

A data-driven framework to inform extreme flood warning: Evaluating dual-tier weighting in contrasting catchments

Ali akbar Eatesam

a.etesam@semnan.ac.ir

Semnan University <https://orcid.org/0009-0009-6614-1811>

Khosrow Hoseini

Semnan University

Hojat Karami

Semnan University

Research Article

Keywords: Extreme flood detection, Gradient boosting, Low-yield catchments, Flood risk management, Weighting scenarios, Metaheuristic optimization

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11122208/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A data-driven framework to inform extreme flood warning: Evaluating dual-tier weighting in contrasting catchments

Ali Akbar Eatesam¹, Khosrow Hoseini¹, Hojat Karami¹

¹Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan, Iran

Corresponding Author: Ali akbar Eatesam

Corresponding author's email: a.etesam@semnan.ac.ir

Corresponding author's ORCID ID: 0009-0009-6614-1811

Abstract

Detecting rare flood extremes in low-yield catchments is difficult because unweighted loss functions give the few high-flow days little influence on model training. This study evaluates whether weighting extreme residuals during calibration improves detection without degrading continuous streamflow simulation. Daily streamflow in two contrasting Iranian basins, responsive Haraz (runoff coefficient 0.57) and attenuated lowland Tajan (0.082), was simulated with two gradient-boosting learners (XGBoost and LightGBM) calibrated in two stages: Bayesian hyperparameter screening followed by metaheuristic tuning, under three weighting scenarios and two objective functions at a 24-hour lead time. Performance was assessed with continuous metrics and with detection, false-alarm, and critical-success indices at the 98th and 99th flow percentiles. In Haraz, single-tier weighting sufficed (Nash-Sutcliffe efficiency 0.98; detection 0.83-0.85 at the 99th percentile). In Tajan, unweighted models detected only 0.20-0.29 of extremes; dual weighting with a residual-based objective doubled detection to 0.40 while keeping efficiency at 0.55, at a false-alarm ratio of 0.57. The same weighting with a variance-based objective degraded continuous accuracy (efficiency 0.27-0.39), identifying the objective function, not the weighting, as the failure source. Run-to-run stability checks confirmed bounded metric spreads. The results indicate that weighting and objective choices should follow catchment attenuation capacity, and that the detection-false-alarm trade-off is acceptable where missed floods carry higher consequences than false alarms.

Keywords: Extreme flood detection, Gradient boosting, Low-yield catchments, Flood risk management, Weighting scenarios, Metaheuristic optimization

1- Introduction

Flood early warning, reservoir operation, and floodplain management all depend on reliable antecedent-based forecasts of extreme flows, especially as the population and assets exposed to flooding continue to increase (Tellman et al., 2021). However, the basins where these decisions are made are increasingly shaped by human activities: reservoir regulation, irrigation abstraction, and land-use changes alter the natural rainfall-runoff relationship. In low-yield catchments, these factors can reduce the streamflow observed at the outlet to a small fraction of the precipitation input (Di Baldassarre et al., 2013). Physics-based models remain central to many operational warning systems, but their real-time application is constrained by computational demands and the need for high-resolution hydro-geomorphological data. In catchments with low runoff coefficients or weak rainfall-runoff signals, uncertainties in boundary conditions and parameterization further limit their reliability (Vieux et al., 2004; Nayak et al., 2005; Hosseiny et al., 2020; Cloke and Pappenberger, 2009). Data-driven frameworks, particularly gradient boosting methods such as XGBoost, have therefore become practical alternatives for reproducing nonlinear, continuous hydrological dynamics. While such models effectively represent average flow conditions, accurate forecasting extreme peak discharges remains a significant challenge, especially in low-yield catchments where extreme-flow observations are sparse (Brunner et al., 2021).

Gradient boosting architectures have become standard tools for operational flow simulation in flood warning services because they model complex, nonlinear hydro-climatic relationships directly from observed data (Ren et al., 2024; Akinsoji et al., 2024). By sequentially correcting prediction errors, these ensembles capture higher-order interactions among hydro-meteorological variables more effectively than single-learner approaches (Wu et al., 2024). Comparative evaluations in hydrology have generally reported that ensemble methods outperform individual learners in high-dimensional predictor spaces (Yu et al., 2020; Tyrallis et al., 2021; Zounemat-Kermani et al., 2021). However, these capabilities do not automatically extend to the upper tail of the flow distribution. When extreme events provide only a limited number of training samples, tree-based ensembles tend to overfit the abundant low- and median-flow conditions (Brunner et al., 2021). Reproducing rare exceedances, such as Q_{99} flood peaks, in a true predictive mode therefore exposes a structural constraint of standard training objectives: because conventional loss functions minimize global error, optimization is biased toward the dominant median-flow behavior

of the record and systematically marginalizes extreme peak responses. For a warning service, this bias is not purely a statistical artifact; it results in missed flood events, whose economic and social consequences are disproportionately large compared to the costs of false alarms—a disparity that, given the inherent fallibility of every link in the warning chain, directly limits the loss-reduction benefits that such systems are designed to deliver (Merz et al., 2010; Parker and Priest, 2012).

Reliable estimation of extreme outflows remains difficult in systems with complex or attenuated hydrological behavior (Ali et al., 2025). Recent studies reinforce this point: Zhang et al. (2025) document substantial shifts in flood-season extremes in the Jinsha River Basin associated with climate change and the operation of large reservoir groups, illustrating how regulated and otherwise altered systems distort flow distributions that warning models must learn. Tree-based algorithms have also been successfully applied to flood susceptibility mapping across the diverse climatic regions of Iran (Mirzaei et al., 2021), although such spatial assessments do not address the temporal detection of extreme exceedances that operational services require. Beyond these physical and statistical challenges, the credibility of data-driven warning tools depends on how they are validated. Standard machine learning workflows often employ randomized data partitioning, which mixes daily streamflow observations between training and testing phases, disrupts chronological dependencies, and causes future-to-past data leakage. Performance inflated by such leakage is particularly misleading for rare Q_{99} exceedances. For water agencies, over-optimistic skill scores translate into unwarranted confidence in operational alerts (Plate, 2002). Therefore, rigorous evaluation requires chronological partitioning and strictly antecedent inputs, that is, predictor sets from which same-day observations are excluded, guaranteeing a genuine operational lead time, with expanding or sliding validation windows that preserve the temporal integrity of the flow record.

Comparative evaluations of machine learning models for daily streamflow prediction consistently report systematic underestimation of extreme flood peaks, with errors reaching 30–50% of the observed peak magnitude (Zhang et al., 2026). Objective-function experiments conducted across 273 catchments have similarly attributed this high-flow underestimation to squared-error training objectives (Lin et al., 2024b). The underlying mechanism is the heteroscedasticity of high-flow residuals: because residual variance increases with discharge magnitude, a uniform training loss allocates most of its gradient budget to the abundant low-flow observations, leaving errors at

extreme magnitudes comparatively under-penalized (Toth, 2016). While synthetic minority oversampling techniques (SMOTE) are prevalent in static spatial susceptibility mapping, where flood and non-flood samples may legitimately be balanced prior to training (Li et al., 2026), their application to continuous time series is physically inconsistent. Generating synthetic discharge values would violate natural temporal autocorrelations and mass-balance constraints. Therefore, bias correction must be embedded within the training objective through asymmetric residual penalties rather than through prior data manipulation.

Modifications to the training objective have been proposed to correct this statistical imbalance. Laimighofer et al. (2022) embedded an expectile-based penalty into XGBoost to reduce the under-prediction of extremely low flows, while multi-objective formulations that include explicit flood-deviation terms have been reported to improve extreme-event accuracy in reservoir inflow forecasting (Jia et al., 2024). Systematic experiments across 273 watersheds further demonstrate that squared-error-based objectives compromise low-flow accuracy because high-magnitude samples dominate the gradient (Lin et al., 2024a). However, such single-tier asymmetry merely reallocates fit quality between the targeted tail and the bulk of the record. For water agencies, this trade-off is operational rather than merely statistical: the continuous baseline underpins low-flow management and environmental-flow assessment, whereas the upper tail governs flood alerts. Three gaps remain unresolved in the literature. First, existing weighting schemes are single-tier: they up-weight extreme observations but lack mechanisms to simultaneously protect baseflow stability. Second, weighting strategies have been evaluated in isolation, without systematic comparison across contrasting hydrological regimes. Third, reporting of algorithmic stability is inconsistent; most studies present a single tuned configuration without quantifying whether reported gains are reproducible across stochastic realizations. No existing study addresses these three limitations simultaneously within a unified, regime-conditional framework.

LightGBM serves as a computationally efficient alternative to XGBoost for continuous runoff simulation, a crucial advantage when models must be retrained within operational time constraints (Bentéjac et al., 2021; Xu et al., 2023). Evidence from data-scarce watersheds indicates that systematic hyperparameter optimization improves predictive accuracy compared to untuned configurations (Tadayon et al., 2025). However, when the target is the highly imbalanced records of rare extreme events, standard tuning procedures struggle to navigate the resulting high-

dimensional hyperparameter spaces (Andonie, 2019). Systematic benchmarking of six hyperparameter optimization strategies for flood modeling has confirmed the superiority of metaheuristic and stochastic search and shown that the choice of optimizer significantly affects predictive accuracy (Razavi-Termeh et al., 2026). Integrating metaheuristic optimization with gradient boosting has become common practice in natural hazard assessment (Bordbar et al., 2022; Khajavi et al., 2025; Ajin et al., 2025; Vu et al., 2026; Liu et al., 2025). In this study, the Tree-structured Parzen Estimator implemented in Optuna serves as a preliminary filter to narrow the parameter bounds, followed by two swarm intelligence algorithms: Particle Swarm Optimization (PSO), retained as a classical benchmark (Kennedy and Eberhart, 1995), and the Nutcracker Optimization Algorithm (NOA), selected for its memory-based search diversity (Abdel-Basset et al., 2023). Both algorithms calibrate the localized penalty weights of extreme observations and their influence within the global objective function. Thus, the optimizer's search space extends from model architecture to the penalty structure itself-encoding a trade-off between hazard detection and baseline reliability that operational practice sets by agency policy as much as by statistical criteria.

Beyond the selection of parameters and objective functions, the statistical reliability of tuned models governs whether they can be trusted in operation. Comprehensive reviews have noted that reliability reporting remains inconsistent, which restricts operational adoption (Abozweita et al., 2026; Zounemat-Kermani et al., 2021). Comparative assessments in water resources applications have documented that gradient boosting models, when insufficiently regularized, tend to overfit the training data and lose accuracy on unseen datasets (Alsulamy et al., 2025). Equally important for deployment is the algorithmic stability of the tuning pipeline itself: because metaheuristic optimizers initialize randomly, a single execution cannot demonstrate that reported detection gains are reproducible rather than artifacts of a favorable random seed-reproducibility that water agencies require before trusting any warning statistic. Therefore, this study quantifies algorithmic robustness through twenty independent stochastic realizations of each configuration, reporting stability envelopes for both continuous and categorical performance metrics.

The novelty of this study lies in a dual-tier weighting framework that simultaneously optimizes two conflicting operational goals—rare-event detection and baseline stability—and in a regime-conditional configuration protocol that prescribes the appropriate weighting scenario for each

hydrological regime. Unlike prior single-tier approaches that sacrifice one service to protect the other, the proposed framework isolates extreme peaks without compromising acceptable baseflow accuracy. Building on recent advances in multi-stage optimization (Abozweita et al., 2026; Ludyawati et al., 2026), the methodology is tested across two contrasting physical environments in northern Iran: the naturally responsive Haraz basin and the attenuated, low-yield Tajan basin. The contributions are threefold. First, a regime-conditional evaluation identifies where additional complexity is warranted. Second, the collapse of variance-based objectives under extreme asymmetric weighting in attenuated basins is documented, with residual-penalty objectives identified as a stable alternative. Third, the detection–false-alarm trade-off is interpreted in operational terms. From a water resources management perspective, the results provide a decision-support protocol: given a basin's runoff coefficient and degree of anthropogenic attenuation, agencies can select the appropriate weighting scenario, thereby avoiding the trial-and-error tuning that currently delays operational deployment. Crucially, whereas one-day-ahead inflow forecasting in regulated Iranian basins has relied on numerical weather prediction inputs (Tanhapour et al., 2025), the present framework operates strictly on lagged meteorological observations and delivers its 24-hour operational lead time without dependence on external forecast products. Across all configurations, twenty independent stochastic runs ensure that reported gains are robust for practical application. The next section describes the study area, data, and the detailed methodological framework used to address the three identified gaps.

