Hydrology and Water Resources
https://www.sciltp.com/journals/hwr

Article

New Insights in Prediction of Daily River Flow Using SWOT Observations and Machine Learning

Shengyu Zou [1,2], Kiril Manevski [2,3,4], Jingyi Tian [5] and Jing Tian [1,*]

[1] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[2] Sino-Danish College, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China
[4] Department of Agroecology, Aarhus University, 8830 Tjele, Denmark
[5] College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China
* Correspondence: tianj.04b@igsnrr.ac.cn

How To Cite: Zou, S.; Manevski, K.; Tian, J.; et al. New Insights in Prediction of Daily River Flow Using SWOT Observations and Machine Learning. *Hydrology and Water Resources* **2026**, *1*(2), 10. https://doi.org/10.53941/hwr.2026.100010

Received: 9 February 2026
Revised: 22 March 2026
Accepted: 23 March 2026
Published: 13 April 2026

Abstract: Accurate river flow estimation is essential for water resources management, yet continuous *in situ* observations remain scarce across rivers globally and particularly in Asia and Africa. This study explores the potential of integrating SWOT satellite–derived river width and water surface elevation with limited gauging data to reconstruct river flow of China's vast river network. The study compared four machine learning models for data integration: Random Forest (RF), XGBoost (XGB), Multi-Layer Perceptron (MLP), and Transformer (TF). A temporally ordered five-fold cross-validation framework was used to evaluate both interpolating and extrapolating performance. Under interpolating mode, RF and XGB effectively reproduced the observed hydrographs, capturing flow variability and extremes. Under extrapolation mode, all models show reduced skill due to short record length, seasonal incompleteness, and zero-flow effects, although neural network models exhibited relatively better performance. These results demonstrate a potential solution for river flow gap filling by using SWOT satellite observations after appropriate data processing, however, the approach requires substantially larger and more diverse training datasets for improving extrapolation performance.

Keywords: deep learning; interpolation; extrapolation; ensemble tree models; river width; water surface elevation

1. Introduction

Gauges for accurate monitoring and forecasting of river flow are fundamental to global water security, disaster management, and climate change adaptation. However, nearly half of the world's major river systems are ungauged or poorly gauged [1–3], creating a critical monitoring and knowledge gap on how much water flows where and when, limiting effective hydrological modeling and water resource management. This information scarcity situation could be mitigated by the rapid development of remote sensing satellite technologies, which have opened new pathways for estimating streamflow and water levels globally [4–7].

There are many satellites orbiting the Earth, however most are not on a mission dedicated to terrestrial water such as riverine land cover. The long-running Landsat program provides multispectral data for river morphology and water quality [8,9], the Sentinel-1 radar mission enables all-weather river discharge estimation [10,11], and the Gravity Recovery and Climate Experiment Follow-On (GRACE-FO) mission quantifies basin-scale total water storage changes impacting river systems [12]. These missions have advantages and disadvantages, but none can retrieve directly river morphological description needed for computation of water flows and water quality [13]. The launch of the Surface Water and Ocean Topography (SWOT) satellite in 2022/2023 represents a paradigm

shift, offering high-resolution observations of water surface elevation (WSE), width, and slope that are practically usable for satellite-based river flow inversion [14–18]. The main instrument of SWOT is the Ka-band Radar Interferometer (KaRIn) [19], which simultaneously senses water surface elevation and total inundated width of river reaches [4]. This remote sensing "revolution", coupled with significant advances in computational capabilities, accelerates the adoption of Machine Learning (ML) and Deep Learning (DL) methods in streamflow prediction [20–22]. Data-driven models demonstrate superior capability in capturing complex nonlinear relationships and spatiotemporal dependencies inherent in hydrological processes, often surpassing traditional empirical or physically based models [23–25]. Specifically, ML/DL models excel at modeling the nonlinear rainfall–runoff transformation and the stage–river flow relationship [26–29], and advanced architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have shown robust performance in forecasting tasks across diverse basins [30,31]. However, due to the inherent variability in catchment characteristics, data availability, and modeling objectives, there is no universally optimal data-driven approach for river flow modeling, making it imperative to screen multiple algorithms and provide knowledge on robustness, predictive performance and generalizability.

Previous SWOT-ML studies generally follow two methodological directions: hybrid physical data-driven approaches that estimate hydraulic parameters prior to river flow calculation, and direct regression-based models that learn end-to-end mappings from SWOT observables to river flow [18,19,32,33]. Although both strategies are promising, their robustness and generalization under realistic observational constraints remain insufficiently understood. In fact, a fundamental challenge arises from the intrinsic characteristics of SWOT observations involving measurement random noise, erroneous data from heterogeneous river morphologies and temporally sparse sampling imposed by the satellite's orbit cycle of 21 days [34]. These factors complicate the reliable reconstruction of hydrological events, particularly when training data are limited or discontinuous. Moreover, most existing studies rely on case-specific experiments or restricted regional domains, and systematic, station-level evaluations across large, heterogeneous river networks are still scarce. Consequently, the relative strengths and limitations of different ML model families in reconstructing hydrological variability from sparse satellite observations remain unclear.

China has a vast river network and long-term gauge records, with many hydrological stations missing observations, which substantially limits the applicability of traditional calibration-based or physics-constrained modeling approaches. Large-scale satellite-driven regression models could offer a viable pathway for river flow reconstruction, temporal gap filling, and hydrological analysis across diverse climatic and geomorphic conditions. To address these challenges, this study presents advanced regression-based framework for SWOT-driven river flow reconstruction across China's heterogeneous river systems. We conduct a systematic benchmark of four representative ML models-Random Forest (RF), XGBoost (XGB), multilayer perceptron (MLP), and Transformer-selected to span a broad spectrum of model complexity and data efficiency. A five-fold random cross-validation strategy is implemented at the station level to evaluate each model's ability to recover hydrological events and maintain predictive stability under noisy and temporally sparse observation conditions. By focusing on a large number of stations distributed across China, this design enables a rigorous assessment of model robustness and generalizability at the national scale, going beyond basin-specific studies. The overarching goal of this work is to develop a transferable and practical approach for river flow reconstruction in regions lacking open or continuous *in situ* measurements. By integrating SWOT observations with limited hydrological data through a consistent regression and validation framework, this study aims to provide reliable temporal supplementation for intermittently observed stations, support river flow estimation at poorly gauged or intermittently monitored sites, and establish a methodological foundation for large-scale, satellite-based hydrological analyses across China's rivers. The findings demonstrate the feasibility of nation-wide flow reconstruction using SWOT data and a small amount of hydrological station data and provide guidance for operational applications in other data-scarce regions [20].

