly, and a scatter plot of the linear regression for the best method

XGB, trained with dry season datasets, and LSTM, trained with wet season datasets, consistently achieved higher accuracy. The authors concluded that the improved performance indicates that classifying wet and dry seasons is beneficial for modeling.

The results obtained in this study, demonstrating the clear advantage of ensemble machine learning models in river flow forecasting, are supported by a growing body of scientific literature. Numerous studies consistently show the superiority of machine learning methods over traditional approaches, including multi-model ensemble (MME) methods (Li et al. 2021; Shetty et al. 2023; Ahmed et al. 2019; Fan et al. 2024; Xiao et al. 2022; Yan et al. 2022). It's important to note that the works by Yu and Yang (2024) and Li et al. (2024) have made significant contributions to understanding the capabilities of machine learning for rainfall-runoff modeling, revealing, in particular, the high effectiveness of LSTM in forecasting monthly runoff. Furthermore, studies by Wang et al. (2023) and Kartal et al. (2024) underscore the potential of combining multi-model climate forecasts

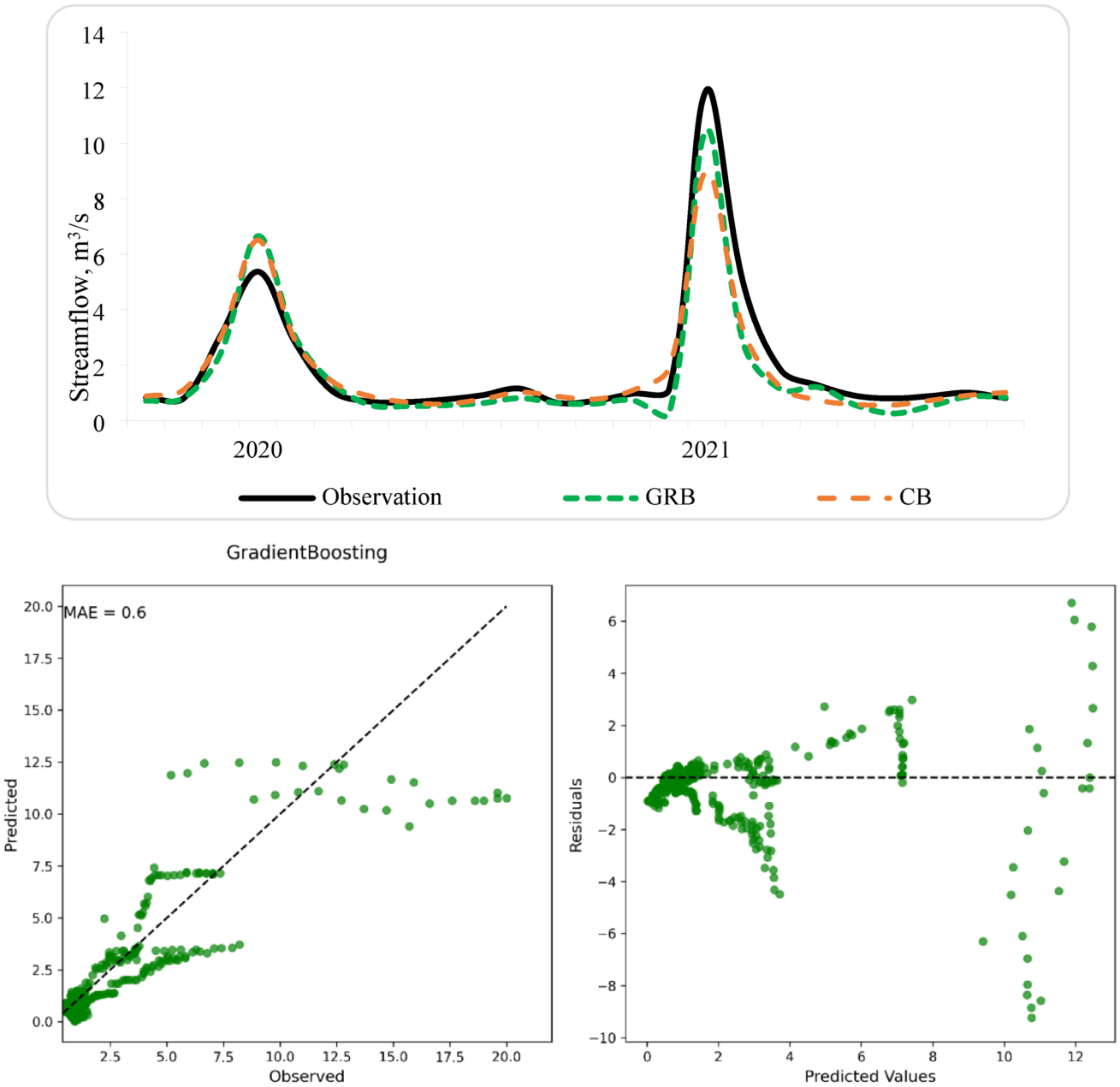


Fig. 16 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Ulken Kobda – Kobda, and a scatter plot of the linear regression for the best method

with machine learning methods to significantly improve precipitation forecast accuracy, which is also corroborated by our results, indicating a substantial increase in precipitation forecast accuracy using ML methods compared to raw CMIP6 model outputs.

Significant variability in the effectiveness of different models (e.g., MAE varied from 0.4 to 22) is attributed to both the inherent characteristics of the algorithms themselves and the hydrological and statistical properties of the studied rivers. Factors influencing forecast accuracy include the feeding regime (rainfall, snowmelt, mixed), flow dynamics (from sharp peaks to steady flow), the distribution and quality of initial data, noise levels, time series length, and the complexity of non-linear relationships between climatic parameters and river flow. Specifically, tree-based models such as CB, RF, and GRB demonstrated the best results on rivers with more stable hydrological regimes and high-quality data. At the same time, the linear LASSO model showed the worst performance on rivers with complex non-linear relationships and increased noise levels, which is explained by this model's limited ability to account for non-linearities. A study (Tyralis et al. 2021) applied a super-ensemble of 10