2- Study Area and Data

The study focuses on five adjacent river basins located in the central and eastern regions of Mazandaran Province, northern Iran: Haraz, Babolrood, Talar, Tajan, and Nekarood. Their geographic locations within the province and the country are illustrated in Figure 1. Due to the steep topography, the area exhibits considerable variation in rainfall-runoff processes, and the hydrological responses of neighboring catchments often differ markedly. These basins were pre-screened to identify catchments with contrasting hydrological response behaviors.

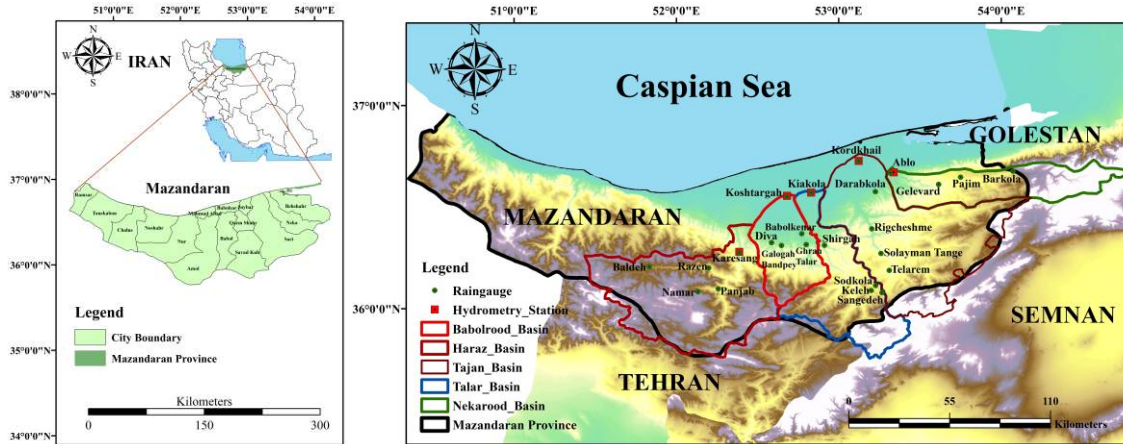


Figure 1. Location and hydro-meteorological network of the five screened basins in Mazandaran Province, Iran.

2.1. Physical characteristics of the screened basins

To define the physical setting of each basin, key morphometric and hydrological descriptors (i.e., drainage area, main channel slope, mean basin elevation, runoff coefficient, and time of concentration) were retrieved from topographic and hydrological data (Table 1). These variables served as the basis for distinguishing between flow regimes of responsive and attenuated types in the following step.

Table 1. Morphometric and hydrological characteristics of the five candidate basins in Mazandaran Province, Iran.

Basin	Hydrometric Station	Basin Area (Km ²)	Basin Slope (%)	Main Canal Slope (%)	Main Canal Length (Km)	Time of Concentration (Hr)
Haraz	Karesang	3987	47	3.72	138.1	16.89
Babolrood	Koshtargah	1625	23.4	1.60	74.4	11.45
Talar	Kiakola	2387	37.3	1.43	126.4	17.04
Tajan	Kordkhail	4027	29.4	1.07	134.5	18.04
Nekarood	Ablo	1905	25.4	1.46	45.22	12.97

To quantify the water-yielding capacity of each candidate basin and identify the most contrasting hydrological regimes, we evaluated the long-term mean annual runoff coefficient (RC). Serving as an integrated proxy for the combined physical influences of climate, topography, and land cover on streamflow generation, the RC was calculated as the ratio of mean annual runoff depth to mean annual precipitation:

$$RC = R/P \quad (1)$$

where R is the mean annual runoff depth (mm) and P is the mean annual areal precipitation (mm) over the historical observation period. The resulting RC values for the five basins are presented in

Table 2. The Haraz basin had the highest coefficient ($RC = 0.57$), whereas the Tajan basin had the lowest ($RC = 0.082$). These two basins, representing the upper and lower bounds of the observed RC range, were therefore selected for detailed analysis.

Table 2. Mean annual runoff coefficients (RC) of the five candidate basins.

Basin	Haraz	Babolrood	Talar	Tajan	Nekarood
RC	0.571	0.272	0.121	0.082	0.085

2.2. Hydrological regimes of the selected basins

The Haraz Basin ($RC = 0.57$) exhibits a responsive rainfall-runoff regime. Streamflow closely follows precipitation patterns, with pronounced peaks in winter and spring and relatively low baseflow during summer. The annual flow hydrograph spans a wide range of magnitudes, from sustained baseflow to short-lived flood peaks, reflecting the basin's steep topography and limited flow regulation (Yahyazadeh Shourabi et al., 2025).

The Tajan Basin ($RC = 0.082$), by contrast, exhibits a markedly attenuated hydrological regime. The rainfall-runoff signal is weak, and the flow hydrograph is dominated by low-magnitude discharges with infrequent and damped peak responses. High-flow events are rare, and the overall flow variability is substantially lower than that of the Haraz Basin.

2.3. Hydrometeorological data and feature engineering

Daily hydrometeorological time series were compiled for both basins using data from the regional meteorological and hydrometric networks operated by the Mazandaran Regional Water Company. The Haraz basin record spans from 1976 to 2024, comprising more than 11,000 daily observations. The Tajan basin record covers the period from 1970 to 2016, with over 14,000 daily observations. The spatial distribution of the monitoring stations is shown in Figure 1.

In the Haraz basin, daily precipitation is recorded at five stations: Baladeh, Namar, Razan, Panjab, and Karesang. Baladeh and Namar are located in high-elevation zones, while Razan and Panjab are situated at lower altitudes. Daily streamflow is measured at the Karesang hydrometric station, which also serves as a rain gauge and is located at the natural outlet of the catchment.

In the Tajan basin, precipitation is recorded at five stations: Telarem, Solayman Tange, Rigcheshme, Darabkola, and Kordkhail. These stations span from the mountainous headwaters to the coastal lowlands. Streamflow is measured at the Kordkhail hydrometric station, located in the low-gradient floodplain near sea level. The geographical coordinates, elevations, and type of all stations in both basins are listed in Table 3.

Table 3. Geographical coordinates, elevations, and station types of the hydrometeorological stations.

Station	Elevation (m)	Latitude (D-M-S)	Longitude (D-M-S)	Basin	Type
Namar	2120	36-06-00	52-05-00	Haraz	Rain gauge
Baladeh	1875	36-11-32	51-47-52	Haraz	Rain gauge
Razan	1240	36-11-47	52-10-49	Haraz	Rain gauge
Panjab	920	36-05-50	52-16-14	Haraz	Rain gauge
Karesang	220	36-18-00	52-21-59	Haraz	Rain gauge/ Hydro station
Telarem	1090	36-12-02	53-14-54	Tajan	Rain gauge
Solayman Tange	500	36-14-30	53-15-09	Tajan	Rain gauge
Rigcheshme	223	36-22-31	53-10-31	Tajan	Rain gauge
Darabkola	140	36-33-41	53-15-06	Tajan	Rain gauge
Kordkhail	-18	36-42-36	53-06-17	Tajan	Rain gauge/ Hydro station

Descriptive statistics of daily streamflow at the basin outlet and spatially averaged daily precipitation are summarized in Table 4, including minimum, maximum, mean, standard deviation, coefficient of variation (CV), Hurst exponent, and the 98th and 99th percentiles.

Table 4. Descriptive statistics, Hurst exponents, and extreme quantiles of daily streamflow and spatially averaged precipitation.

Variable	Basin	Count	Min	Max	Mean	Std	CV	Hurst	Perc98	Perc99
Q (m ³ /s)	Haraz	11092	0	164	27.38	17.44	0.64	0.89	80.99	87.40
P (mm)	Haraz	11092	0	38	1.31	2.97	2.27	0.64	11.38	14.88
Q (m ³ /s)	Tajan	14059	0	410.73	12.08	20.09	1.66	0.84	67.98	89.85
P (mm)	Tajan	14059	0	45.84	2.05	3.49	1.70	0.62	12.99	16.14

The coefficient of variation of daily streamflow in the Tajan basin is approximately three times higher than in the Haraz basin, and both records exhibit positive skewness. The maximum recorded discharge in the Haraz basin is approximately six times the historical mean, whereas the corresponding ratio in the Tajan basin is approximately 34. These statistics confirm the high degree of imbalance between median-flow and extreme-flow observations in both basins.

The Hurst exponent (H) was calculated to quantify temporal persistence. All series exhibited long-range dependence ($H > 0.5$). The streamflow series showed $H = 0.89$ and 0.84 for the Haraz and Tajan basins, respectively, compared to $H = 0.64$ and 0.62 for the corresponding precipitation series. The higher persistence in discharge relative to precipitation indicates that each catchment acts as a hydrological filter, and that antecedent conditions carry information not contained in same-day rainfall alone.

The predictor space was organized into three groups:

(I) Station-specific precipitation with one-day and two-day lags (P_{t-1} , P_{t-2}) from each of the five rain gauges. Same-day precipitation was excluded to ensure a minimum 24-h operational lead time.

(II) Basin-scale antecedent moisture, including spatially averaged precipitation and cumulative totals over rolling windows of 3, 10, 30, and 90 days prior to day t .

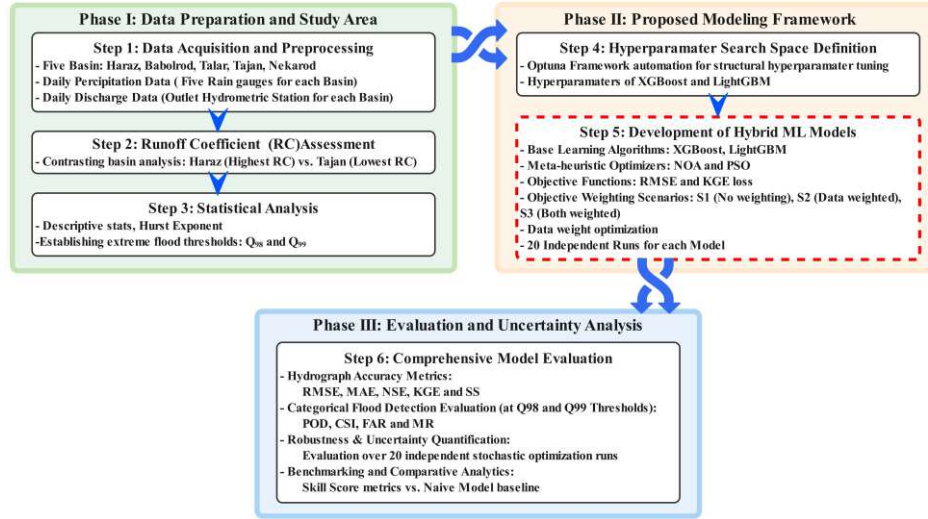
(III) Autoregressive and cyclical variables: previous-day streamflow (Q_{t-1}), sine and cosine transform of the day of the year ($DOY_{\sin}$, $DOY_{\cos}$), and a discrete month index.

For each basin, the first 80% of the chronological record was allocated to model training and hyperparameter optimization, and the remaining 20% was reserved as an independent testing set. Partitioning followed strict chronological order to prevent future-to-past data leakage.