2. Data and Methods

Figure 1 summarizes the overall methodology. First, a spatiotemporally matched dataset was constructed by integrating SWOT satellite observations with *in situ* hydrological station measurements, which is not a trivial task. Spatial matching was performed using the geographic coordinates of SWOT observation points (Lon_s, Lat_s) and hydrological stations (Lon_h, Lat_h). Temporal alignment was achieved by matching SWOT overpass times (time_s) with hydrological observation times (time_h) within predefined temporal windows. The resulting dataset included remotely sensed river attributes—WSE, river width, and water surface slope—alongside observed river flow (Q_{obs}) from gauging stations. The matched dataset was then organized into a unified sample set and split into training and validation subsets using five-fold cross-validation to ensure robust model evaluation. Machine

learning models—including Random Forest (RF), XGBoost (XGB), Multi-Layer Perceptron (MLP), and Transformer (TF)—were trained to directly estimate river flow (Q_{sim}) from SWOT-derived and *in situ* features [35]. This framework enables systematic assessment of different machine learning approaches for reconstructing river flow from satellite and sparse *in situ* observations, providing a scalable method for supplementing temporal gaps in gauged river records. Detailed procedures are presented in Sections 2.3.1–2.3.4.

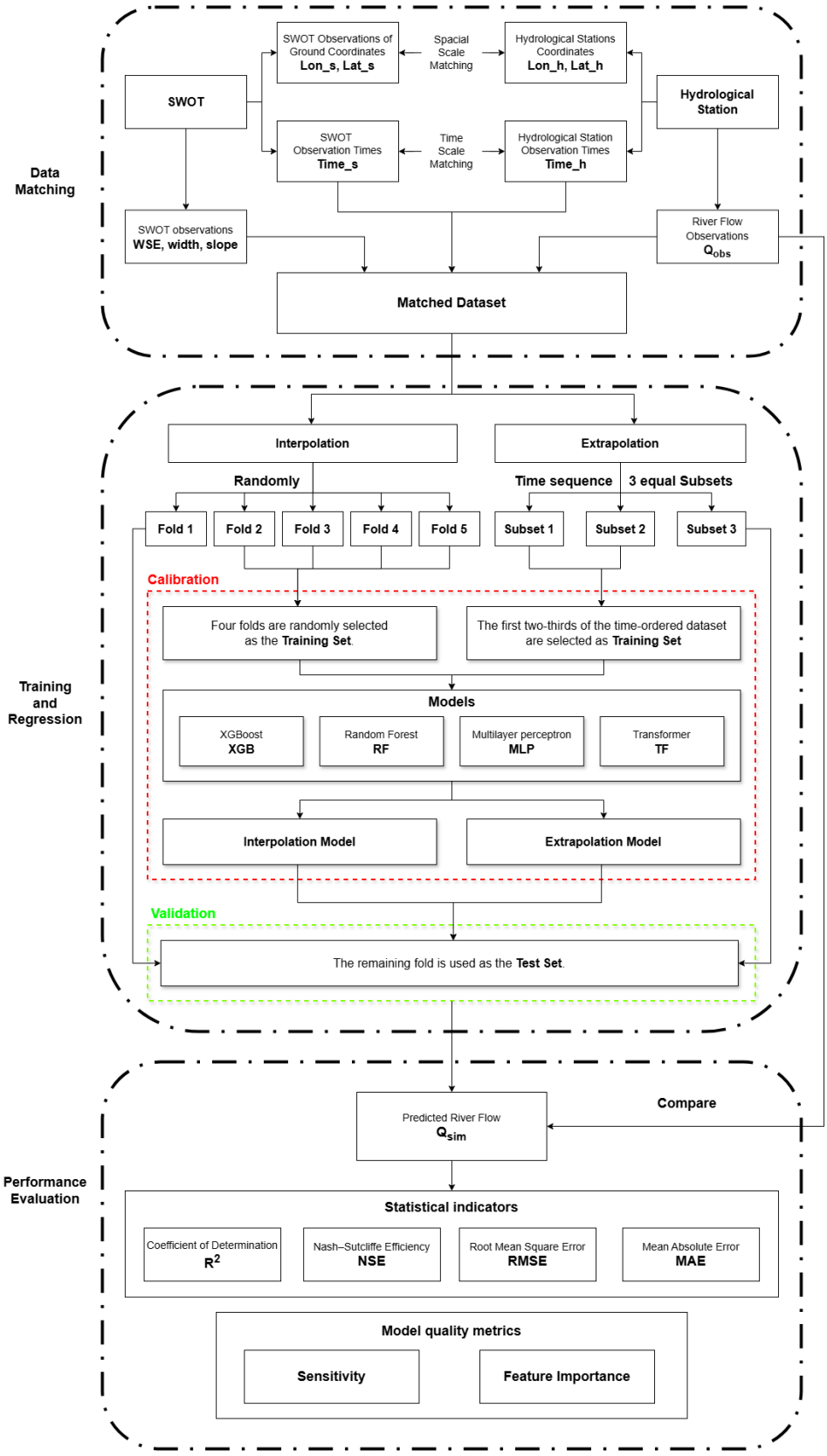

Figure 1. Workflow of spatiotemporal data matching and machine learning–based river flow prediction using SWOT observations.

2.1. SWOT Data

This study uses SWOT Level-2 River Single-Pass Vector Data Product, Version C (SWOT_L2_HR_RiverSP_2.0) at 21-day resolution and covering April 2023 to March 2025. Derived from the KaRIn measurements, SWOT product provides WSE, river width, and slope for both reaches and nodes. The data, originally distributed as ESRI Shapefiles through NASA's PO.DAAC, were converted to CSV via the Hydrocron API for streamlined processing. Within the Chinese domain, the SWORD database identifies 23,485 reaches and 553,279 nodes (Figure 2), and only SWOT observations with quality flags reach_q and node_q equal to 0 or 1 were retained to ensure reliability.

Complementing the satellite dataset, an extensive *in situ* hydrological dataset was compiled from 1448 major gauging stations, yielding 1,483,887 measurements of water level and river flow collected from 2019 to 2024, although the temporal overlap with the SWOT observations was from 2023 to 2024. These stations are predominantly located in China's central and southern river basins, including the Yangtze, Yellow, Pearl, Huai, and Hai Rivers, while coverage is notably sparse in the rest of the country (Figure 2). Additionally, cross-sectional information for 22 river locations was obtained from the China Sediment Bulletin, of which 14 sites were suitable for analysis.