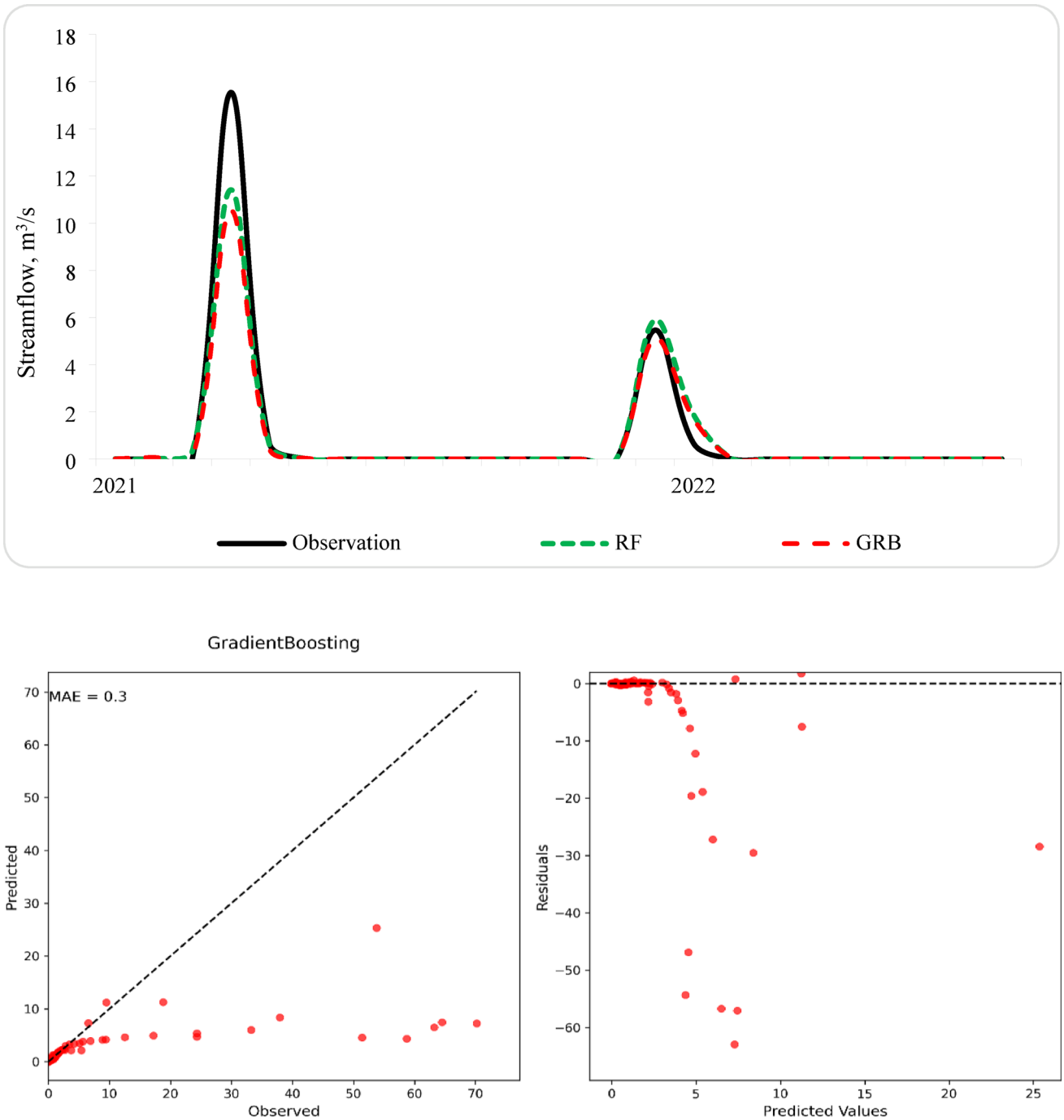


Fig. 17 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Olenty – Zhympity, and a scatter plot of the linear regression for the best method

different machine learning algorithms for daily river flow forecasting. The results showed better accuracy compared to linear regression by 20.06%, with neural networks, XGB, and Extremely Randomized Trees demonstrating the best performance. Thus, the authors conclude the effectiveness of ensemble methods in daily flow forecasting tasks.

Overall, tree-based algorithms demonstrated high effectiveness in modeling complex non-linear dependencies, while linear models proved more suitable for identifying long-term linear trends. These results underscore the necessity of selecting models with consideration for data specifics and hydrological conditions, as well as the importance of individually tuning algorithms for specific river basins. Similar conclusions are corroborated by several scientific works (Terzi et al. 2023; Ikram et al. 2023; Katipoğlu and Sarıgöl 2023).