3- Methodology

The methodological workflow, summarized in Figure 2, proceeds in three phases. Phase I covers hydrometeorological data screening, temporal synchronization, descriptive and persistence statistics, and the definition of extreme-flow thresholds (Q_{98} and Q_{99}). Phase II constructs the predictive architecture in two stages: Bayesian search-space reduction (Stage 1), followed by metaheuristic optimization of the dual-weighting structure under three weighting scenarios (Stage 2). Phase III assesses performance with continuous and categorical metrics, benchmarks skill

258 against a naïve baseline, and quantifies robustness through twenty independent stochastic runs of
 259 each configuration.



260

261 Figure 2. Methodological workflow from hydrometeorological data preparation to predictive model evaluation and
 262 robustness assessment.

263 3.1. Base Machine Learning Architectures

264 3.1.1. Extreme Gradient Boosting (XGBoost)

265 Extreme Gradient Boosting (XGBoost) utilizes a sequential ensemble of classification and
 266 regression trees (CARTs) to map complex nonlinear relationships (Chen and Guestrin, 2016). For
 267 a hydrological dataset comprising n daily observations and m engineered features, the final
 268 streamflow prediction $\hat{y}_i$ is computed through the additive expansion of K trees:

$$269 \quad \hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad f_k \in F \quad (2)$$

270 where F denotes the functional space of all possible CARTs. To mitigate overfitting during the
 271 continuous mapping of complex hydro-meteorological thresholds, XGBoost minimizes a
 272 regularized objective function (L):

$$273 \quad L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (3)$$

where l is a differentiable convex loss function measuring the discrepancy between predicted and observed streamflow values, and:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (4)$$

$\Omega(f_k)$ is the regularization penalty that restrains tree complexity based on the number of leaves (T) and leaf weights (w). This structural constraint provides XGBoost with substantial generalization power over traditional tree ensembles when forecasting highly non-linear hydrological states.

3.1.2. Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) serves as the second base learner (Ke et al., 2017). Two design choices distinguish it from level-wise boosting. First, trees grow leaf-wise, splitting whichever leaf yields the largest loss reduction, so a given accuracy is reached with fewer leaves. Second, Gradient-based One-Side Sampling retains the instances with the largest gradients and randomly down-samples the remainder, with a correction factor applied to the retained proportions so that the gradient distribution is preserved. In daily streamflow records, the largest gradients typically arise from poorly reproduced high-flow events; this sampling therefore keeps rare extremes such as Q_{99} peaks influential during training while the abundant low-gradient baseflow days are thinned. Exclusive Feature Bundling further merges mutually exclusive predictors, reducing the dimensionality of the engineered input space. LightGBM has been reported as a computationally efficient alternative for continuous runoff simulation (Bentéjac et al., 2021; Xu et al., 2023), and that efficiency matters here because each configuration is re-trained twenty times under the stability protocol of Section 3.5.

3.2. Evaluation Metrics and Objective Functions

Model performance was evaluated using three categories of metrics to capture both continuous hydrograph dynamics and discrete extreme events: continuous statistical errors, categorical threshold indices, and relative skill scores.

3.2.1. Continuous Hydrological Metrics and Objective Functions

300 The continuous predictive accuracy of the models was evaluated using four standard statistical
 301 criteria: Root Mean Square Error (RMSE), Weighted RMSE, Mean Absolute Error (MAE), Nash-
 302 Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency (KGE) and Weighted KGE, calculated as
 303 follows:

$$304 \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Q_{\text{obs},i} - Q_{\text{sim},i})^2} \quad (5)$$

$$305 \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |Q_{\text{obs},i} - Q_{\text{sim},i}| \quad (6)$$

$$306 \quad \text{NSE} = 1 - \frac{\sum_{i=1}^n (Q_{\text{obs},i} - Q_{\text{sim},i})^2}{\sum_{i=1}^n (Q_{\text{obs},i} - \bar{Q}_{\text{obs}})^2} \quad (7)$$

$$307 \quad \text{KGE} = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (8)$$

$$308 \quad \text{Weighted RMSE} = \sqrt{\frac{\sum_{i=1}^n w_i (Q_{\text{obs},i} - Q_{\text{sim},i})^2}{\sum_{i=1}^n w_i}} \quad (9)$$

$$309 \quad \text{Weighted KGE} = \sqrt{(r_w - 1)^2 + (\alpha_w - 1)^2 + (\beta_w - 1)^2} \quad (10)$$

310 where $Q_{\text{obs},i}$ and $Q_{\text{sim},i}$ are the observed and simulated discharge at time i , respectively; n is the total
 311 number of observations; r represents the Pearson correlation coefficient; α is the ratio of simulated
 312 to observed standard deviations and β is the ratio of simulated to observed means.

313 To intensify the model's sensitivity toward extreme flood events, both objective functions were
 314 reformulated using a continuous flow-proportional weighting coefficient (w_i), which assigns
 315 progressively greater weights to timesteps with higher observed discharges. The weighted
 316 statistical components (r_w , α_w , β_w) embedded within the Weighted KGE, along with the complete
 317 derivation of w_i , are provided in Appendix A.

While all four metrics were used for general model evaluation, RMSE and KGE act as the primary objective functions during the optimization procedure (Stage II) of the metaheuristic search. From a mathematical perspective, RMSE is less tolerant of high-magnitude errors due to the squared term, which made it particularly responsive to peak flows. Meanwhile, KGE provides an optimal compromise for the joint control on the correlation, variability, and mean biases. By introducing such two objective functions, it can be seen that the proposed dual-weighting framework effectively constrains the optimization procedure to avoid the underestimation of extreme flows in the tail of the distribution.

3.2.2. Categorical Flood Warning Metrics

Continuous performance measures should assess overall variability, but they can mask predictive errors during rare, but intense, peaks in flows. To evaluate how well the algorithms identify such extreme events, a contingency table analysis was performed for specific high-flow quantiles (the 98th and 99th percentiles). Forecasts of continuous values that exceeded the thresholds were converted into binary outcomes to calculate the Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI):

$$POD = \frac{H}{H+M} \quad (11)$$

$$FAR = \frac{FA}{H + FA} \quad (12)$$

$$CSI = \frac{H}{H + M + FA} \quad (13)$$

Here, H (Hits) is the number of correctly predicted extreme threshold exceedances; M (Misses) are the observed extreme events that were not predicted; and FA (False Alarms) are the predicted extremes that did not occur. The best possible predictive model would have $POD = 1$, $CSI = 1$, and $FAR = 0$.

3.2.3. Relative Performance and Skill Scores

To quantify the predictive advantage of the machine learning algorithms relative to a baseline forecast, the Skill Score (SS) was calculated. The SS measures the algorithmic performance gain

of each optimized model compared to a naive reference baseline. In this study, the evaluation was applied to the RMSE metric (SS_{RMSE}) across different extreme-flow thresholds, formulated as:

$$SS_{RMSE} = (1 - \frac{RMSE_{Model}}{RMSE_{Naive}}) \times 100 \quad (14)$$

where $RMSE_{Model}$ is the error of the implemented gradient boosting scenario, and $RMSE_{Naive}$ represents the error of the naive baseline projection. A positive SS_{RMSE} indicates that the trained machine learning model yields superior forecasting accuracy compared to the naive assumption, validating the model's actual predictive intelligence under normal and extreme flow conditions.

3.3. Stage 1: Search Space Reduction via Bayesian Optimization

Boosting algorithms expose a large set of interdependent hyperparameters, and exhaustive search over this space is computationally prohibitive when every evaluation re-trains the model on a multi-decadal daily record. Stage 1 therefore used the Tree-structured Parzen Estimator (TPE) implemented in the Optuna library (Akiba et al., 2019) as a preliminary filter, not as the final tuning step. Its role was to discard unpromising regions of the initial bounds and return a reduced, focused domain to the metaheuristic stage that follows.

TPE is a sequential model-based sampler: it maintains two density estimators over the hyperparameter space, one built from the best trials and one from the remainder, and draws new candidates where the ratio of the two densities is highest (Bergstra et al., 2011). This approach has been applied to XGBoost-based flood estimation (Jarajapu et al., 2022), though typically as the sole tuning step. In this study, the sampler was directed at the unweighted baseline RMSE (Scenario S_1), so that the reduced bounds reflect the intrinsic structure of the rainfall–runoff mapping rather than any particular weighting scheme.

Random K-fold cross-validation would break the chronological ordering of the daily series and allow future information to influence the validation folds. The Optuna trials therefore used a time-series split with an expanding training window that advances strictly forward in time, so that every trial was scored only on observations later than its training period. This keeps the tuning phase consistent with the chronological partitioning described in Section 2.3.

Two hundred trials were run for each basin and each base learner, giving 800 Bayesian trials in total. Optuna's fANOVA importance evaluator (Hutter et al., 2014) then ranked the hyperparameters by their contribution to the loss, and those with negligible influence were fixed at their default values. The retained variables govern tree architecture, learning rate, and row and column sampling; their initial and reduced bounds are listed in Table 5. These reduced bounds initialized the metaheuristic optimization of Stage 2 (Section 3.4).

3.4. Stage 2: The Metaheuristic and Weighting Framework

Using the reduced parameter bounds from Stage 1, Stage 2 applies two swarm intelligence algorithms to tune the remaining hyperparameters and the weighting parameters. Particle Swarm Optimization (PSO) serves as the classical baseline, while the Nutcracker Optimization Algorithm (NOA) (Abdel-Basset et al., 2023) is selected for its memory-based search mechanism, which helps maintain diversity in complex spaces. Both algorithms simultaneously calibrate the tree hyperparameters and the penalty weights defined in the following scenarios.

3.4.1. Nutcracker Optimization Algorithm (NOA)

To mitigate the risk of premature convergence in the complex hyperparameter space, the Nutcracker Optimization Algorithm (NOA) was employed as the primary metaheuristic engine (Abdel-Basset et al., 2023). NOA mathematically simulates the foraging and spatial memory strategies of Clark's nutcrackers. Rather than relying on standard velocity-based updates, its search mechanism is partitioned into two distinct algorithmic stages: a global exploration phase (modeled as food hoarding) to broadly sample the parameter space, and an intensive localized exploitation phase (modeled as cache recovery) to finely tune the optimal solutions.

The positional updating of candidate hyperparameter solutions relies on a dynamic memory matrix that archives previously discovered high-performing configurations. During the global exploration phase, the position vector (X_i) of the i -th search agent at the next iteration ($t+1$) is fundamentally governed by a reference-based spatial update protocol:

$$X_i^{new} = X_i + r_1 \cdot (X_{best} - X_i) + r_2 \cdot (RP_j - X_k) \quad (15)$$

where X_{best} represents the elite configuration, RP_j is a reference position extracted from the cache memory matrix, and r_1, r_2 are randomized search scaling coefficients. This unique mathematical

reliance on dynamic memory storage enables NOA to systematically navigate the highly complex and discontinuous search terrains associated with the dual-weighted multi-objective fitness function. By preserving the diversity of populations and keeping the spatial boundaries of promising solutions from converging prematurely, NOA is particularly good at avoiding local errors and finding the absolute global optimum for extreme flood predictions.

3.4.2. Particle Swarm Optimization (PSO)

PSO is included as a comparative baseline, as it is widely used in hydrological optimization (Cheng and Jin, 2015; Bui et al., 2019). The algorithm updates the trajectory of each particle based on its personal best position (P_{best}) and the swarm's global best position (G_{best}). While computationally efficient, standard PSO is more susceptible to premature convergence in high-dimensional spaces than memory-based algorithms like NOA, providing a useful contrast for evaluating the proposed framework.

In order to obtain an unbiased comparison baseline, both metaheuristics were allowed to use the same computational budget. Specifically, a population size of 20 was used for a maximum of 20 iterations. Since the search space boundaries had been narrowed substantially in Stage I, this budget (a total of 400 function evaluations per run) was sufficient for both algorithms to converge to an acceptable solution without incurring an unrealistically large computational burden. To systematically evaluate the framework's ability to resolve the underestimation of Q_{99} flood peaks, the metaheuristic tuning was executed under three comparative scenarios:

3.4.3. Scenario 1: Baseline (Unweighted)

In the baseline configuration (S_1), the metaheuristics calibrated only the tree hyperparameters identified in Stage 1. All daily streamflow observations were assigned a uniform sample weight ($W_i=1$) during training. This unweighted setup serves as the control experiment, reflecting the default behavior of standard boosting architectures, which tend to optimize primarily for median baseflow conditions.