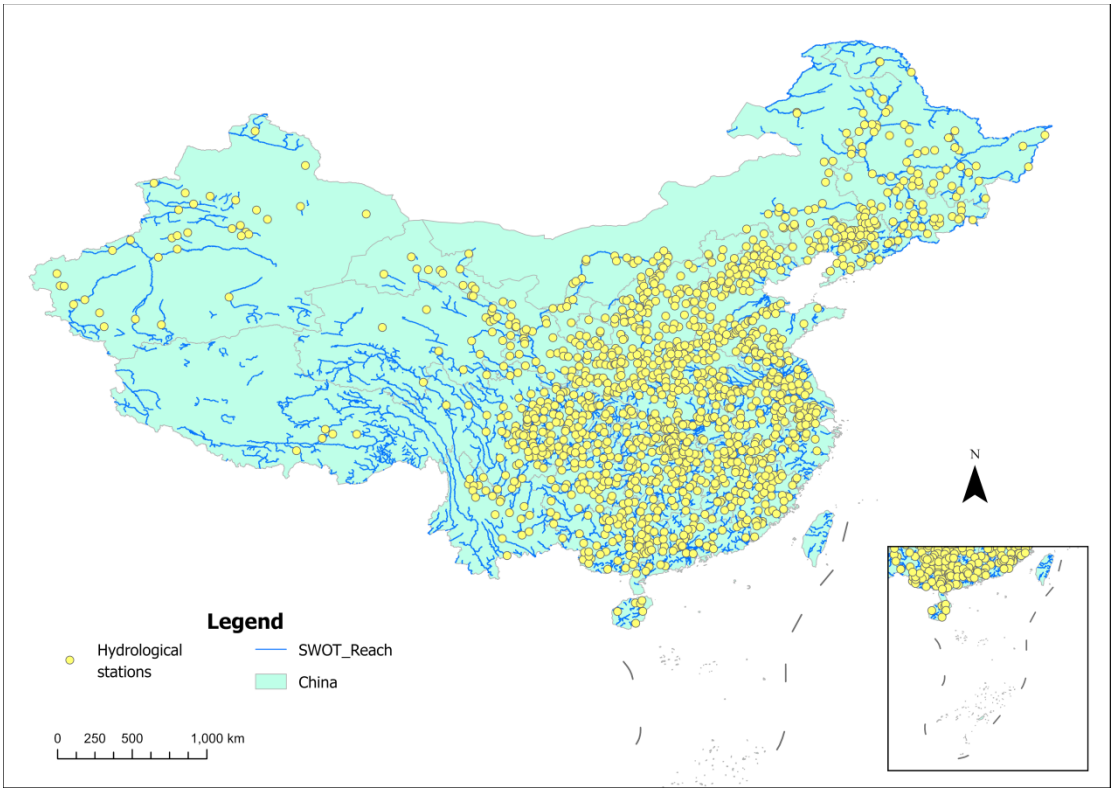

Figure 2. Geographical distribution of the hydrological stations and sections used in this study. Yellow triangles indicate hydrological stations with mapped hydrographs.

2.2. Data Processing

Ensuring spatiotemporal consistency is a prerequisite for integrating SWOT observations with ground-based hydrological data [4,36]. To preserve data reliability, only SWOT records with high-quality flags (reach_q and node_q equal to 0 or 1) were retained. Spatial correspondence between hydrological stations and SWOT observations was then established using the "generate neighbor table" algorithm in ArcMap, which identifies the closest SWOT reach or node to each station. This spatial matching step is essential because the SWOT-provided node locations represent the river centerline and may not coincide spatially with gauging stations [8].

Temporal matching required additional consideration due to the 21-day revisit cycle of SWOT and the asynchronous nature of *in situ* measurements. After unifying the time zones of the two datasets, a 24-h tolerance was adopted. The choice of this one-day matching window was supported by two factors. First, large and medium-sized rivers typically experience modest sub-daily fluctuations in river flow and water surface elevation, with diurnal variability much smaller than seasonal or synoptic changes. Matching within ±1 day therefore introduces minimal hydrological inconsistency while substantially increasing the number of valid pairings. Second, previous studies have used similar windows when reconciling satellite altimetry with *in situ* observations. Satellite altimetry

has emerged as an effective tool for estimating river discharge, particularly in regions with limited *in situ* observations (e.g., Tarpanelli et al. [37]). Discharge can be derived from altimetry-based water levels using stage–discharge relationships (rating curves). In this study, satellite-derived water levels were matched with *in situ* observations within a temporal tolerance of ±1 day to ensure near-synchronous conditions. These precedents confirm that a one-day window is an established and effective approach for mitigating timing mismatches between satellite overpasses and ground measurements. Based on both hydrological reasoning and methodological practice, this study therefore adopted the 24-h matching criteria for SWOT–station pairing.

To further improve spatial robustness and reduce potential mismatches between satellite footprints and station sensor locations, the dataset for each hydrological station was expanded to include its two nearest SWOT nodes. This expansion can yield multiple candidate SWOT records within the ±24 h matching window for a single gauge observation. In such cases, node values were not aggregated but only one record was retained for final match by selecting the SWOT observation with the smallest absolute time difference to the gauge timestamp. If multiple candidates had the same timestamp (from different nearby nodes), duplicate records were removed.

Previous studies suggested that a two-stage modeling framework—first predicting flow occurrence using a classification model and subsequently predicting flow magnitude using a regression model—coukd improve predictive accuracy for intermittent hydrological processes [38]. Such approaches separate occurrence and magnitude components of streamflow generation and are conceptually related to hurdle or zero-inflated models. However, errors introduced in the classification stage may propagate to the regression stage, particularly when the model is applied outside the training distribution or under extrapolation conditions. Misclassification of flow occurrence (e.g., predicting zero flow during a wet period) can therefore produce substantial prediction bias that outweighs the potential benefits of separate modeling stages. For this reason, zero-flow observations were retained and modeled directly within the regression framework in this study, allowing the machine-learning models to learn the full range of hydrological states and enabling a comprehensive evaluation of model robustness across diverse flow conditions. Intermittent or zero-flow conditions are not uncommon in hydrological time-series, resulting in zero-inflated or semi-continuous distributions that can challenge conventional regression algorithms [39]. Modern ML models are capable of implicitly learning the transition boundary between zero-flow and non-zero-flow states when trained on sufficiently representative data [22], as deemed in this study.

Following this preprocessing workflow, 950 stations retained complete and usable records, yielding a total of 50,740 matched data points.

River discharge typically exhibits a highly right-skewed distribution, where flood peaks may exceed baseflow by several orders of magnitude. Such skewed hydrological variables are commonly transformed using logarithmic functions to reduce skewness and approximate a Gaussian-like distribution [40]. Log transformation compresses extreme values and improves the statistical properties of the target variable for modeling. Accordingly, discharge Q was transformed using

$$\log(1+Q) \tag{1}$$

This type of transformation is widely adopted in streamflow modeling studies to mitigate the influence of extreme flood values and improve model stability.