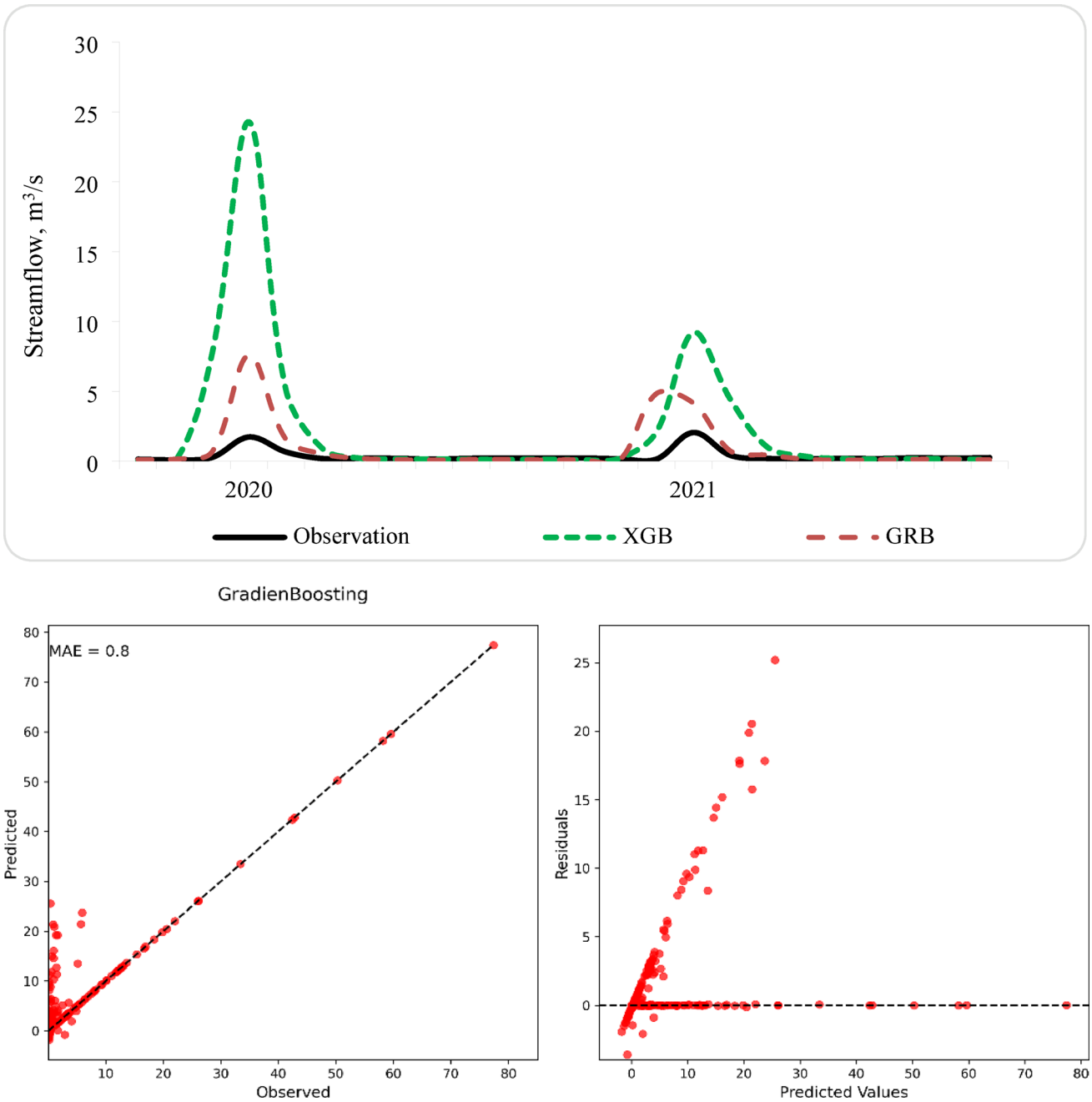


Fig. 18 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Or – Bogetsay, and a scatter plot of the linear regression for the best method

The obtained results have significant practical value, allowing for increased accuracy in forecasting water resources in the transboundary Zhaiyk-Caspian region. This, in turn, contributes to more effective management of water management systems. From a scientific perspective, this work contributes to the development of machine learning application methods in hydrology, demonstrating the influence of data structure and algorithm characteristics on forecast accuracy. Furthermore, the study expands knowledge about the region's hydrological regime and its sensitivity to climatic factors.

Future research should focus on expanding the input dataset to include information on land use, evaporation, and soil moisture, as well as accounting for the spatial heterogeneity of climatic factors. The application of modern neural network time series models, such as Darts and LSTM, which can effectively process multiple interconnected series, appears particularly promising. It's also relevant to develop

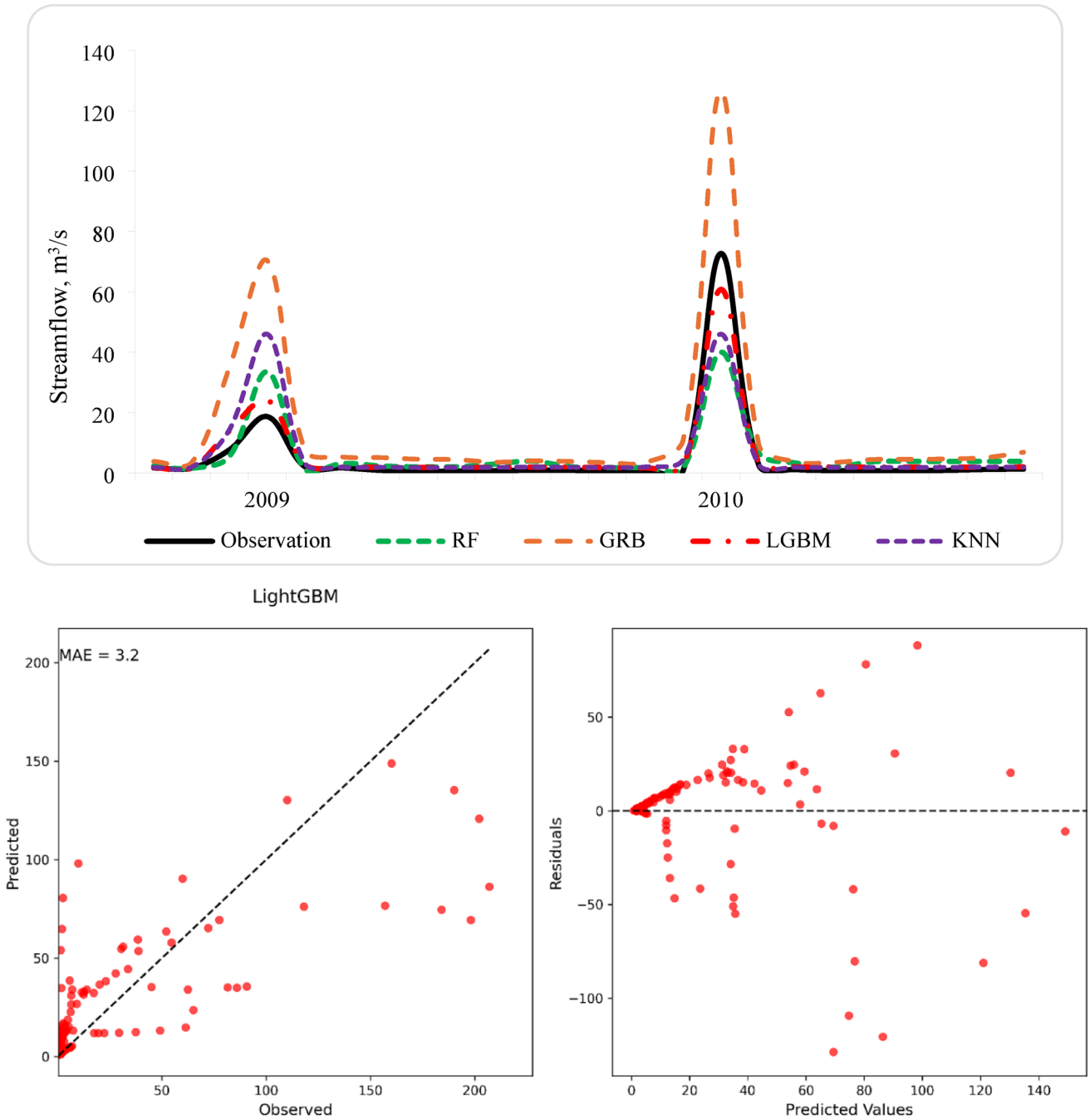


Fig. 19 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Shagan – Kamenny, and a scatter plot of the linear regression for the best method

adaptive ensemble models that can dynamically adjust their parameters in response to changing climatic conditions. Integrating machine learning models with spatially distributed hydrological models can provide a more comprehensive understanding of hydrological processes and enhance the reliability of forecasts under climate change conditions.