3.4.4. Scenario 2: Single-tier Weighting (Data-weighted)

Scenario 2 (S_2) introduces a single-tier data weighting mechanism. Daily observations exceeding the extreme flow threshold ($Q_{obs} \geq Q_{99}$) were assigned an amplified sample weight ($W_{data} > 1$),

while non-extreme observations retained a weight of 1. Rather than setting W_{data} to a fixed empirical value, it was treated as a continuous decision variable and optimized by the metaheuristic algorithms. Although this approach increases the gradient contribution of rare peaks, it can induce a trade-off, where improved extreme-event detection is achieved at the expense of baseflow accuracy (Laimighofer et al., 2022). In Scenario 2, the sample weights enter only the training loss; the fitness function guiding the metaheuristics remains the unweighted form of the objective function.

3.4.5. Scenario 3: Dual-tier Weighting Framework (Data and Objective weighted)

To address the trade-offs associated with single-tier approaches, Scenario 3 implements the proposed dual-tier weighting framework. In this configuration, the metaheuristics simultaneously optimize two interconnected components: (i) the structural tree hyperparameters, and (ii) the localized data penalty weight for extreme events (W_{data}).

Scenario 3 (S_3) implements the proposed dual-tier weighting framework. As in Scenario 2, observations exceeding the Q_{99} threshold receive an amplified sample weight (W_{data}) during training, forming the data-level tier. The second tier extends the weighting into the objective functions themselves: candidate configurations are scored with the weighted RMSE of Equation (9) in variant S_3 -RMSE, or with the weighted KGE distance of Equation (10) in variant S_3 -KGE. The two objective functions are never combined, linearly or otherwise. The metaheuristics therefore optimize the tree hyperparameters and W_{data} against a fitness landscape in which extreme events carry explicit weight both in the training loss and in the scoring criterion, while the unweighted metrics of Section 3.2.1 remain the basis for performance reporting.

3.5. Algorithmic Stability Protocol

Metaheuristic optimizers such as PSO and NOA are stochastic: identical configurations can follow different optimization trajectories depending on the random initialization of the search agents (Pasche et al., 2026), so a single execution is not representative. To characterize this run-to-run variability, each configuration (basin, learner, optimizer, objective variant, and weighting scenario) was executed twenty times with independent seeds in the Haraz basin and fifteen times in the Tajan basin, the reduced count reflecting the higher per-run computational cost in Tajan; a supplementary check confirmed that extending the Tajan protocol to twenty runs did not alter the

reported means or spreads. Performance is reported as the mean across runs (Tables 6 and 7), while the run-to-run spread is summarized through medians, interquartile ranges, and min–max envelopes (Figures 4 and 5); together these support the interpretation that the reported gains of the dual-weighting framework reflect the weighting and objective choices rather than favorable random initializations.

During manuscript preparation, the authors used large language model (Qwen3.8-Max) as editorial assistants to refine wording, sentence structure, and scientific tone. All technical content, numerical results, methodological choices, and interpretations were generated, verified, and approved exclusively by the authors, who retain full accountability for the final manuscript.

4- Results

The evaluation compares predictive performance in the two contrasting physical environments introduced in Section 2: the responsive Haraz regime and the attenuated Tajan regime. Following the methodological workflow, the analysis begins with the hyperparameter sensitivity and the search-space reduction achieved in Stage 1, then proceeds to scenario performance and to the stability of the configurations across independent runs.

4.1. Search Space and Hyperparameter Analysis

Figure 3 shows the fANOVA importance scores of the hyperparameters for both base learners, computed from the Stage 1 trials (200 per basin and per learner). The importance rankings differ between the two basins, indicating that the physical regime shapes which parameters control the fit. In the responsive Haraz basin, both models assign the highest importance to parameters governing tree complexity and learning rate (for example, `max_depth` and the learning rate), which is consistent with the need to resolve fast, nonlinear rainfall–runoff transformations. In the attenuated and skewed Tajan basin, importance shifts toward regularization and data-sampling parameters (for example, `min_child_weight` and `subsample`). This pattern suggests that, where the rainfall–runoff signal is weak, the models rely on regularization to avoid overfitting the dominant baseflow rather than on additional structural depth.

In the responsive Haraz basin, the clear rainfall–runoff signal shifted importance toward parameters that control the overall convergence of the boosting process rather than toward internal tree regularization. For LightGBM, the learning rate alone accounted for 81.3% of the variance in

the validation objective, suggesting that, when the physical signal is strong, calibration of the gradient step size matters more than the detailed tree topology. For XGBoost, 59.0% of the total importance was carried by the ensemble size (`n_estimators`), consistent with a reliance on the number of sequential boosting rounds to reproduce the continuous hydrograph.

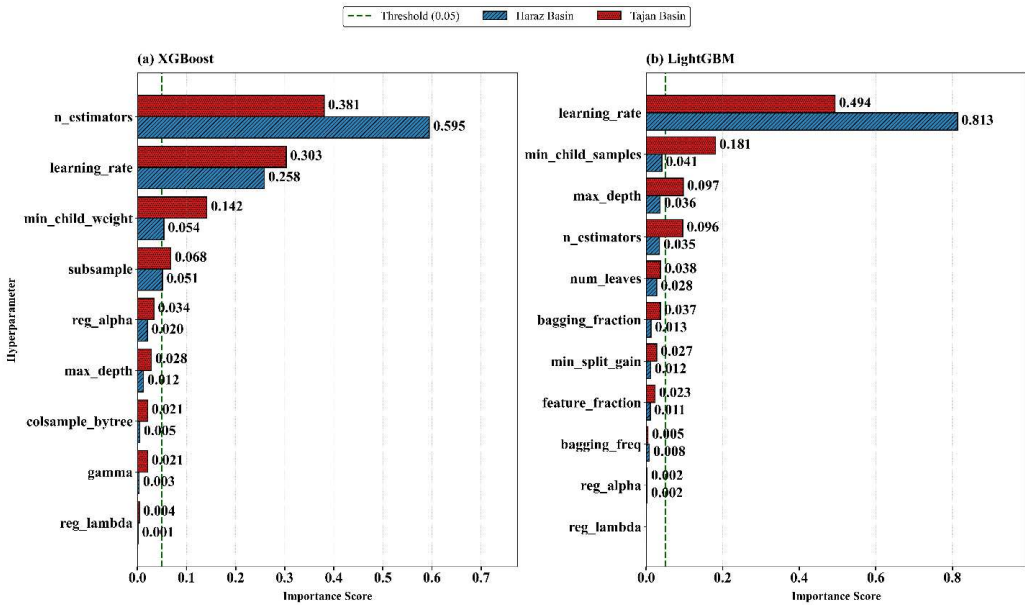


Figure 3. fANOVA hyperparameter importance scores from Stage 1 optimization for (a) XGBoost and (b) LightGBM in the Haraz and Tajan basins.

In the low-yielding Tajan basin ($RC = 0.082$), the importance profile differs markedly. Only about 8% of the basin precipitation appears as streamflow at the outlet; the remainder is lost largely to evapotranspiration and subsurface transfer, and the resulting flow series is dominated by slow, low-magnitude components with rare event responses. Consistent with this weak signal, Stage 1 assigned higher importance to complexity-penalizing parameters: for LightGBM, `min_child_samples` (18.1%) and `max_depth` (9.7%), and for XGBoost, `min_child_weight` (14.2%). This pattern suggests that, in signal-attenuated catchments, the boosting learners rely on leaf-structure regularization to avoid fitting the dominant low-flow noise at the expense of the rare extreme responses.

To stabilize the metaheuristic optimization of Stage 2, a 5% importance threshold was applied (Figure 3). Hyperparameters whose fANOVA score fell below this threshold, such as the regularization coefficients (`reg_lambda`, `reg_alpha`) and the subsampling fractions (`subsample`, `colsample_bytree`, `feature_fraction`, `bagging_fraction`), were held fixed at their best-trial values

from Stage 1. Only the parameters whose importance exceeded the threshold were retained as active search variables and passed to the NOA and PSO algorithms for calibration under the weighting scenarios. The retained bounds and the fixed values are summarized in Table 5.

Table 5. Fixed values and reduced search bounds established for the Stage 2 metaheuristic optimization.

Hyperparameter	Haraz Basin		Tajan Basin	
	XGBoost	LightGBM	XGBoost	LightGBM
learning_rate	0.037 - 0.076	0.017 – 0.138	0.004 – 0.008	0.027 – 0.117
max_depth	3 - 15	4	8 - 10	8 - 15
min_child_weight	15 - 20	-	13 - 15	-
n_estimators	624 - 1000	250	640 - 820	629 - 995
subsample	0.530	-	0.665	-
colsample_bytree	0.950	-	0.657	-
gamma	1.240	-	1.190	-
reg_lambda	0.144	3.240	1.170	0.0000002
reg_alpha	0.065	0.003	0.00015	6.170
feature_fraction	-	0.89 – 1.0	-	0.570
num_leaves	-	146 - 222	-	189
min_child_samples	-	16	-	80 - 97
min_split_gain	-	0.940	-	0.150
bagging_fraction	-	0.790	-	0.570
bagging_freq	-	6	-	9

4.2. Continuous and Extreme-Event Performance

Tables 6 and 7 summarize, for each configuration, the mean of each performance metric across the twenty independent runs described in Section 3.5; the run-to-run spread is presented in Section 4.3. The three weighting scenarios (unweighted baseline, single-tier, and dual-tier) are compared using the continuous hydrograph metrics computed over the full test record (RMSE, MAE, NSE, and KGE) and the categorical extreme-event indices computed at the Q_{98} and Q_{99} thresholds (POD, FAR, and CSI), together with Skill Scores relative to the naïve baseline, reported in Table 6.

Performance differs markedly between the two catchments, in line with their physical contrast. In the responsive Haraz basin ($RC = 0.57$), the unweighted baseline (S_1) already reproduces the main flow dynamics without any weighting scheme: across all learner–optimizer combinations and both

objective functions, $NSE = 0.98$ and $KGE = 0.96-0.97$, with RMSE between 2.65 and 2.74 m³/s (Table 6).

Table 6. Continuous performance metrics (means over twenty runs) for all model configurations and weighting scenarios.