For neural network–based models, the log-transformed discharge was further standardized using Z-score normalization to improve numerical stability during gradient-based optimization. The standardized target variable was defined as

$$z = \frac{\log(1+Q) - \mu}{\sigma}, \tag{2}$$

where μ and σ are the mean and standard deviation of $\log(1+Q)$ computed from the training dataset. Standardization is commonly applied in deep learning–based hydrological models to stabilize training and facilitate efficient gradient optimization [30,41].

Tree-based models were trained directly on $\log(1+Q)$, as these algorithms are generally insensitive to variable scaling. During inference, model outputs were converted back to discharge units by reversing the Z-score normalization and subsequently applying the inverse logarithmic transformation.

2.3. Machine Learning Approaches

2.3.1. Random Forest

Random Forest (RF) is an ensemble learning method based on bagging and decision-tree regression [42]. It constructs a large number of decision trees using bootstrap resampling of the training data and random subsets of

predictor variables at each split, and produces final predictions by averaging across all trees. Owing to its non-parametric nature and inherent resistance to overfitting, RF is well suited for hydrological applications characterized by nonlinear relationships, heterogeneous river conditions, and noisy satellite-derived inputs. In this study, RF models were trained using normalized satellite and auxiliary predictors to directly estimate river flow, without requiring explicit assumptions regarding the functional form of the input–output relationship. It was implemented using an ensemble of 300 decision trees (n_estimators = 300), which was found sufficient to stabilize prediction variance while maintaining computational efficiency. Tree depth was not explicitly constrained (max_depth = None), allowing the model to fully capture nonlinear relationships inherent in river flow processes. Default node-splitting parameters were retained (min_samples_split = 2, min_samples_leaf = 1) to avoid excessive smoothing of extreme flow conditions. A fixed random seed was applied to ensure reproducibility across repeated experiments.

2.3.2. XGBoost

Extreme Gradient Boosting (XGB) is a boosting-based ensemble algorithm that builds decision trees sequentially, where each new tree is trained to correct the residuals of the previous ensemble [27]. By optimizing a regularized objective function using gradient descent, XGB achieves strong predictive performance while effectively controlling model complexity. Compared with bagging-based methods, XGB is particularly effective at capturing complex nonlinear interactions and subtle feature contributions, making it suitable for modeling river flow variability driven by multiple correlated satellite observations. In this study, XGB was implemented as a direct regression model for river flow estimation using the same predictor set as RF, enabling a consistent comparison between ensemble strategies. The XGB model was configured with 500 boosting iterations (n_estimators = 500) and a moderate tree depth (max_depth = 6) to balance model expressiveness and generalization. A relatively small learning rate (learning_rate = 0.05) was adopted to ensure gradual optimization and reduce the risk of overfitting. To enhance robustness against noisy satellite-derived predictors, stochastic subsampling was applied at both the sample and feature (colsample_bytree = 0.8) levels. L2 regularization (reg_lambda = 1.0) was retained to further constrain model complexity.

2.3.3. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) networks are feedforward artificial neural networks composed of multiple fully connected layers and nonlinear activation functions, capable of approximating highly complex continuous mappings. In contrast to tree-based models, MLPs explicitly learn distributed feature representations and are therefore more sensitive to input feature design. In this study, two MLP configurations were evaluated:

(1) a baseline model using only physical and spatial predictors, and
(2) an augmented model incorporating a station encoding variable to represent site-specific characteristics.

This design allows assessment of whether introducing implicit station identity information improves river flow prediction, thereby providing insight into the trade-off between model generalizability and site-specific learning. The MLP consisted of two hidden layers with 128 and 64 neurons, respectively, using ReLU activation functions to model nonlinear river flow responses. The Adam optimizer was employed for efficient gradient-based optimization, with L2 regularization (alpha = 1×10^{-4}) applied to mitigate overfitting.

2.3.4. Transformer

Transformer (TF), originally developed for sequence modeling, relies on self-attention mechanisms to capture long-range dependencies across input features. Unlike recurrent architectures, the TF processes all inputs in parallel and dynamically weight feature interactions through attention scores.

In this study, the TF architecture was adapted for river flow prediction by jointly modeling hydrological variables, spatial coordinates, and temporal encodings. Similar to the MLP framework, two TF variants were implemented—with and without station encoding—to systematically evaluate the contribution of site-specific information. Both employed an embedding dimension of 64 (d_model = 64) and four attention heads (nhead = 4) to capture interactions among hydrological, spatial, and temporal features. Two stacked self-attention layers were used to balance representational capacity and computational cost. A feedforward dimension of 128 and a dropout rate of 0.1 were applied to improve generalization. Similarly to the MLP experiments, station encoding was optionally included as an additional input feature to examine its impact on model performance. By comparing TF performance against both tree-based models and MLPs, this study assesses the relative advantages of attention-based learning for satellite-driven hydrological inference.

2.4. Interpolation and Extrapolation Evaluation Design

To explicitly distinguish between interpolation and extrapolation performance, two complementary validation schemes were implemented in this study. The dataset consists of satellite-based discharge observations from a large number of hydrological stations distributed across China, spanning diverse climatic zones, river morphologies, and hydrological regimes. As a result, model performance reflects not only temporal generalization but also robustness to pronounced spatial heterogeneity.

Interpolation assessment was conducted using five-fold random cross-validation. In this scheme, all available samples from all stations were randomly partitioned into five subsets of approximately equal size. In each fold, four subsets were used for model calibration, while the remaining subset was used for validation. Because both the calibration and validation samples were drawn from the same overall temporal and spatial distribution, the validation data were statistically similar to the training data. Under this setting, the models were evaluated on their ability to interpolate discharge within the range of hydrological states represented in the training set, including comparable flow magnitudes, seasonal conditions, and station-specific characteristics.

Extrapolation assessment was implemented through a time-based split. All observations from all stations were first sorted chronologically, and the earliest two-thirds of the time series were used for model calibration, while the most recent one-third was reserved for validation. This setup emulates a realistic forecasting scenario in which models are trained on historical satellite observations and applied to predict future river flows at the same set of stations. In contrast to random cross-validation, the validation samples in this scheme may contain hydrological regimes, flow extremes, or seasonal patterns that were not fully represented in the training data, thereby requiring the models to extrapolate beyond their calibration domain.

By jointly applying these two validation strategies, the present framework separates a model's ability to reconstruct discharge under statistically and spatially familiar conditions (interpolation) from its ability to generalize to future hydrological states under strong spatial heterogeneity (extrapolation). This distinction is particularly critical for large-scale SWOT-based river monitoring across China, where both temporal non-stationarity and cross-basin variability pose substantial challenges for machine-learning models.