The results of this study confirm the effectiveness of applying ensemble machine learning algorithms for river flow forecasting in a transboundary basin. Comparison with international studies shows high consistency of the obtained results, which emphasizes the universality and reliability of the chosen approach. Despite some limitations, such as the need for large volumes of data and the complexity of interpretation, the advantages of ensemble methods make them promising for solving problems.

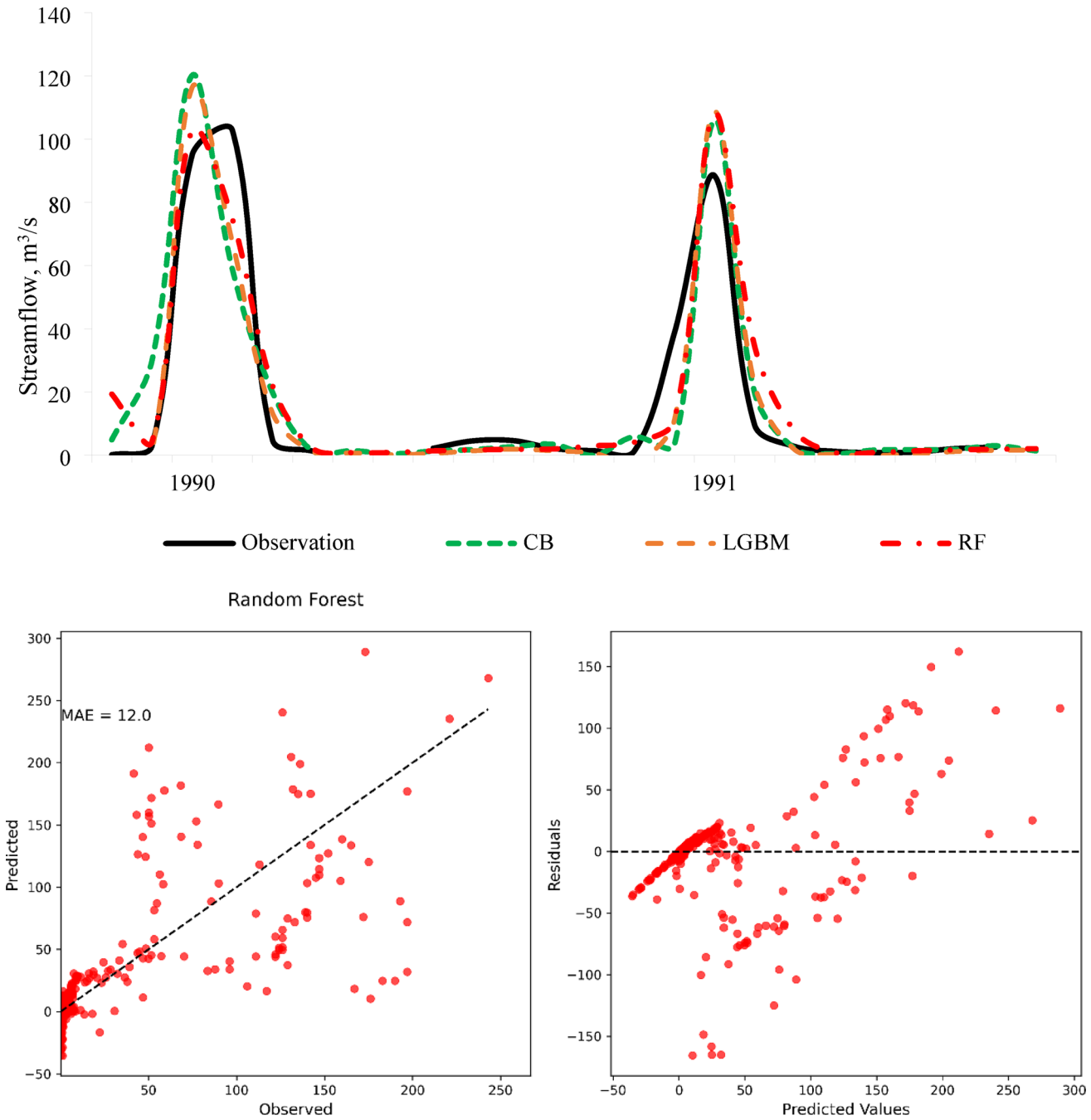


Fig. 20 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Zhem – Zharkamys, and a scatter plot of the linear regression for the best method

6 Conclusion

In the present study, a comprehensive evaluation of the effectiveness of eight machine learning algorithms for modeling river flow was conducted across six rivers in the Zhaiyk-Caspian water management basin. The analysis of the results revealed the following:

1. Variability in Algorithm Effectiveness: Mean Absolute Error (MAE) values varied across a wide range, indicating a significant influence of hydrological conditions and data characteristics of individual catchments on modeling accuracy.
2. Advantage of Ensemble Methods: Tree-based ensemble algorithms (RF, C, GRB, LGBM) demonstrated the best accuracy metrics (MAE≈0.3), confirming their ability