Model		XGBoost-PSO (RMSE)					LightGBM-PSO (RMSE)				
Metric		RMSE	MAE	NSE	KGE	SS	RMSE	MAE	NSE	KGE	SS
Basin	Scenario										
Haraz	S ₁	2.65	1.32	0.98	0.96	0.22	2.68	1.27	0.98	0.96	0.21
	S ₂	2.55	1.31	0.98	0.97	0.28	2.56	1.23	0.98	0.97	0.28
	S ₃	2.57	1.32	0.98	0.97	0.27	2.57	1.25	0.98	0.97	0.27
Tajan	S ₁	9.38	4.20	0.62	0.63	0.50	9.13	3.71	0.64	0.73	0.53
	S ₂	9.38	4.18	0.62	0.64	0.50	9.15	3.72	0.64	0.73	0.53
	S ₃	10.99	4.96	0.48	0.55	0.32	9.93	4.52	0.58	0.69	0.44
Model		XGBoost-NOA (RMSE)					LightGBM-NOA (RMSE)				
Metric		RMSE	MAE	NSE	KGE	SS	RMSE	MAE	NSE	KGE	SS
Basin	Scenario										
Haraz	S ₁	2.65	1.31	0.98	0.96	0.23	2.68	1.26	0.98	0.96	0.21
	S ₂	2.55	1.30	0.97	0.97	0.28	2.55	1.23	0.98	0.97	0.28
	S ₃	2.56	1.32	0.97	0.97	0.27	2.59	1.27	0.98	0.97	0.26
Tajan	S ₁	9.38	4.19	0.62	0.63	0.50	9.14	3.72	0.64	0.73	0.53
	S ₂	9.38	4.17	0.62	0.64	0.50	9.22	3.78	0.63	0.73	0.52
	S ₃	11.09	5.06	0.47	0.54	0.30	10.17	4.47	0.55	0.67	0.41
Model		XGBoost-PSO (KGE loss)					LightGBM-PSO (KGE loss)				
Metric		RMSE	MAE	NSE	KGE	SS	RMSE	MAE	NSE	KGE	SS
Basin	Scenario										
Haraz	S ₁	2.68	1.33	0.98	0.96	0.21	2.74	1.30	0.98	0.96	0.17
	S ₂	2.70	1.37	0.98	0.97	0.20	2.90	1.38	0.97	0.97	0.07
	S ₃	2.71	1.39	0.98	0.97	0.19	2.88	1.39	0.98	0.97	0.08
Tajan	S ₁	9.54	3.89	0.61	0.69	0.49	9.24	3.82	0.63	0.74	0.52
	S ₂	9.71	3.96	0.59	0.68	0.47	9.27	3.84	0.63	0.73	0.51
	S ₃	11.98	5.39	0.38	0.47	0.19	12.91	6.16	0.28	0.47	0.05
Model		XGBoost-NOA (KGE loss)					LightGBM-NOA (KGE loss)				
Metric		RMSE	MAE	NSE	KGE	SS	RMSE	MAE	NSE	KGE	SS
Basin	Scenario										
Haraz	S ₁	2.70	1.34	0.98	0.97	0.20	2.71	1.28	0.98	0.96	0.19
	S ₂	2.71	1.40	0.98	0.97	0.19	2.98	1.41	0.97	0.97	0.02
	S ₃	2.67	1.36	0.98	0.97	0.22	2.98	1.43	0.97	0.97	0.02
Tajan	S ₁	9.53	3.89	0.61	0.69	0.49	9.23	3.82	0.63	0.74	0.52
	S ₂	9.63	3.95	0.60	0.69	0.48	9.29	3.85	0.63	0.73	0.51
	S ₃	11.94	5.35	0.39	0.47	0.20	12.98	6.26	0.27	0.47	0.04

As Table 6 shows, the baseline configurations already achieve high continuous accuracy in the Haraz basin ($NSE = 0.98$, $KGE = 0.96-0.97$), and the dual-tier scenario (S₃) alters these values only marginally. Detection skill at Q₉₉ likewise changes little, with POD rising from 0.77-0.84 in S₁ to 0.83-0.85 in S₃ while CSI and FAR remain within 0.70-0.77 and 0.10-0.13 (Table 7). This pattern suggests that, in a responsive catchment where high-flow events are comparatively frequent

in the record, conventional loss functions already receive sufficient extreme-flow information, and the added complexity of dual weighting yields little additional benefit.

Table 7. Extreme-event detection metrics at the Q_{98} and Q_{99} thresholds (means over twenty runs) for all configurations and scenarios.

Model		XGBoost-PSO (RMSE)						LightGBM-PSO (RMSE)					
Metric		POD		CSI		FAR		POD		CSI		FAR	
Basin	Scenario	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}
Haraz	S_1	0.92	0.82	0.85	0.74	0.09	0.11	0.91	0.77	0.83	0.70	0.10	0.11
	S_2	0.93	0.85	0.84	0.76	0.10	0.12	0.92	0.83	0.82	0.75	0.11	0.12
	S_3	0.93	0.85	0.84	0.76	0.10	0.12	0.93	0.84	0.83	0.75	0.11	0.12
Tajan	S_1	0.26	0.22	0.22	0.20	0.44	0.25	0.35	0.20	0.25	0.17	0.51	0.49
	S_2	0.27	0.23	0.22	0.21	0.44	0.26	0.35	0.19	0.26	0.16	0.51	0.51
	S_3	0.45	0.33	0.29	0.22	0.56	0.58	0.43	0.40	0.28	0.27	0.56	0.55
Model		XGBoost-NOA (RMSE)						LightGBM-NOA (RMSE)					
Metric		POD		CSI		FAR		POD		CSI		FAR	
Basin	Scenario	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}
Haraz	S_1	0.92	0.84	0.85	0.75	0.08	0.12	0.91	0.78	0.83	0.83	0.10	0.11
	S_2	0.93	0.83	0.85	0.75	0.10	0.13	0.92	0.83	0.83	0.83	0.11	0.13
	S_3	0.93	0.84	0.85	0.75	0.10	0.12	0.92	0.83	0.82	0.82	0.11	0.13
Tajan	S_1	0.27	0.21	0.22	0.20	0.44	0.26	0.35	0.20	0.26	0.26	0.51	0.48
	S_2	0.27	0.22	0.23	0.20	0.44	0.27	0.34	0.20	0.25	0.25	0.54	0.52
	S_3	0.45	0.33	0.28	0.22	0.57	0.60	0.45	0.40	0.29	0.29	0.55	0.57
Model		XGBoost-NOA (KGE)						LightGBM-NOA (KGE)					
Metric		POD		CSI		FAR		POD		CSI		FAR	
Basin	Scenario	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}
Haraz	S_1	0.92	0.87	0.86	0.77	0.08	0.12	0.91	0.81	0.82	0.73	0.10	0.12
	S_2	0.93	0.84	0.82	0.75	0.12	0.12	0.95	0.84	0.78	0.75	0.19	0.13
	S_3	0.93	0.85	0.83	0.76	0.11	0.12	0.96	0.83	0.77	0.74	0.20	0.13
Tajan	S_1	0.34	0.29	0.25	0.22	0.51	0.50	0.37	0.25	0.27	0.20	0.51	0.52
	S_2	0.34	0.29	0.24	0.22	0.52	0.51	0.36	0.22	0.25	0.17	0.53	0.55
	S_3	0.45	0.38	0.23	0.21	0.67	0.67	0.53	0.54	0.26	0.24	0.66	0.70
Model		XGBoost-PSO (KGE loss)						LightGBM-PSO (KGE loss)					
Metric		POD		CSI		FAR		POD		CSI		FAR	
Basin	Scenario	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}	Q_{98}	Q_{99}
Haraz	S_1	0.92	0.84	0.85	0.75	0.09	0.12	0.91	0.81	0.83	0.73	0.10	0.12
	S_2	0.93	0.85	0.83	0.76	0.12	0.12	0.94	0.84	0.78	0.75	0.18	0.12
	S_3	0.94	0.84	0.83	0.75	0.12	0.13	0.95	0.83	0.78	0.74	0.18	0.13
Tajan	S_1	0.33	0.29	0.25	0.22	0.51	0.52	0.37	0.25	0.27	0.20	0.52	0.50
	S_2	0.35	0.29	0.25	0.21	0.53	0.53	0.38	0.23	0.27	0.18	0.51	0.54
	S_3	0.45	0.39	0.24	0.21	0.67	0.69	0.54	0.53	0.26	0.24	0.67	0.70

In the attenuated Tajan basin, detection is far harder. Under the unweighted baseline (S_1), POD at Q_{99} remains between 0.20 and 0.29 across configurations, indicating that most extreme events are missed when the natural class imbalance is left unaltered. The dual-tier scenario (S_3) raises POD at Q_{99} to 0.33-0.40 under the RMSE objective and to 0.39-0.54 under the KGE objective, with the highest value (0.54, LightGBM-NOA, KGE objective) accompanied by FAR = 0.70 and CSI =

0.24-0.26 (Table 7). The accompanying rise in false alarms is expected in a low-yield basin, where amplifying the weight of rare peaks shifts the detection behavior toward more triggered events. For operational early warning, missed floods carry higher socio-economic costs than false alarms, so a POD-oriented trade-off remains defensible despite the elevated FAR; this asymmetry is examined further in Section 5.3.

The choice of optimizer changed the results only slightly. Across both catchments, the absolute differences between PSO and NOA in KGE and POD did not exceed 0.03, and were below 0.02 in all but one configuration (LightGBM, Tajan, S₃ under the KGE objective; Tables 6 and 7). This parity suggests that, after the search-space reduction of Stage 1, the constrained bounds left both swarm-intelligence algorithms with a landscape simple enough to reach comparable configurations. Optimizer choice is therefore not a decisive factor here: the differences induced by the weighting scenarios and the objective function are several times larger.

This trade-off exposes a limitation of variance-based objectives under strong data imbalance. In Tajan under S₃ with the KGE objective, continuous accuracy deteriorates sharply: NSE falls to 0.27-0.28 for LightGBM and 0.38-0.39 for XGBoost, and the RMSE-based Skill Score drops to 0.04-0.05 and 0.19-0.20, respectively (Table 6). The mechanism is structural: KGE aggregates correlation, variability, and bias ratios over the whole record, so assigning large weights to a few extreme days steers the optimization toward the weighted moments and away from the unweighted balance that the reported metrics measure. The same weighting under the RMSE objective keeps NSE at 0.47-0.58, confirming that the deterioration is tied to the objective function rather than to the weighting itself, and within each learner the pattern held for both optimizers. Dual weighting with the KGE objective therefore buys the highest detection (POD 0.53-0.54 at Q₉₉) at the cost of continuous fidelity, whereas the RMSE objective preserves a workable balance.

Under the RMSE objective, the S₃ configurations in Tajan pair this detection gain with acceptable continuous accuracy: for LightGBM, NSE 0.55-0.58 with POD 0.40, FAR 0.55-0.57, and CSI 0.27 at Q₉₉; XGBoost keeps NSE 0.47-0.48 with POD 0.33 and CSI 0.22 (Tables 6 and 7). Detection is slightly lower than under the KGE objective (POD 0.53-0.54), but the continuous hydrograph is preserved. Within the dual-weighting framework, the residual-penalty objective therefore offers the more balanced operating point for low-yield basins.

Across the two catchments, the metrics point to different preferred configurations. In Haraz, XGBoost optimized by NOA under the single-tier scenario with the RMSE objective (S₂-RMSE) offers the best balance: NSE = 0.98 and KGE = 0.97 (Table 6), with POD 0.83 at Q₉₉ (Table 7), matching the dual-tier scenario without its added complexity. Because the two optimizers differed by no more than 0.02–0.03 in this basin, NOA is retained as the representative engine for the stability analysis of Section 4.3.

In Tajan, the best-performing configuration is LightGBM optimized by NOA under the dual-tier scenario with the RMSE objective (S₃-RMSE): NSE = 0.55 with POD 0.40, FAR 0.57, and CSI 0.27 at Q₉₉ (Tables 6 and 7). This accepts a moderate false-alarm rate in exchange for roughly doubling the detection of the unweighted baseline, while keeping the continuous hydrograph usable, in contrast to the KGE-guided variant, whose NSE falls to 0.27–0.28 as reported above.

Taken together, the two catchments favor different configurations, S₂ in the responsive Haraz and S₃ in the attenuated Tajan, which suggests that the weighting strategy should follow the catchment's yield and memory characteristics rather than fixed a priori. The implications for operational warning design and the transferability of the framework are discussed in Section 5.

4.3. Algorithmic Stability Across Runs

Following the protocol of Section 3.5, the run-to-run stability of the two selected configurations (Haraz: XGBoost-NOA, S₂-RMSE; Tajan: LightGBM-NOA, S₃-RMSE) was examined. Because the training time per run is much higher in Tajan than in Haraz, owing to the longer and more irregular flow record, the stability protocol was applied with twenty independent runs in Haraz and fifteen in Tajan. Preliminary checks confirmed that extending the Tajan runs from 15 to 20 did not change the mean or spread of the metrics in any configuration. Figure 4 shows, for the RMSE objective, the mean convergence trajectory, the best run, and the min-max band across the independent runs. Within the allotted budget the mean trajectory approaches a plateau and the band narrows, indicating that the computational budget defined in Section 3.4.2 is sufficient for convergence in both basins. Reporting the full band rather than a single execution makes the variability induced by the random initialization of the search agents explicit.