2.5. Model Evaluations and Statistical Metrics

For all models, their performance was evaluated using four common statistical metrics: coefficient of determination (R^2), Nash–Sutcliffe efficiency (NSE), mean absolute error (MAE), and root mean squared error (RMSE) [43,44]. The metrics are defined as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(Q_{obs,i} - \overline{Q_{obs}})(Q_{sim,i} - \overline{Q_{sim}})}{\sqrt{\sum_{i=1}^{n}(Q_{obs,i} - \overline{Q_{obs}})^2}\sqrt{\sum_{i=1}^{n}(Q_{sim,i} - Q_{sim})^2}}\right)^2 \tag{3}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i} - Q_{\mathrm{sim},i}\right)^2}{\sum_{i=1}^{n}\left(Q_{\mathrm{obs},i} - Q_{\mathrm{obs}}\right)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, Q_{obs,i} - Q_{sim,i}| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\, Q_{obs,i} - Q_{sim,i})^2} \tag{6}$$

where $Q_{obs,i}$ and $Q_{sim,i}$ denote the observed and simulated river flow at time i, respectively, $\overline{Q_{obs}}$ and $\overline{Q_{sim}}$ represent the mean values of observed and simulated river flow, and n is the number of samples.

Each metric reflects different aspects of model performance: R^2 measures the strength of the linear relationship between observations and simulations, while NSE evaluates the predictive skill relative to the mean of observations; RMSE emphasizes large errors due to its quadratic formulation, while MAE provides a more robust measure of average prediction error that is less sensitive to extreme values [45]. However, these metrics also have limitations. R^2 primarily reflects correlation rather than prediction accuracy and may remain high even when systematic bias exists. NSE is sensitive to extreme values and may produce overly pessimistic scores for highly variable hydrological series. RMSE is strongly influenced by large deviations, while MAE treats all errors

equally and may underestimate the influence of extreme events. Therefore, using multiple metrics together provides a more comprehensive evaluation and also understanding of the model performance [24,46].

2.6. Sensitivity

Sensitivity analysis for the ML models was conducted using a validation-set perturbation inference approach to quantify how observational uncertainties propagate into discharge predictions. For each trained model, baseline discharge predictions Q were first obtained for the validation dataset. Subsequently, individual input variables (WSE, width, and slope) were perturbed independently while keeping the remaining inputs unchanged, and model inference was repeated to produce new predictions Q'. Perturbations were applied only to samples classified as non-zero flow conditions.

Model sensitivity was quantified using both the absolute prediction change $\Delta Q = Q' - Q$ and the relative change

$$\frac{\Delta Q}{Q} = \frac{Q' - Q}{\max(Q, \varepsilon)} \tag{7}$$

where ε is a small constant preventing numerical instability near zero flow. Perturbations were implemented as truncated Gaussian noise ($\epsilon \sim \mathrm{N}(0, \sigma)$, truncated to $\pm 3\sigma$).

The perturbation magnitude was determined based on the design precision of the SWOT mission and assessments of observation error levels in related studies. Specifically, the standard deviation of the perturbation for water surface elevation (WSE) was set at $\sigma = 0.10\,\mathrm{m}$, which is comparable to the precision level of approximately 10 cm (1σ) achieved by the SWOT mission in inland water surface elevation inversion [4,47]. The perturbation for river width was set as $\sigma = 0.15 \times \mathrm{width}$, and this relative error range referenced estimates of uncertainty in river width inversion from several SWOT river observation studies, generally considered to be within the range of 10–20% [4,48]. The standard deviation of the perturbation for the slope was set to $\sigma = 1.7 \times 10^{-5}\,\mathrm{m/m}$, which corresponds to the design requirement of the SWOT mission for the accuracy of river slope observations, namely $1.7\,\mathrm{cm/km}$ [49]. This setting allows the sensitivity analysis to assess the impact of uncertainties in different variables on the flow estimation results under conditions close to the actual satellite observation error level. The sensitivity results were aggregated at station level using median sensitivity, with number of valid samples recorded for robustness.

2.7. Feature Importance

To interpret the contribution of individual predictors to discharge estimation, feature importance was quantified using SHAP (SHapley Additive exPlanations) [50], a model-agnostic interpretation framework derived from cooperative game theory that attributes the prediction of a model to individual input features based on Shapley values. In this framework, each feature is assigned a contribution representing its marginal effect on the predicted output relative to a reference baseline. The SHAP value for a feature reflects the average contribution of that feature across all possible subsets of predictors, thereby providing a theoretically consistent and locally accurate explanation of model predictions.

In this study, SHAP values were computed using the KernelExplainer, a model-agnostic approximation that estimates Shapley values through weighted linear regression in the neighborhood of each prediction [51]. The explainer treats the trained model as a black-box function and repeatedly evaluates predictions under different feature subsets to approximate the marginal contribution of each variable. A subset of the training data was used as the background dataset to define the reference distribution of inputs, and SHAP values were calculated for samples in both the calibration and validation datasets.

To obtain global feature importance, the mean absolute SHAP value for each predictor was computed across all evaluated samples:

$$\mathrm{Importance}_j = \frac{1}{N}\sum_{i=1}^{N} |\,\phi_{ij}| \tag{8}$$

where ϕ_{ij} denotes the SHAP value of feature j for sample i, and N is the number of evaluated samples. This metric quantifies the overall magnitude of each feature's contribution to model predictions, allowing comparison of the relative importance of hydrological and geometric variables across different model architectures.

3. Results

3.1. Overall Performance

Figures 3 and 4 summarize model performance during calibration and validation for interpolation and extrapolation, respectively, using station-level distributions of R^2, NSE, RMSE, and MAE. Performance metrics are computed independently for each station, and the median value across stations is used to represent the central tendency, while the interquartile range characterizes inter-station variability. Clear performance contrasts were evident among different model classes, i.e., interpolation performing better compared to the extrapolation, when the validation was often with poor performance. Overall under temporal interpolation testing, XGB and RF consistently showed high calibration performance (high $R^2$and NSE with low RMSE and MAE), whereas XGB and RF showed considerable variation and similar performance for both calibration and validation stages, which is overall not desirable. The situation was more complex for the temporal extrapolation, where all models captured poorly the temporal dynamics (low R^2; Figure 4a,b) and absolute accuracy was also not as good as for the interpolation model, although much better for XGB and RF compared to XGB and RF (Figure 4c,d).

A closer inspection of the station-level distributions revealed several systematic patterns that helped explain the differences among model classes. Under the interpolation setting (Figure 3), tree-based models (XGB and RF) not only achieve higher median skill but also show relatively compact interquartile ranges across stations, indicating more stable predictive behavior. For example, the median R^2 values for XGB and RF during validation remain close to 0.8–0.9, while RMSE and MAE remain comparatively low. In contrast, neural network models (MLP and Transformer) exhibit substantially larger inter-station variability, as reflected by wider interquartile ranges and long whiskers. This suggests that these models are more sensitive to differences in station characteristics and data density. One possible explanation is that tree ensembles partition the feature space through hierarchical decision rules, effectively capturing spatial heterogeneity among river basins. By contrast, neural networks attempt to learn a single global nonlinear mapping across all stations, which can be difficult when hydrological regimes vary strongly across the study domain.