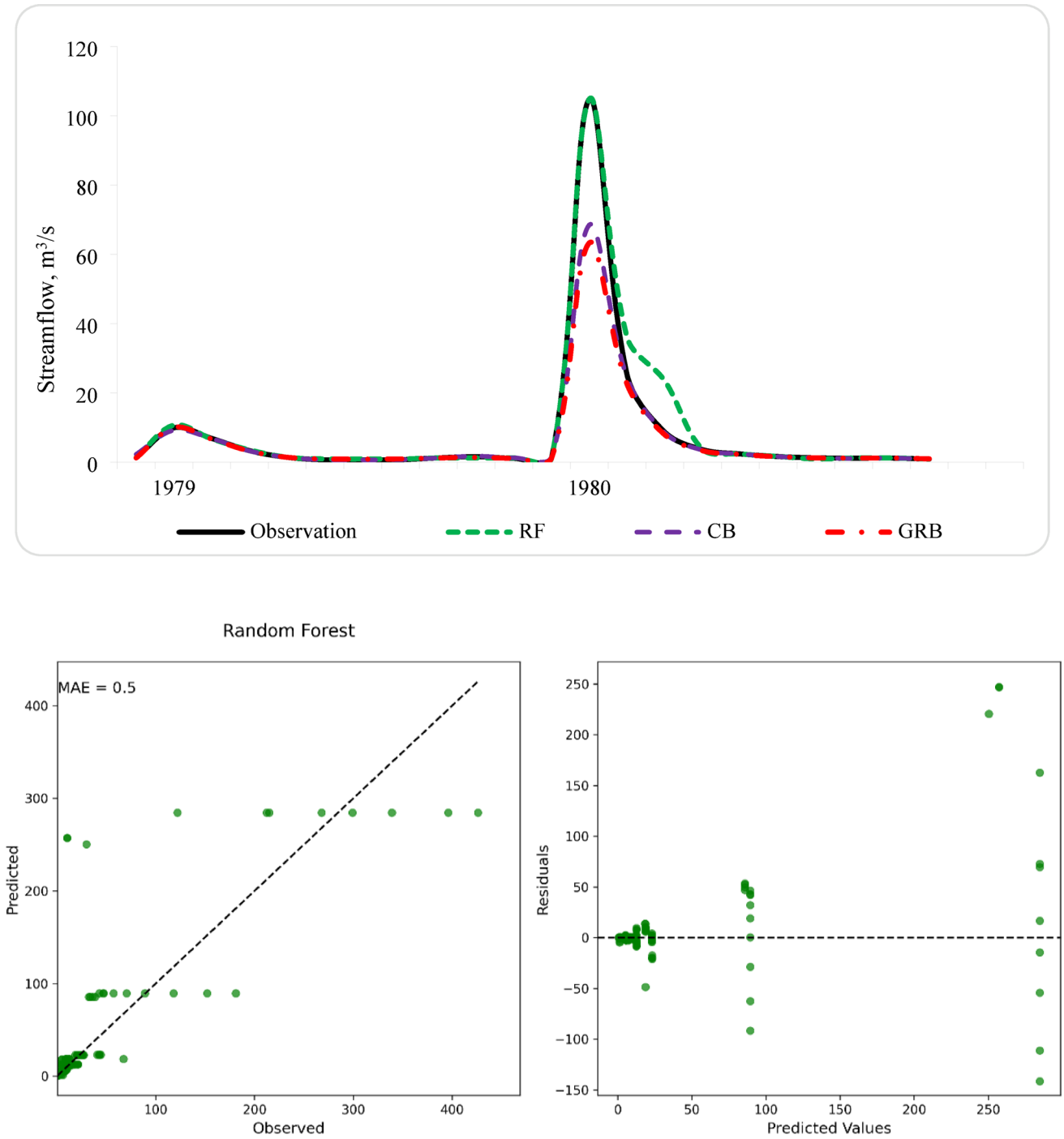
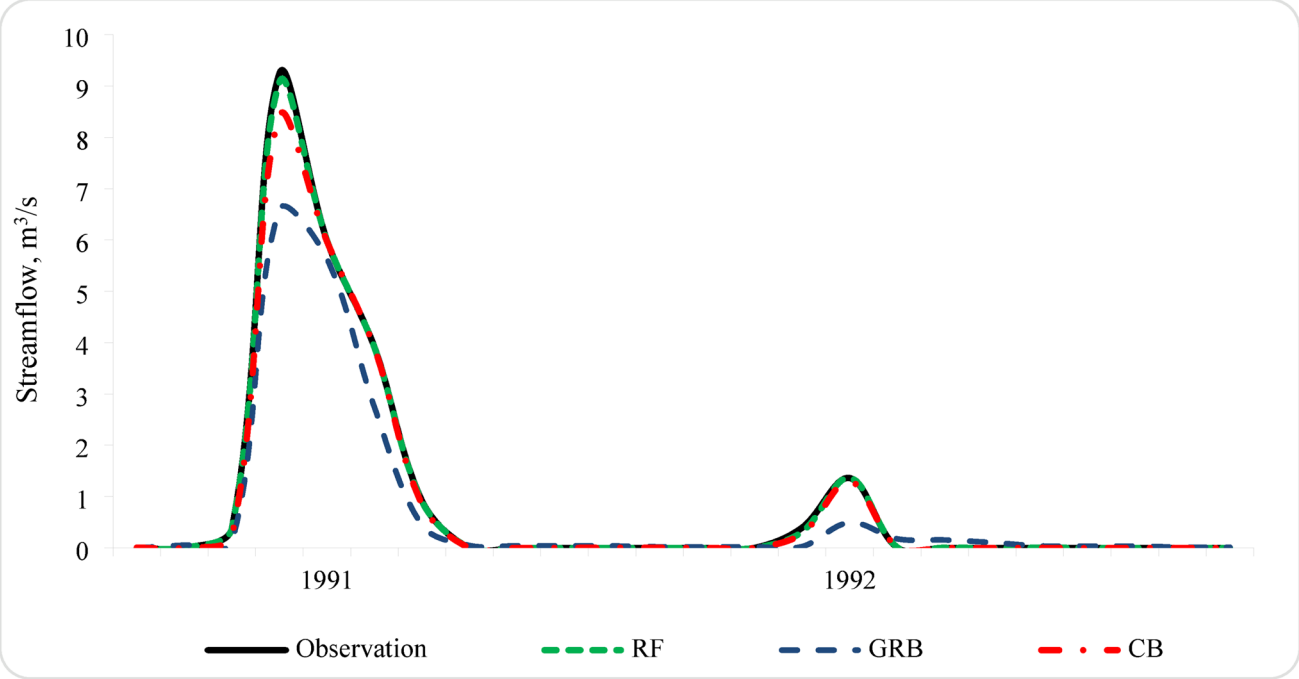


Fig. 21 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Oiyl – Taltogay, and a scatter plot of the linear regression for the best method

to effectively handle complex non-linear dependencies typical of hydrological processes.