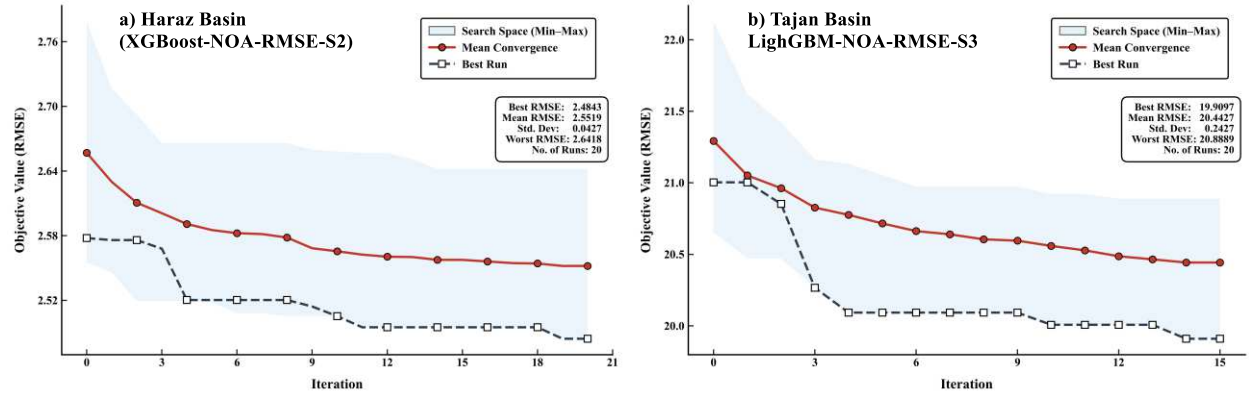


Figure 4. Convergence of the NOA objective (RMSE) for the two selected configurations, showing the mean trajectory, the best run, and the min–max band across the independent runs (20 in Haraz, 15 in Tajan): (a) Haraz, XGBoost-NOA, S₂-RMSE; (b) Tajan, LightGBM-NOA, S₃-RMSE. Insets report the best, mean, standard deviation, and worst objective values.

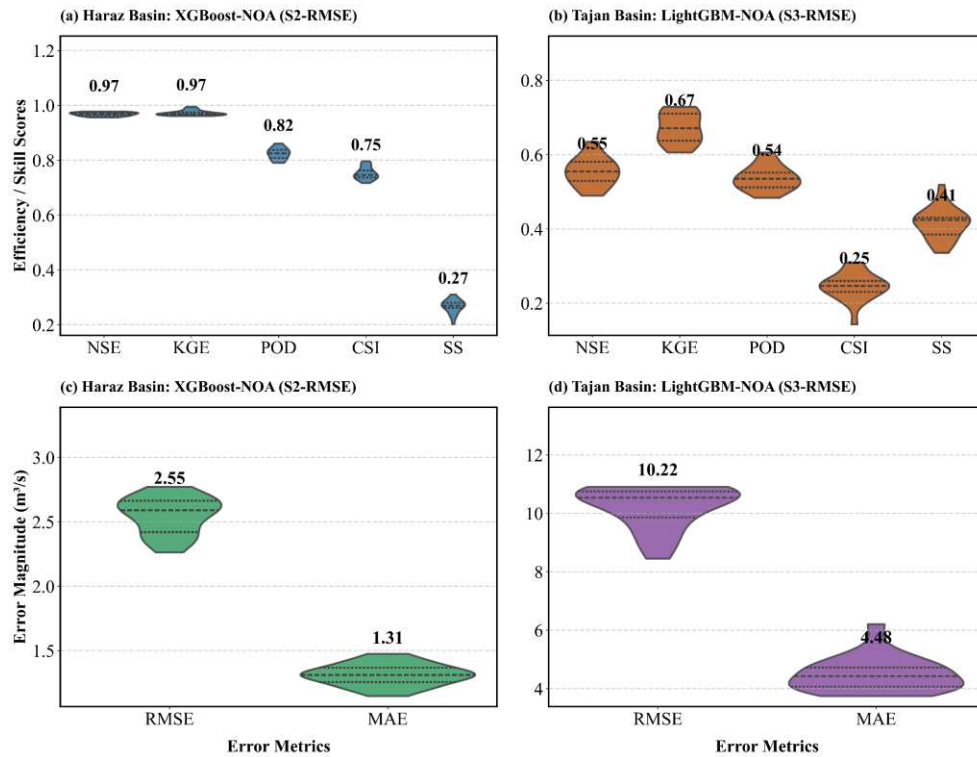


Figure 5. Run-to-run distributions of the continuous and categorical metrics for the two selected configurations, shown as violin plots with median and interquartile lines: (a, c) Haraz, XGBoost–NOA, S₂–RMSE (20 runs); (b, d) Tajan, LightGBM–NOA, S₃–RMSE (15 runs). Annotated values report the [median / mean] across runs.

Figure 5 summarizes these run-to-run distributions as violin plots, in which the dashed lines mark the median and the interquartile range and the local width of each shape traces the empirical density of the metric.

The distributions in Figure 5 separate the two basins clearly. In Haraz (panels a and c), the selected XGBoost-NOA configuration (S_2 -RMSE) shows narrow run-to-run spreads: medians of 0.97 for both NSE and KGE with tight interquartile ranges, and a median RMSE of 2.55 m³/s across the twenty runs. With a strong rainfall-runoff signal ($RC = 0.57$), the optimizer therefore reaches solutions of comparable predictive quality under different random initializations. These plots establish the stability of predictive performance; they do not imply that all runs converge to identical parameter vectors.

In Tajan (panels b and d), the same configuration family shows wider run-to-run spreads, consistent with the weaker and more irregular signal of this basin. The medians remain close to the tabulated means (NSE 0.55 against 0.58; RMSE 10.22 m³/s against 10.17 m³/s; MAE 4.48 against 4.43 m³/s), but the interquartile ranges of KGE and the Skill Score are visibly broader than in Haraz, indicating that individual runs land on solutions of more variable quality. This is the expected behavior in a low-yield catchment, where a small number of extreme events dominates the objective and modest changes in the fitted parameters shift the event-scale errors. Even so, the lower tails of the NSE and KGE violins stay well above the failure levels of the KGE-guided variant (NSE 0.27-0.28), so the selected configuration is stable enough to support the comparative conclusions of Section 4.2.

Figure 6 condenses the comparison into a single multi-metric profile per basin. At both thresholds, the Haraz configuration (XGBoost-NOA, S_2 -RMSE) traces a polygon close to the outer edge on all five axes, confirming that single-tier weighting already yields a balanced performance in a responsive catchment. The Tajan configuration (LightGBM-NOA, S_3 -RMSE) traces a smaller, uneven polygon: moderate NSE and KGE, contracted POD and CSI, and a reduced (1-FAR). Read together, the two polygons show what the weighting framework buys and what it costs in a low-yield basin, and frame the management implications discussed in Section 5.

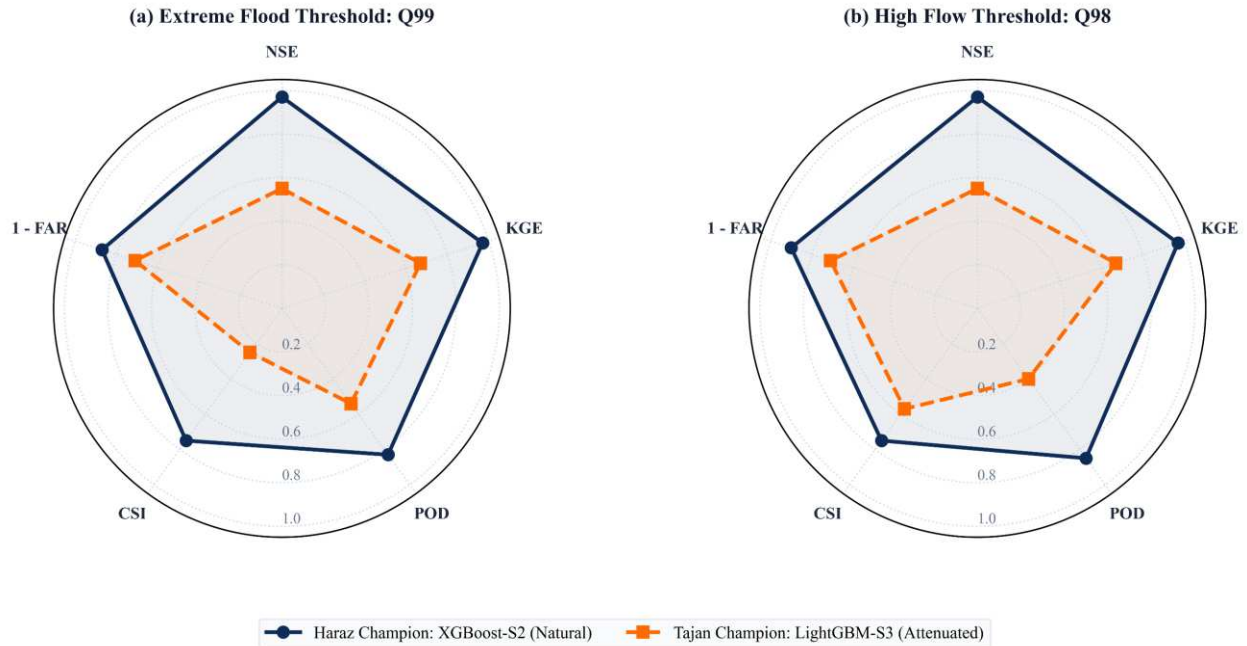


Figure 6. Multi-metric profiles of the selected configurations at the (a) Q99 and (b) Q98 thresholds: Haraz, XGBoost–NOA, S2–RMSE; Tajan, LightGBM–NOA, S3–RMSE. NSE and KGE are whole-record values; FAR is plotted as $(1 - \text{FAR})$ so that all axes are outward-maximizing.

5. Discussion

5.1. Physical hydrology and model complexity

The sensitivity profiles of Stage 1 link the physical runoff generation of each catchment to the structural needs of the boosting models. In the responsive Haraz basin ($\text{RC} = 0.57$), importance concentrated on global convergence parameters, above all the learning rate. In high-yield mountainous catchments with rapid rainfall-runoff response and seasonal snowmelt, the hydrological signal is coherent although nonlinear, so the ensemble can reduce residuals without deep internal partitioning. Responsive catchments also exhibit strong temporal persistence (Koutsoyiannis, 2002; Montanari and Koutsoyiannis, 2012), which supplies consistent learning gradients and reduces the need for heavy regularization. Tree-based models remain piecewise-constant and cannot extrapolate beyond the trained range; the good performance in this regime is therefore supported by the relatively rich record of high-flow events, which keeps predictions inside the observed target space.

In the attenuated Tajan basin ($RC = 0.082$) the optimization landscape shifts. In lowland and regulated catchments of this type, transmission losses, evapotranspiration, and anthropogenic abstractions, in the Tajan basin including regulation by the Shahid Rajaee Dam and irrigation withdrawals, damp the amplitude of flow variations. Days above Q_{99} form about 1% of the record by construction, so the learning task is heavily imbalanced. Consistently, Stage 1 favored parameters that restrict split complexity, namely `min_child_samples` for LightGBM and `min_child_weight` for XGBoost, which limit overfitting to the dominant low-flow class. This behaviour is in line with reports of reduced data-driven skill in catchments with weak or skewed signals (Kratzert et al., 2019; Nearing et al., 2021). For water-resources practice, the implication is direct: in low-yield, regulated basins the effective modelling investment lies in weighting and regularization strategies rather than in deeper ensembles.