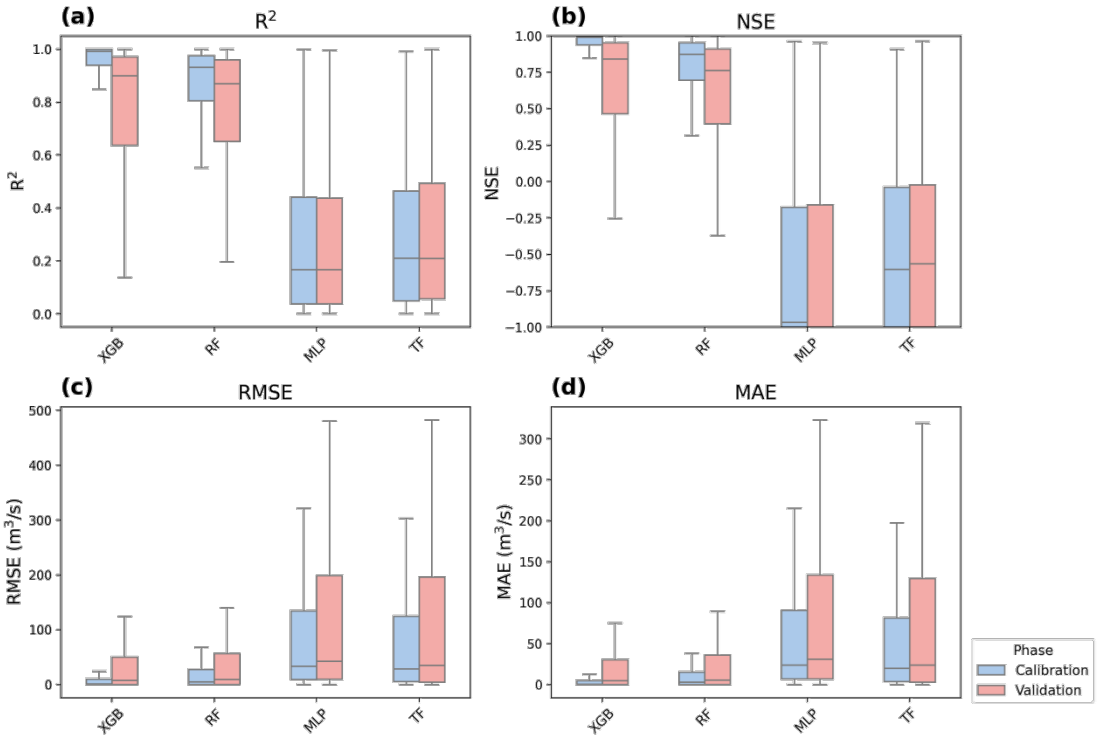

Figure 3. Machine learning model performance for river flow prediction under the interpolation mode. Panels (**a**–**d**) show station-level distributions of R^2, NSE, RMSE, and MAE, respectively, for the calibration and validation phases. Here, R^2 denotes the coefficient of determination, NSE the Nash–Sutcliffe efficiency, RMSE the root mean squared error, and MAE the mean absolute error. For each station, performance metrics are computed independently. The horizontal line within each box indicates the median value across stations, the box spans the interquartile range (25th–75th percentiles), and the whiskers represent the full range, characterizing inter-station variability.

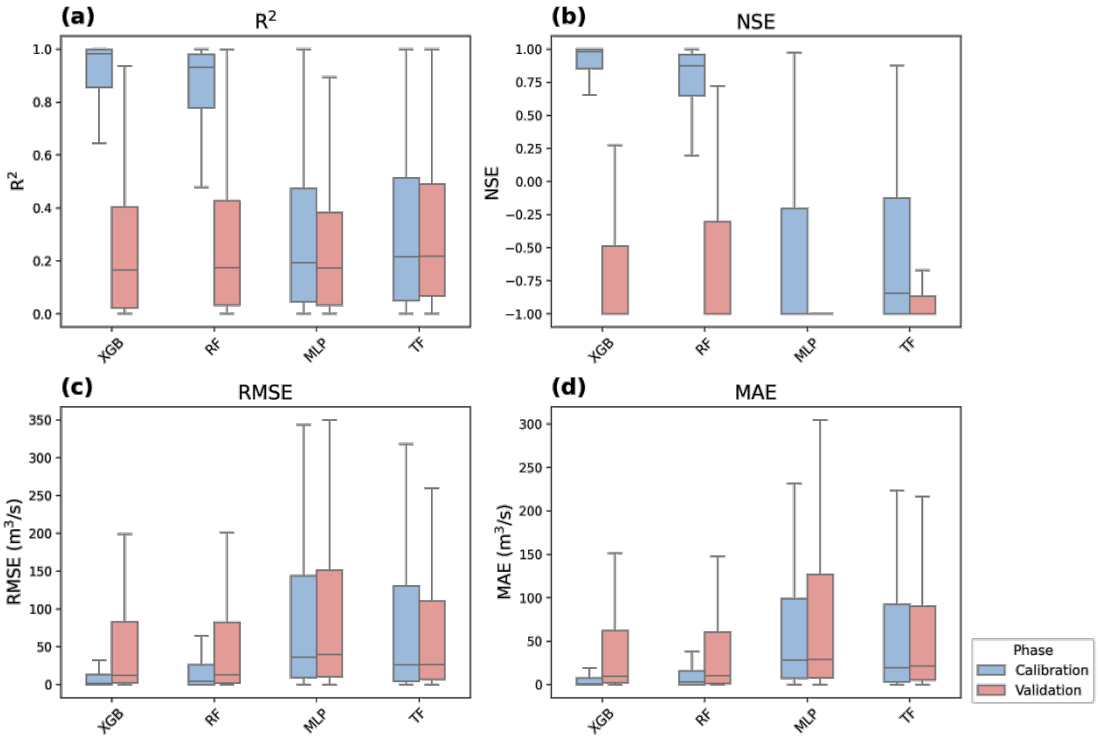

Figure 4. Machine learning model performance for river flow prediction under the extrapolation mode. Panels (**a**–**d**) present station-level distributions of R^2, NSE, RMSE, and MAE, respectively, for the calibration and validation phases, where validation samples fall outside the range of the training data. Here, R^2 denotes the coefficient of determination, NSE the Nash–Sutcliffe efficiency, RMSE the root mean squared error, and MAE the mean absolute error. The median value across stations is shown by the horizontal line within each box, the interquartile range by the box length, and the full range of inter-station variability by the whiskers.