3. Limitations of Weaker Models: LASSO and ADB models showed the worst overall performance (MAE≈22), likely due to their inability to adequately describe non-linear dependencies and their limited robustness to noise in the data.
4. Practical Significance: The obtained results emphasize the high importance of model selection based on the characteristics of a specific region. Reliable river flow forecasts are crucial for effective water resource management, especially under conditions of climate change and a lack of observed data.



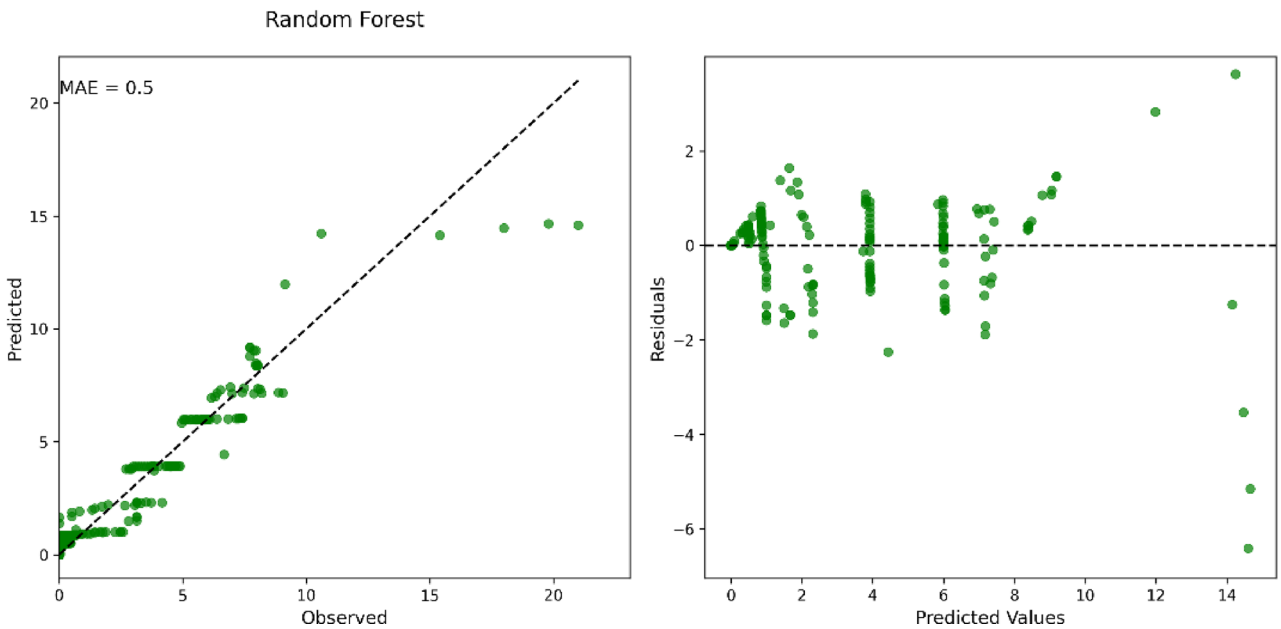


Fig. 22 Comparison of mean monthly observed water discharges with modeling results for the two-year test period for Sagyz – Sagyz, and a scatter plot of the linear regression for the best method

Directions for Future Research:

- Investigating the influence of data volume, noise level, and feature distribution on model performance;
- Developing specialized ensemble models tailored to the hydrological characteristics of specific basins;
- Applying advanced neural network time series architectures (e.g., Darts) to enhance prediction accuracy;
- Integrating machine learning methods with physically based hydrological models.

Overall, this study demonstrates the high potential of applying machine learning methods for river flow modeling.

Reliable flow assessment is critically important for mitigating the consequences of natural disasters (such as droughts and floods) and making informed decisions in water resource management and climate change adaptation.

Author Contribution Sayat Alimkulov: conceptualization, formal analysis; Lyazzat Makhmudova: conceptualization, methodology writing-review and editing,; Bekzat Satenova: validation, writing-review and editing, writing – original draft, software; Aisulu Tursunova: conceptualization, formal analysis and investigation; Lyazzat Birimbayeva: data acquisition, writing – original draft, writing-review and editing; Elmira Talipova: data acquisition, editing;Darkhan Abdibekov visualization, software; Zhanibek Smagulov and Oirat Alzhanov: data acquisition, visualization. (All authors contributed to the article and approved the submitted version).

Funding This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Grant No. BR28713279 "Scientific and Applied Foundations for Flood Risk Management in the Flat and Low Hilly Territories of Kazakhstan under Modern Climate Change Conditions".

Declarations

Conflict of interest The authors declare no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abbass K, Qasim MZ, Song H et al (2022) A review of the global climate change impacts, adaptation, and sustainable mitigation measures. Environ Sci Pollut Res 29:42539–42559. https://doi.org/10.1007/s11356-022-19718-6

Ahmed K, Sachindra DA, Shahid S (2019) Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. Atmos Res. https://doi.org/10.1016/j.atmosres.2019.104806

Akiyanova FZ, Abitbayeva AD, Yegemberdiyeva K, Temirbayeva RK (2014) Problems of desertification of the territory of kazakhstan: status and forecast. Life Sci J 11(10):341–345

Alimkulov S, Makhmudova L, Talipova EK, Baspakova G, Tigkas D, Gulsaira I (2024a) Response of the water level of the Balkash lake to the distribution of meteorological and hydrological droughts under the conditions of climate change. J Water Clim Change 15(7):3395–3408. https://doi.org/10.2166/wcc.2024.271

Alimkulov S, Makhmudova L, Talipova EK, Baspakova G, Tigkas D, Gulsaira I (2024b) Response of the water level of Balkash lake to the distribution of meteorological and hydrological droughts under the conditions of climate change. J Water Clim Change 15(7):3395–3408. https://doi.org/10.2166/wcc.2024.271

Alimkulov S, Makhmudova L, Talipova E, Baspakova G, Monkayeva G (2025) Assessment of the impacts of climate change on drought intensity and frequency using SPI and SPEI in the Southern Pre-Balkash region, Kazakhstan. Watershed Ecol Environ 7:11–22. https://doi.org/10.1016/j.wsee.2024.12.001

Bangira C (2019) Food security as a water grand challenge. J Contemp Water Res Educ 165(1):59–66. https://doi.org/10.1111/j.1936-704X.2018.03293.x