The stability and cost results of Section 4.3 complete this physical picture. In Haraz, the coherent signal yields a smooth objective landscape: convergence plateaus early (Figure 4) and the run-to-run spreads of the efficiency metrics remain tight (Figure 5), with twenty runs completed at low computational cost. In Tajan, each run required markedly more time, which limited the protocol to fifteen runs; a supplementary check confirmed that extending the protocol to twenty runs did not alter the reported means or spreads. The wider interquartile ranges in Tajan reflect a rougher landscape shaped by a small number of dominant events, while the difference between PSO and NOA remained within 0.02-0.03 for KGE and POD, making the optimizer a secondary decision once the search space is reduced. For operational deployment, therefore, the computational budget of such frameworks is set by the low-yield, regulated basins, and savings are better sought in search-space reduction and weighting design than in the choice of metaheuristic.

5.2. Selection of Objective Functions

The dual-weighting experiments (S_3) expose a clear limitation in the choice of optimization loss. KGE is widely used because it evaluates correlation, variability ratio, and bias ratio separately (Gupta et al., 2009), and its variability component explicitly tracks how well the simulated spread matches the observed one (Kling et al., 2012). When large weights are assigned to the upper 1% of flows (Q_{99}), however, the weighted moments seen by the objective are dominated by those few days, and the variability ratio the optimizer tries to match no longer represents the unweighted record. The optimizer then sacrifices the fit over the remaining days to satisfy the weighted ratios,

which is consistent with the collapse observed in Tajan: under the KGE objective, NSE falls to 0.27 for LightGBM-NOA, whereas the RMSE objective keeps it at 0.55 for the same configuration. This behavior adds to the documented sensitivity of distribution-matching metrics to outliers and skewed error distributions (Alonso Pizarro and Jorquera, 2024) and argues against their use as objectives under aggressive asymmetric weighting.

Under the same weighting, the RMSE objective remains stable. It does not enforce distributional matching; it accumulates squared residuals, so large errors on the weighted days receive proportionally larger penalties while the unweighted days continue to contribute their ordinary residuals. In this sense RMSE acts as a targeted penalty mechanism: extreme peaks are captured while low-flow performance is largely preserved. The comparison supports a practical guideline rather than a universal rule: KGE remains well suited to unweighted continuous simulation, whereas residual-based objectives are the safer companion for asymmetric extreme-value weighting. Moving to a residual-based objective does not remove the detection-false-alarm trade-off (POD versus FAR, Section 4.2), but for agencies configuring operational flood-forecast chains in low-yield basins it offers an operationally defensible design: rare extremes are detected without losing the continuous baseline on which water-allocation and reservoir-operation decisions rely.

5.3. Flood detection and early-warning implications

The dual-weighting framework in signal-attenuated catchments is an operational compromise between baseline accuracy and extreme-event detection. In Tajan, S_3 approximately doubled the POD of Q_{99} floods relative to the unweighted baseline, at the cost of a corresponding rise in FAR. This behavior is rooted in the physics of low-yield catchments: with RC below 0.10, most precipitation is absorbed by deep percolation, evapotranspiration, transmission losses, and agricultural abstractions and reservoir regulation (Di Baldassarre et al., 2013), so even substantial rainfall events often fail to produce streamflow peaks. When the loss function assigns large weights to the few historical events that did generate extreme runoff, the fitted models become markedly more responsive to rainfall inputs. In effect, the learner treats a larger share of major rainfall events as flood-generating, under-representing the catchment's buffering storage, which is precisely why detection rises together with false alarms.

An elevated FAR does degrade categorical efficiency, yet within operational flood-risk management the trade-off is defensible: warning systems are designed to limit false negatives,

because the consequences of a missed flood exceed the costs of a false alarm (Roulin, 2007; Cloke and Pappenberger, 2009). This principle matters most for downstream communities in regulated, low-yield catchments, including those of the Tajan basin (Fouladi Semnan and Ostad Mirza Tehrani, 2026). The unweighted models behaved as bulk-flow estimators, reproducing the dominant low flows while missing most extremes, a pattern consistent with the systematic underestimation of high flows reported for data-driven models (Zhang et al., 2026). This vulnerability is particularly pronounced in northern Iranian basins, where recent national-scale assessments indicate that flood dynamics are heavily governed by specific geomorphological constraints (e.g., watershed perimeter and length), making them highly susceptible to such underestimation unless explicitly weighted for extremes (Kanani-Sadat et al., 2024). The dual-weighting framework does not remove this difficulty, but with a 24-h lead time it raises the probability that critical thresholds are detected in time for catchment-scale warnings and reservoir pre-release decisions, shifting the learning priority from bulk-fit precision toward the hazards that warning systems exist to capture. This aligns with recent advances in data-scarce, semi-arid catchments, where hybrid frameworks integrating Bayesian optimization and machine learning have proven essential for reducing false alarm rates and enhancing climate-resilient flood risk management (Jan Ahmadi et al., 2026).

5.4. Limitations and future work

Despite the improved detection of extreme events, several methodological limitations remain. The weighting scheme and the objective functions operate on statistical residuals only, so the class imbalance is addressed mathematically rather than mechanistically. Without explicit constraints on catchment retention and transmission dynamics, the primary drivers of attenuation in low-yield systems, the models overcompensate for rare peaks, which is a primary contributor to the elevated FAR observed in testing. A second limitation is the reliance on point gauge precipitation, which cannot represent the spatial variability of rainfall, particularly in mountainous terrain. A third is that the evidence base comprises two contrasting basins; whether the reported regime-dependent behavior holds elsewhere remains to be tested. In addition, the observed flow series of the Tajan basin embeds reservoir operation, and without release records the framework cannot separate the natural from the regulated component of the response.

To mitigate these limitations, future architectures could move toward physics-informed machine learning. Embedding mass-balance constraints or soil-moisture deficit thresholds into the loss function would keep simulated peaks within physically plausible ranges, for example suppressing flood predictions when antecedent storage is far from saturation. Coupling the framework with causal-inference tools could additionally encode routing delays and cause-effect dependencies of attenuated systems. Beyond model structure, future work should test the transferability of the weighting framework in other low-yield, regulated basins and evaluate spatially distributed remote-sensing inputs (soil moisture, radar precipitation) as a route to reducing false alarms in noise-dominated regimes.

6. Conclusion

Predicting extreme flood events (Q_{99}) across contrasting regimes motivates a shift from generic machine-learning setups toward regime-specific weighting frameworks. In responsive systems with strong rainfall-runoff signals (Haraz, $RC = 0.57$), conventional algorithms already provide sufficient learning gradients: a single-tier XGBoost-NOA configuration achieved high continuous accuracy ($NSE = 0.98$) with tight run-to-run spreads and strong detection at Q_{99} (POD 0.83-0.85, FAR 0.10-0.13), so dual weighting added little value.

In attenuated systems (Tajan, $RC = 0.082$), unweighted models missed most extremes (POD 0.20-0.29). The dual-weighting framework doubled detection, but only when paired with a residual-based objective: under the variance-based KGE objective the weighted moments dominated the fit and continuous accuracy fell to NSE 0.27-0.39, whereas the RMSE objective raised POD to 0.40 while keeping NSE at 0.55 for LightGBM-NOA.

This detection gain comes with more false alarms (FAR 0.57 at Q_{99}), a trade-off that operational warning practice accepts because missed floods cost more than false alarms; run-to-run stability checks confirmed that the reported metrics remain bounded across independent runs. The study therefore supports targeted residual penalties over global variance preservation when asymmetric weights are applied in imbalanced, low-yield catchments, and indicates that the detection-stability trade-off should be set by the catchment's attenuation capacity rather than by default configurations. For water-resources management, the framework offers a configurable path toward early flood warnings and informed pre-release decisions in regulated, data-scarce basins.

References

- Abdel-Basset M, Mohamed R, Jameel M (2023) Nutcracker optimization algorithm: a new nature-inspired metaheuristic algorithm for global optimization and engineering design problems. *Knowledge-Based Systems* 260:110143. <https://doi.org/10.1016/j.knsys.2022.110143>
- Abozweita OA, Sidek L, Sherif M, Fardush Sham FA, El-Shafie A, Najah Ahmed A (2026) A decade of progress in streamflow forecasting: a comprehensive review of single and hybrid machine learning models (2014-2024). *Arch Comput Methods Eng*. <https://doi.org/10.1007/s11831-026-10630-7>
- Ajin RS, Costache R, Bărbulescu A, Fanti R, Segoni S (2025) Flood susceptibility assessment using multi-tier feature selection and ensemble boosting machine learning models. *Water* 17(14):2041. <https://doi.org/10.3390/w17142041>
- Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 2623-2631. <https://doi.org/10.1145/3292500.3330701>
- Akinsoji AH, Adelodun B, Adeyi Q, et al. (2024) Integrating machine learning models with comprehensive data strategies and optimization techniques to enhance flood prediction accuracy: a review. *Water Resour Manage* 38:4735-4761. <https://doi.org/10.1007/s11269-024-03885-x>
- Ali AM, Abdallah M, Mohammadi B, Elzain HE (2025) Three-stage hybrid modeling for real-time streamflow prediction in data-scarce regions. *J Hydrol Reg Stud* 59:102337. <https://doi.org/10.1016/j.ejrh.2025.102337>
- Alonso Pizarro J, Jorquera J (2024) Advancing objective functions in hydrological modelling: integrating knowable moments for improved simulation accuracy. *J Hydrol* 634:131071. <https://doi.org/10.1016/j.jhydrol.2024.131071>
- Alsulamy S, Kumar V, Kisi O, et al. (2025) Enhancing water level prediction using ensemble machine learning models: a comparative analysis. *Water Resour Manage* 39:3995-4014. <https://doi.org/10.1007/s11269-025-04142-5>
- Andonie R (2019) Hyperparameter optimization in learning systems. *J Membr Comput* 1(4):279-291. <https://doi.org/10.1007/s41965-019-00023-0>

793 Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting
794 algorithms. *Artif Intell Rev* 54(3):1937-1967. <https://doi.org/10.1007/s10462-020-09896-5>

795 Bergstra J, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization.
796 In: *Advances in Neural Information Processing Systems* 24, pp 2546-2554

797 Bordbar M, Aghamohammadi H, Pourghasemi HR, Azizi Z (2022) Multi-hazard spatial modeling
798 via ensembles of machine learning and meta-heuristic techniques. *Sci Rep* 12(1):1451.
799 <https://doi.org/10.1038/s41598-022-05364-y>

800 Bui DT, Ngo PTT, Pham TD, Jaafari A, Minh NQ, Hoa PV, Samui P (2019) A novel hybrid
801 approach based on a swarm intelligence optimized extreme learning machine for flash flood
802 susceptibility mapping. *Catena* 179:184-196. <https://doi.org/10.1016/j.catena.2019.04.009>

803 Brunner MI, Slater L, Tallaksen LM, Clark M (2021) Challenges in modeling and predicting floods
804 and droughts: a review. *WIREs Water* 8:e1520. <https://doi.org/10.1002/wat2.1520>

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

808 Cheng R, Jin Y (2015) A social learning particle swarm optimization algorithm for scalable
809 optimization. *Inf Sci* 291:43-60. <https://doi.org/10.1016/j.ins.2014.08.039>

810 Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: a review. *J Hydrol* 375(3-4):613-
811 626. <https://doi.org/10.1016/j.jhydrol.2009.06.005>

812 Di Baldassarre G, Kooy M, Kemerink JS, Brandimarte L (2013) Towards understanding the
813 dynamic behaviour of floodplains as human-water systems. *Hydrol Earth Syst Sci* 17(8):3235-
814 3244. <https://doi.org/10.5194/hess-17-3235-2013>

815 Fouladi Semnan A, Ostad Mirza Tehrani MJ (2026) Fully automated risk and hazard assessment
816 for decision-support (FARHAD): a cloud-native machine-learning framework for worst-case
817 scenario flood-risk mapping via weakline analysis. *Water Resour Manage* 40:426.
818 <https://doi.org/10.1007/s11269-026-04786-x>

819 Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error
820 and NSE performance criteria: implications for improving hydrological modelling. J Hydrol
821 377(1-2):80-91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>