The contrast becomes more pronounced under the extrapolation scenario (Figure 4), where validation samples fall outside the temporal range of the training data. Here, the median validation R^2 for all models decreases markedly, and NSE values frequently become negative, indicating that the predictions are often less informative than using the observed mean discharge as a benchmark. The deterioration is particularly evident for the tree-based models, whose calibration performance remains high but whose validation skill drops sharply. This behavior suggests that these models rely heavily on patterns present in the training period and struggle to reproduce unseen hydrological states during extrapolation. Neural networks also show poor extrapolation skill, but their performance degradation is somewhat more gradual, likely because their distributed representations allow limited generalization beyond the training range. Overall, these results indicate that while machine learning models can effectively reconstruct discharge dynamics when observations overlap in time, their ability to extrapolate to unseen hydrological conditions remains limited.

3.2. Probability Distributions for SWOT-Predicting Station Gauge Data

Figures 5 and 6 present the cumulative distribution functions (CDFs) of the station-level R^2 values for the calibration and the validation phases, offering a complementary distributional perspective that highlights the proportion of stations achieving a given level of predictive skill and facilitates direct comparison of interpolation and extrapolation behavior across models. In the interpolation setting, all models demonstrate high predictive accuracy on validation data, with performance metrics nearly as good as during calibration. XGB and RF achieve the highest median validation R^2 (approximately 0.8–0.85, compared to ~0.9 in calibration) and NSE (median ~0.6–0.7), indicating that these models can capture a large portion of discharge variance in random splits. In contrast, neural-network-based models show lower overall accuracy and larger dispersion. The MLP model demonstrates limited explanatory power, characterized by low median R^2 (generally below 0.4), near-zero or weakly positive NSE values, and wide interquartile ranges. The Transformer model yields modest improvements over MLP, reflected by slightly higher central tendencies and reduced dispersion, but remains inferior to tree-based models under interpolation. The relatively small gap between calibration and validation metrics in this scenario suggests minimal overfitting: model predictions remain robust when tested on unseen data drawn from the same distribution as the training set.

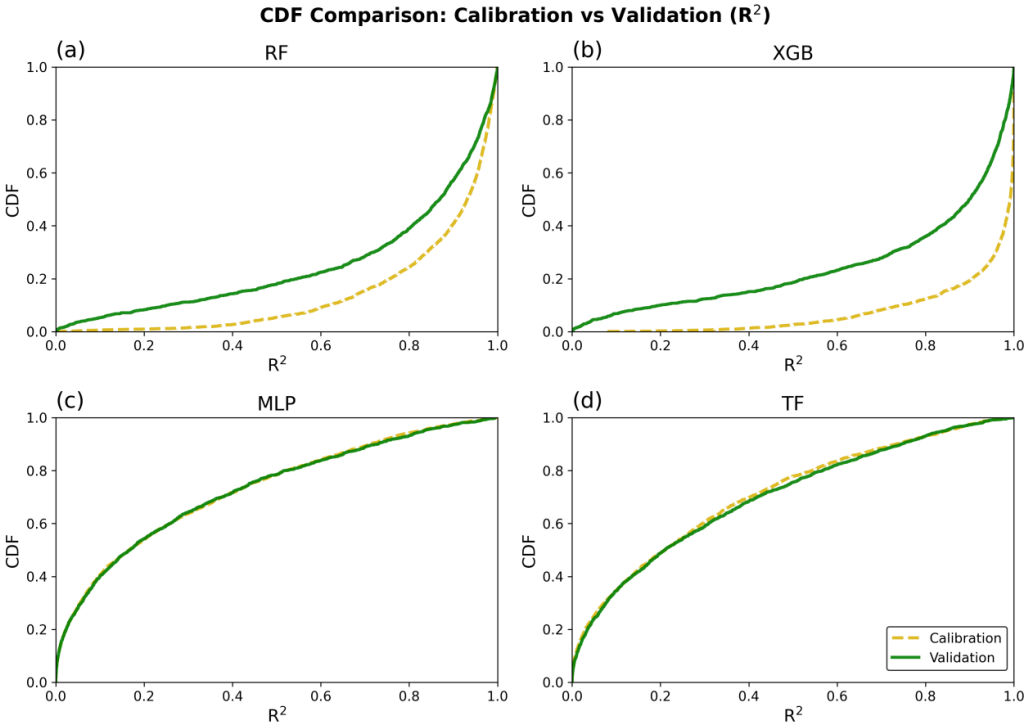

Figure 5. Cumulative distribution functions (CDFs) of performance metrics for four machine learning models under the interpolating mode. Solid green lines denote validation performance within the calibration period, while dashed yellow lines represent calibration results. Panels (**a**–**d**) correspond to: (**a**) Random Forest (RF), (**b**) XGBoost (XGB), (**c**) Multi-Layer Perceptron (MLP), and (**d**) TensorFlow (TF). Note that RF and XGB show larger discrepancies between calibration and validation CDFs, while MLP and TF exhibit closer agreement between the two distributions.