Birimbayeva L, Makhmudova L, Alimkulov S, Tursunova A, Mussina A, Tigkas D, Beksultanova Z, Rodrigo-Clavero M-E, Rodrigo-Ilarri J (2024) Analysis of the Spatiotemporal variability of hydrological drought regimes in the lowland rivers of Kazakhstan. Water 16(16):2316. https://doi.org/10.3390/w16162316

Birimbayeva L, Makhmudova L, Alimkulov S et al (2025) Reaction of minimal streamflow to natural factors in the context of climate uncertainty. Water Resour Manage. https://doi.org/10.1007/s11269-025-04238-y

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Ciampittiello M, Marchetto A, Boggero A (2024) Water resources management under climate change: A review. Sustainability 16(9):3590. https://doi.org/10.3390/su16093590

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27. https://doi.org/10.1109/TIT.1967.1053964

Davletgaliev SK, Alimkulov SK, Talipova EK (2020) The possibility of applying simulated series for compile scenario forecasting river runoff. Environ Earth Sci 79:397. https://doi.org/10.1007/s12665-020-09138-5

Dong N, Hao H, Yang M, Wei J, Xu S, Kunstmann H (2025) Deep-learning-based sub-seasonal precipitation and streamflow ensemble forecasting over the source region of the Yangtze river. Hydrol Earth Syst Sci 29:2023–2042. https://doi.org/10.5194/hess-29-2023-2025

Fan Y, Fu X, Kan G, Liang K, Yu H (2024) Combining multiple machine learning methods based on CARS algorithm to implement runoff simulation. Water 16:2397. https://doi.org/10.3390/w16172397

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139. https://doi.org/10.1006/jcss.1997.1504

Hao R, Bai Z (2023) Comparative study for daily streamflow simulation with different machine learning methods. Water 15(6):1179. https://doi.org/10.3390/w15061179

Hydrology and Climate Group (H2K). (n.d.). HBV model. University of Zurich. Retrieved April 28 (2025) from https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html

HydroSolutions (2025) SAPPHIRE Central Asia Project – Harnessing deep learning for operational discharge forecasting in Central Asia. Available at: https://www.hydrosolutions.ch/projects/sapphire-central-asia-project---harnessing-deep-learning-for-operational-discharge-forecasting-in-central-asia (Accessed 08 September 2025)

Ikram RMA, Hazarika BB, Gupta D, Heddam S, Kisi O (2023) Streamflow prediction in mountainous region using new machine learning and data preprocessing methods: A case study. Neural Comput Appl 35:9053–9070. https://doi.org/10.1016/j.jher.2023.07.004

IPCC (2022) Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment

Report of the Intergovernmental Panel on Climate Change (H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, Eds.). Cambridge University Press. https://doi.org/10.1017/9781009325844

Jahanbani H, Ahmed K, Gu B (2024) Data-driven artificial intelligence–based streamflow forecasting: a review of methods, applications, and tools. J Am Water Resour Assoc 60(6):1095–1119. https://doi.org/10.1111/1752-1688.13229

Kabdrakhova SS, Seilkhan A, Assan ZZ (2024) Flood forecasting in Malaya Almatinka river via machine learning and deep learning with oversampling. J Probl Comput Sci Inform Technol 2(1):15–24. https://doi.org/10.26577/jpcsit2024020102

Kartal V (2024) Machine learning-based streamflow forecasting using CMIP6 scenarios: assessing performance and improving hydrological projections and climate change. Hydrol Process 38(6):e15204. https://doi.org/10.1002/hyp.15204

Katipoğlu OM, Sarıgöl M (2023) Application of boosted tree algorithm with new data preprocessing techniques in the forecasting one-day-ahead streamflow values in the Tigris basin, Türkiye. J Hydroinformatics 50:13–25. https://doi.org/10.1016/j.ecoinf.2023.102455

Kazhydromet (2025) Climate of Kazakhstan by regions. RSE Kazhydromet. Available at: https://www.kazhydromet.kz/ru/klimat/klimat-kazahstana-po-oblastyam (Accessed 01 September 2025)

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3149–3157). Long Beach, CA. https://doi.org/10.5555/3294996.3295074

Kumar V, Kedam N, Sharma KV, Mehta DJ, Caloiero T (2023) Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models. Water 15(14):2572. https://doi.org/10.3390/w15142572

Li P, Wu J (2024) Water resources and sustainable development. Water 16(1):134. https://doi.org/10.3390/w16010134

Li T, Jiang Z, Le Treut H, Li L, Zhao L, Ge L (2021) Machine learning to optimize climate projection over China with multi-model ensemble simulations. Environ Res Lett 16(9):094028. https://doi.org/10.1088/1748-9326/ac1d0c

Li C, Jiao Y, Kan G, Fu X, Chai F, Yu H, Liang K (2024) Comparisons of different machine learning-based rainfall–runoff simulations under changing environments. Water 16(2):302. https://doi.org/10.3390/w16020302

Lu D, Konapala G, Painter SL, Kao S, Gangrade S (2021) Streamflow simulation in data-scarce basins using bayesian and physics-informed machine learning models. J Hydrometeorol 22:1421–1438. https://doi.org/10.1175/JHM-D-20-0082.1

Maity R, Srivastava A, Sarkar S, Khan MI (2024) Revolutionizing the future of hydrological science: impact of machine learning and deep learning amidst emerging explainable AI and transfer learning. Appl Comput Geosci 24:100206. https://doi.org/10.1016/j.acags.2024.100206

Matanó A, Hamed R, Brunner MI, Barendrecht MH, Van Loon AF (2025) Drought decreases annual streamflow response to precipitation, especially in arid regions. Hydrol Earth Syst Sci 29(12):2749–2764. https://doi.org/10.5194/hess-29-2749-2025

Menze BH, Kelm BM, Masuch R et al (2009) A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10:213. https://doi.org/10.1186/1471-2105-10-213

MIKE SHE integrated hydrological modeling. Retrieved April 28 (2025) from https://www.dhigroup.com/technologies/mikepoweredbydhi/mike-she

National Oceanic and Atmospheric Administration, Office for Water Prediction. (n.d.). SAC-SMA Model [GitHub repository]. GitHub. Retrieved April 28 (2025) from https://github.com/NOAA-OWP/sac-sma