822 Hosseiny H, Nazari F, Smith V, Nataraj C (2020) A framework for modeling flood depth using a
823 hybrid of hydraulics and machine learning. Sci Rep 10:8222. [https://doi.org/10.1038/s41598-020-](https://doi.org/10.1038/s41598-020-65232-5)
824 [65232-5](https://doi.org/10.1038/s41598-020-65232-5)

825 Hutter F, Hoos H, Leyton-Brown K (2014) An efficient approach for assessing hyperparameter
826 importance. In: Proceedings of the 31st International Conference on Machine Learning (ICML
827 2014), PMLR 32(1), pp 754-762

828 Jan Ahmadi M, Saidi S, Khajenoori L (2026) A validated hybrid modelling framework for climate-
829 resilient flood risk management in data-scarce semi-arid catchments. Water Resour Manage
830 40:353. <https://doi.org/10.1007/s11269-026-04603-5>

831 Jarajapu DC, Rathinasamy M, Agarwal A, Bronstert A (2022) Design flood estimation using
832 extreme gradient boosting-based on Bayesian optimization. J Hydrol 613:128341.
833 <https://doi.org/10.1016/j.jhydrol.2022.128341>

834 Jia W, Chen M, Yao H, et al. (2024) Improving sub-daily runoff forecast based on the multi-
835 objective optimized extreme learning machine for reservoir operation. Water Resour Manage
836 38:6173-6189. <https://doi.org/10.1007/s11269-024-03953-2>

837 Kanani-Sadat Y, Safari A, Nasserli M, Homayouni S (2024) A novel explainable PSO-XGBoost
838 model for regional flood frequency analysis at a national scale: exploring spatial heterogeneity in
839 flood drivers. J Hydrol 638:131493. <https://doi.org/10.1016/j.jhydrol.2024.131493>

840 Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T (2017) LightGBM: a highly
841 efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154

842 Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95 -
843 International Conference on Neural Networks. IEEE, pp 1942-1948

844 Khajavi H, Rastgoo A, Masoumi F (2025) Enhanced streamflow forecasting for crisis management
845 based on hybrid extreme gradient boosting model. Iran J Sci Technol Trans Civ Eng 49:5233-
846 5254. <https://doi.org/10.1007/s40996-024-01710-4>

847 Kling H, Fuchs M, Paulin M (2012) Runoff conditions in the upper Danube basin under an
848 ensemble of climate change scenarios. *J Hydrol* 424-425:264-277.
849 <https://doi.org/10.1016/j.jhydrol.2012.01.011>

850 Koutsoyiannis D (2002) The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrol*
851 *Sci J* 47(4):573–595. <https://doi.org/10.1080/02626660209492961>

852 Kratzert F, Klotz D, Shalev G, Klambauer G, Hochreiter S, Nearing G (2019) Towards learning
853 universal, regional, and local hydrological behaviors via machine learning applied to large-sample
854 datasets. *Hydrol Earth Syst Sci* 23:5089-5110. <https://doi.org/10.5194/hess-23-5089-2019>

855 Laimighofer J, Melcher M, Laaha G (2022) Low-flow estimation beyond the mean - expectile loss
856 and extreme gradient boosting for spatiotemporal low-flow prediction in Austria. *Hydrol Earth*
857 *Syst Sci* 26:4553-4573. <https://doi.org/10.5194/hess-26-4553-2022>

858 Lin Y, Wang D, Jiang T, Kang A (2024a) Assessing objective functions in streamflow prediction
859 model training based on the naïve method. *Water* 16:777. <https://doi.org/10.3390/w16050777>

860 Lin Y, Wang D, Zhu J, Sun W, Shen C, Shangguan W (2024b) Development of objective function-
861 based ensemble model for streamflow forecasts. *J Hydrol* 632:130861.
862 <https://doi.org/10.1016/j.jhydrol.2024.130861>

863 Liu Y, Wang H, Guan X, et al. (2025) Urban flood depth prediction and visualization based on the
864 XGBoost-SHAP model. *Water Resour Manage* 39:1353-1375. [https://doi.org/10.1007/s11269-](https://doi.org/10.1007/s11269-024-04020-6)
865 [024-04020-6](https://doi.org/10.1007/s11269-024-04020-6)

866 Ludyawati DI, Kim BH, Kam J (2026) An ensemble learning framework for streamflow
867 reconstruction incorporating aleatoric and epistemic uncertainty. *J Hydrol Reg Stud* 66:103604.
868 <https://doi.org/10.1016/j.ejrh.2026.103604>

869 Merz B, Kreibich H, Schwarze R, Thielen A (2010) Review article "Assessment of economic
870 flood damage". *Nat Hazards Earth Syst Sci* 10:1697-1724. [https://doi.org/10.5194/nhess-10-1697-](https://doi.org/10.5194/nhess-10-1697-2010)
871 [2010](https://doi.org/10.5194/nhess-10-1697-2010)

872 Mirzaei S, Vafakhah M, Pradhan B, Alavi SJ (2021) Flood susceptibility assessment using extreme
873 gradient boosting (EGB), Iran. *Earth Sci Inform* 14:169-180. [https://doi.org/10.1007/s12145-020-](https://doi.org/10.1007/s12145-020-00530-0)
874 [00530-0](https://doi.org/10.1007/s12145-020-00530-0)

875 Montanari A, Koutsoyiannis D (2012) A glimpse into the statistical properties of hydrological
876 processes: from the Hurst phenomenon to the Hurst-Kolmogorov stochastic process. *Hydrol Earth*
877 *Syst Sci* 16(10):3587-3599. <https://doi.org/10.5194/hess-16-3587-2012>

878 Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS (2005) Short-term flood forecasting with a
879 neurofuzzy model. *Water Resour Res* 41:W04004. <https://doi.org/10.1029/2004WR003562>

880 Nearing GS, Kratzert F, Sampson AK, Pelissier CS, Klotz D, Frame JM, Prieto C, Gupta HV
881 (2021) What role does hydrological science play in the age of machine learning? *Water Resour*
882 *Res* 57:e2020WR028091. <https://doi.org/10.1029/2020WR028091>

883 Parker DJ, Priest SJ (2012) The fallibility of flood warning chains: can Europe's flood warnings
884 be effective? *Water Resour Manage* 26:2927-2950. <https://doi.org/10.1007/s11269-012-0057-6>

885 Pasche OC, Lam H, Engelke S (2026) Extreme conformal prediction: reliable intervals for high-
886 impact events. *Extremes* 29:129-155. <https://doi.org/10.1007/s10687-026-00536-9>

887 Plate EJ (2002) Flood risk and flood management. *J Hydrol* 267:2-11.
888 [https://doi.org/10.1016/S0022-1694\(02\)00135-X](https://doi.org/10.1016/S0022-1694(02)00135-X)

889 Razavi-Termeh SV, Sadeghi-Niaraki A, Pourzangbar A, et al. (2026) Exploring the impact of
890 hyperparameter tuning on the accuracy of machine learning-based flood susceptibility mapping.
891 *Water Resour Manage* 40:469. <https://doi.org/10.1007/s11269-026-04827-5>

892 Ren H, Pang B, Bai P, Zhao G, Liu S, Liu Y, Li M (2024) Flood susceptibility assessment with
893 random sampling strategy in ensemble learning (RF and XGBoost). *Remote Sens* 16(2):320.
894 <https://doi.org/10.3390/rs16020320>

895 Roulin E (2007) Skill and relative economic value of medium-range hydrological ensemble
896 predictions. *Hydrol Earth Syst Sci* 11(2):725-737. <https://doi.org/10.5194/hess-11-725-2007>

897 Tadayon A, Nazari M, Kerachian R (2025) Enhancing long-lead rainfall forecasting in data-scarce
898 large watersheds using multi-model fusion. *J Hydrol Reg Stud* 62:102936.
899 <https://doi.org/10.1016/j.ejrh.2025.102936>

900 Tanhapour M, Soltani J, Shakibian H, et al. (2025) Development of a multi-objective optimal
 901 operation model of a dam using meteorological ensemble forecasts for flood control. *Water Resour*
 902 *Manage* 39:2743-2761. <https://doi.org/10.1007/s11269-024-04089-z>

903 Tellman B, Sullivan JA, Kuhn C, Kettner AJ, Do CS, Winsemius HC, et al. (2021) Satellite
 904 imaging reveals increased proportion of population exposed to floods. *Nature* 596:80-86.
 905 <https://doi.org/10.1038/s41586-021-03695-w>

906 Toth E (2016) Estimation of flood warning runoff thresholds in ungauged basins with asymmetric
 907 error functions. *Hydrol Earth Syst Sci* 20:2383-2394. <https://doi.org/10.5194/hess-20-2383-2016>

908 Tyrallis H, Papacharalampous G, Langousis A (2021) Super ensemble learning for daily
 909 streamflow forecasting: large-scale demonstration and comparison with multiple machine learning
 910 algorithms. *Neural Comput Appl* 33:3053–3068. <https://doi.org/10.1007/s00521-020-05172-3>

911 Vieux BE, Cui Z, Gaur A (2004) Evaluation of a physics-based distributed hydrologic model for
 912 flood forecasting. *J Hydrol* 298(1-2):155-177. <https://doi.org/10.1016/j.jhydrol.2004.03.035>

913 Vu QH, Bui QD, Vu DC, Luong ND, Duong CH, Luu C, Ha H (2026) Metaheuristic-enhanced
 914 XGBoost models for improving flood and landslide susceptibility assessment. *Earth Syst Environ*.
 915 <https://doi.org/10.1007/s41748-026-01051-4>

916 Wu Y, Zhang Z, Qi X, Hu W, Si S (2024) Prediction of flood sensitivity based on logistic
 917 regression, eXtreme gradient boosting, and random forest modeling methods. *Water Sci Technol*
 918 89(10):2605–2624. <https://doi.org/10.2166/wst.2024.146>

919 Xu K, Han Z, Xu H, Bin L (2023) Rapid prediction model for urban floods based on a light gradient
 920 boosting machine approach and hydrological-hydraulic model. *Int J Disaster Risk Sci* 14:79-97.
 921 <https://doi.org/10.1007/s13753-023-00465-2>

922 Yahyazadeh Shourabi K, Niksokhan MH, Nikoo MR (2025) Enhancing flood frequency
 923 predictions under climate change and uncertainty using machine learning model fusion and
 924 wavelet transform. *Earth Syst Environ*. <https://doi.org/10.1007/s41748-025-00859-w>

925 Yu Y, Yin J, Zhong H (2020) Improved hydrology simulation using ensemble machine learning
 926 techniques. *J Hydrol* 582:124546. <https://doi.org/10.1016/j.jhydrol.2020.124546>

Zhang Y, Zhang Z, Zhang Q, Zhang X, Xu Y, Liu X, Mao J, Xu C (2025) Cumulative and offsetting effects of streamflow response to climate change and large reservoir group in the Jinsha River Basin, China. *J Hydrol Reg Stud* 60:102480. <https://doi.org/10.1016/j.ejrh.2025.102480>

Zhang Z, Xiao Y, Chen R, Long K, Deng H, Zheng Z, Liao J (2026) Comparative assessment of machine learning models for daily streamflow prediction in a subtropical monsoon watershed. *Sci Rep* 16:7341. <https://doi.org/10.1038/s41598-026-38969-8>

Zounemat-Kermani M, Batelaan O, Fadaee M, Hinkelmann R (2021) Ensemble machine learning paradigms in hydrology: a review. *J Hydrol* 598:126266. <https://doi.org/10.1016/j.jhydrol.2021.126266>

Statements and Declarations:

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Ali Akbar Eatesam led the conceptualization of the research, developed the methodology, curated and prepared the datasets, performed the formal analysis, produced the visualizations, and wrote the original draft of the manuscript. Khosrow Hoseini supervised the research process and contributed to the critical review and editing of the manuscript. Hojat Karami co-supervised the work, validated the analytical results, and participated in the review and editing of successive versions. All authors commented on previous versions of the manuscript, and all read and approved the final version.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)