Nearing GS, Kratzert F, Sampson AK, Pelissier CS, Klotz D, Frame JM et al (2021) What role does hydrological science play in the age of machine learning? Water Resour Res 57:e2020WR028091. https://doi.org/10.1029/2020WR028091

Niazkar M, Menapace A, Brentan B, Piraei R, Jimenez D, Dhawan P, Righetti M (2024) Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environ Model Softw 174:105971. https://doi.org/10.1016/j.envsoft.2024.105971

Nurbatsina A, Salavatova Z, Tursunova A, Didovets I, Huthoff F, Rodrigo-Clavero M-E, Rodrigo-Ilarri J (2025) Flood modelling of the Zhabay river basin under climate change conditions. Hydrology 12(2):35. https://doi.org/10.3390/hydrology12020035

Pelikan M (2005) Hierarchical Bayesian optimization algorithm: Toward a new generation of evolutionary algorithms (Studies in Fuzziness and Soft Computing, Vol. 170). Springer. https://doi.org/10.1007/11007937

Pham LT, Luo L, Finley A (2021) Evaluationof random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. Hydrol Earth Syst Sci 25:2997–3015. https://doi.org/10.5194/hess-25-2997-2021

Ringler C, Agbonlahor M, Barron J, Baye K, Meenakshi JV, Mekonnen DK, Uhlenbrook S (2022) The role of water in transforming food systems. Global Food Secur 33:100639. https://doi.org/10.1016/j.gfs.2022.100639

Roth V (2004) The generalized LASSO. IEEE Trans Neural Networks 15(1):16–28. https://doi.org/10.1109/TNN.2003.809398

Rozos E (2023) Assessing hydrological simulations with machine learning and statistical models. Hydrology 10(2):49. https://doi.org/10.3390/hydrology10020049

Sadrtdinova R, Corzo Perez GA, Solomatine DP (2024) Improved drought forecasting in Kazakhstan using machine and deep learning: A non-contiguous drought analysis approach. Hydrol Res 55(2):237–261. https://doi.org/10.2166/nh.2024.154

Sarsekov AS, Visloguzova AV, Vladimirov NM, Gus'kova AI, Medeu AR, Nurmambetov EI, Potapova GM (2017) Relief of Kazakhstan (Explanatory note to the Geomorphological map of Kazakhstan, scale 1:1,500,000) (2nd ed., 236 p.). Almaty. [In Russian]

Sheikh MR, Coulibaly P (2024) Review of recent developments in hydrologic forecast merging techniques. Water 16(2):301. https://doi.org/10.3390/w16020301

Shetty S, Umesh P, Shetty A (2023) The effectiveness of machine learning-based multi-model ensemble predictions of CMIP6 in Western Ghats of India. Int J Climatol 43(11):5029–5054. https://doi.org/10.1002/joc.8131

Singh D, Vardhan M, Sahu R, Chatterjee D, Chauhan P, Liu S (2023) Machine-learning- and deep-learning-based streamflow prediction in a hilly catchment for future scenarios using CMIP6 GCM data. Hydrol Earth Syst Sci 27:1047–1065. https://doi.org/10.5194/hess-27-1047-2023

Szczepanek R (2022) Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and catboost. Hydrology 9(12):226. https://doi.org/10.3390/hydrology9120226

Terzi Ö, Küçüksille EU, Baykal T, Taylan ED (2023) Deep and machine learning for daily streamflow estimation: A focus on LSTM, RFR and XGBoost. Water Pract Technol 18(10):2401–2414. https://doi.org/10.2166/wpt.2023.144

Tokbergenova A, Skorintseva I, Ryskeldiyeva A, Kaliyeva D, Salmurzauly R, Mussagaliyeva A (2025) Assessment of anthropogenic disturbances of landscapes: West Kazakhstan region. Sustainability 17:573. https://doi.org/10.3390/su17020573

Tyralis H, Papacharalampous G, Langousis A (2021) Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms.

Neural Comput Appl 33:3053–3068. https://doi.org/10.1007/s00521-020-05172-3

Wang D, Liu J, Luan Q et al (2023) Projection of future precipitation change using CMIP6 multimodel ensemble based on fusion of multiple machine learning algorithms: A case in Hanjiang river Basin, China. Meteorol Appl 30(5):e2144. https://doi.org/10.1002/met.2144

Wortmann M (n.d.) (ed) (2025) SWIM: Soil and Water Integrated Model. PIK Potsdam. Retrieved April 28, from https://www.pik-potsdam.de/~wortmann/swim/

Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H (2019) Hyperparameter optimization for machine learning models based on bayesian optimization. J Electron Sci Technol 17(1):26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Xiao L, Zhong M, Zha D (2022) Runoff forecasting using machine-learning methods: case study in the middle reaches of Xijiang river. Front Big Data 4:752406. https://doi.org/10.3389/fdata.2021.752406

Xu T, Liang F (2021) Machine learning for hydrologic sciences: an introductory overview. Wiley Interdisciplinary Reviews: Water 8(5):e1533. https://doi.org/10.1002/wat2.1533

Xu N, Wang Z, Dai Y, Li Q, Zhu W, Wang R et al (2023) Prediction of higher heating value of coal based on gradient boosting regression tree model. Int J Coal Geol 274:104293. https://doi.org/10.1016/j.coal.2023.104293

Yan Y, Wang H, Li G, Xia J, Ge F, Zeng Q, Ren X, Tan L (2022) Projection of future extreme precipitation in China based on the CMIP6 from a machine learning perspective. Remote Sens 14(16):4033. https://doi.org/10.3390/rs14164033

Yu H, Yang Q (2024) Applying machine learning methods to improve rainfall–runoff modeling in subtropical river basins. Water 16(15):2199. https://doi.org/10.3390/w16152199

Zhao X, Wang H, Bai M, Xu Y, Dong S, Rao H, Ming W (2024) A comprehensive review of methods for hydrological forecasting based on deep learning. Water 16(10):1407. https://doi.org/10.3390/w16101407

Ziyatbekova G, Darkenbayev D, Shekerbayeva S, Zhaksymbet A (2025) Integrating satellite data and machine learning algorithms into a flood prediction system. Int J Innovative Res Sci Stud 8(5):461–472. https://doi.org/10.53894/ijirss.v8i5.8678

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.