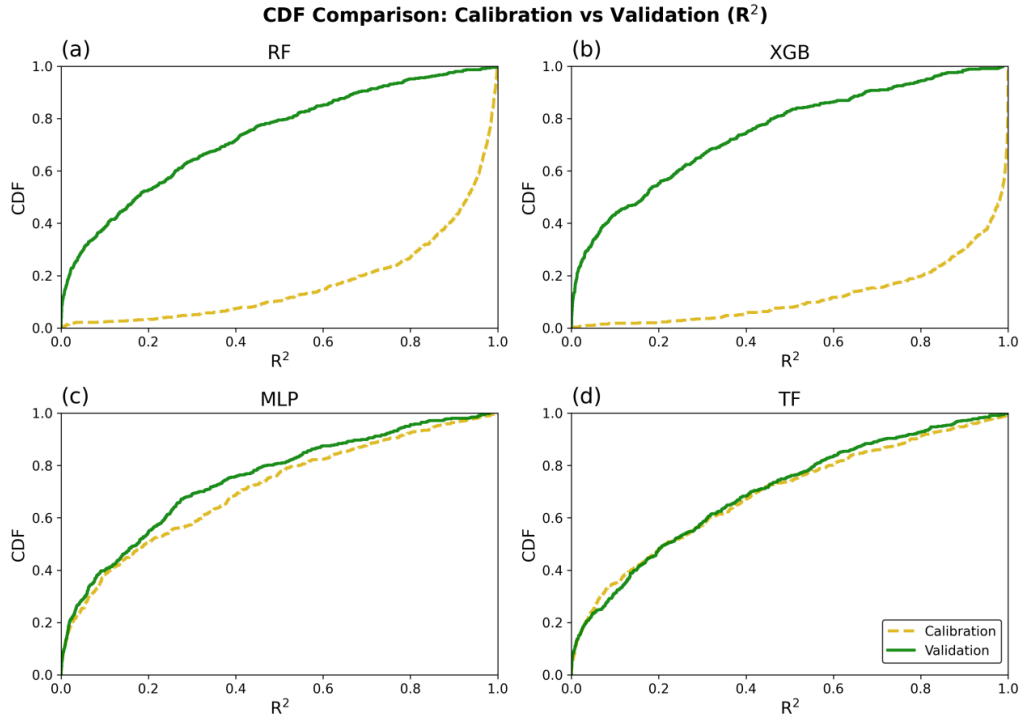

Figure 6. Cumulative distribution functions (CDFs) of performance metrics for four machine learning models under the extrapolation mode. Solid green lines denote validation performance beyond the calibration period, while dashed yellow lines represent calibration results. Panels (**a**–**d**) correspond to: (**a**) Random Forest (RF), (**b**) XGBoost (XGB), (**c**) Multi-Layer Perceptron (MLP), and (**d**) TensorFlow (TF). Notably, RF and XGB exhibit substantial divergence between calibration and validation CDFs, whereas MLP and TF show relatively closer agreement between the two distributions.

Under temporal extrapolation, model behavior changes markedly for all algorithms. Validation performance deteriorates substantially relative to calibration, reflecting the strong non-stationarity of river flow processes and the challenge of predicting outside the temporal support of the training data. This degradation is most pronounced for tree-based models. Despite near-optimal calibration performance, RF and XGB experience sharp declines in validation skill, with median R^2 values dropping to approximately 0.2–0.4 and NSE frequently approaching or falling below zero. The large divergence between calibration and validation distributions indicates that tree-based models rely heavily on temporally local patterns and exhibit limited robustness under extrapolative conditions.

The neural-network-based models displayed a different, though not necessarily superior, extrapolation behavior. For both MLP and Transformer, validation R^2 distributions are broadly comparable to their calibration counterparts, with similar medians and interquartile ranges. However, this apparent stability primarily reflects their already limited calibration skill rather than strong extrapolative capability. While the absence of a sharp performance collapse suggests reduced sensitivity to temporal shifts, absolute predictive accuracy remains modest, with NSE often remaining negative. Among the evaluated neural models, the Transformer shows slightly improved consistency relative to the MLP, indicating a marginally stronger ability to learn temporally transferable representations, though its overall extrapolation skill remains constrained.

3.3. Feature Analysis of Different ML Methods

As shown in Figure 7, the number of effective samples across all stations exhibits a pronounced right-skewed (positively skewed) distribution, with the mode concentrated in the 20–60 range. In contrast, the Wulong station, with 428 valid samples, lies in the high-value tail (>400), indicating a substantially higher data density than most other stations.

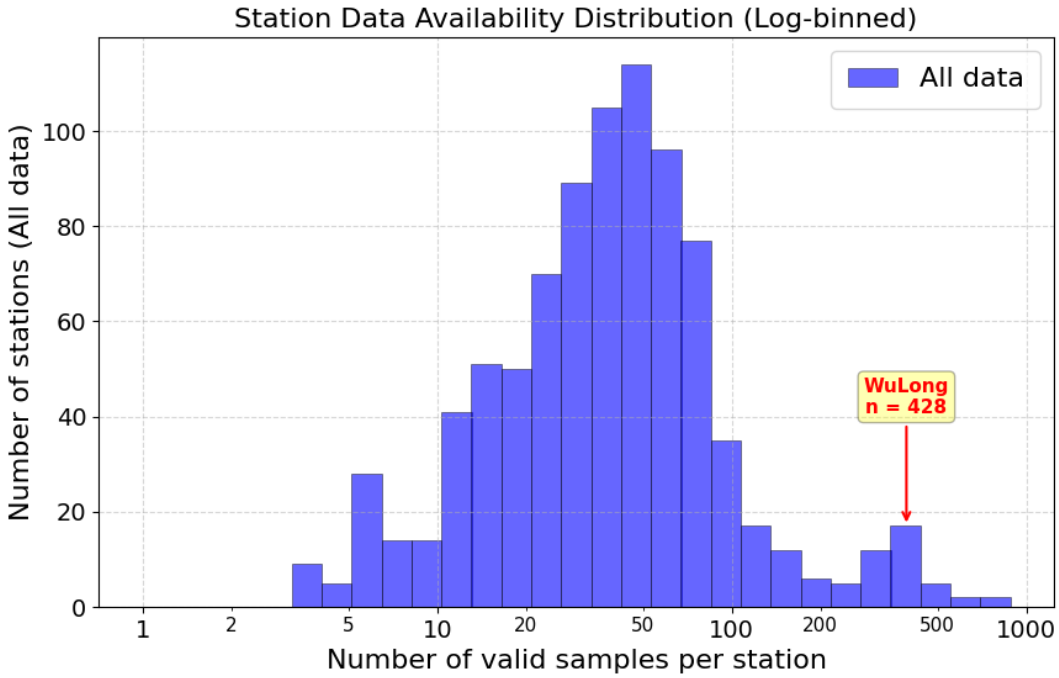

Figure 7. Distribution of data availability across hydrological stations in the study area (log-binned). The x-axis denotes the number of valid samples per station (log-scaled bins), and the y-axis indicates the count of stations. The red annotation marks the Wulong Station, which has 428 valid samples with zero missing values, positioned at the right tail representing high data completeness.

In addition, the discharge dynamic range at this station exceeds 95.48% of the other matched stations, covering a wide spectrum of hydrological conditions from low-flow periods to peak flood events. This combination ensures both dense observational coverage and substantial hydrological variability, enabling robust and visually interpretable evaluation of machine learning models across diverse regimes, including rising limbs, recession periods, and quasi-steady flow conditions.

Overall, this combination of high data completeness and broad hydrological variability makes the Wulong station an ideal benchmark for model evaluation. The large sample size (approximately ten times the regional average) provides strong support for model training and reliable generalization assessment, while the wide

discharge range introduces diverse flow conditions that effectively test model performance across different hydrological regimes, particularly under extreme events.

As a key control station in the upper Yangtze River, Wulong integrates the characteristics of mountainous river systems with anthropogenic influences. Its data therefore represent a typical hydrological response under complex real-world conditions. Consequently, this station enables an objective assessment of model accuracy and generalization in capturing hydrological dynamics, while also providing a reliable basis for evaluating model stability across varying flow conditions.

The predictive performance of all ML models involved in the study is illustrated in Figure 8 for Wulong hydrological station. Under interpolating conditions, tree-based ensemble models, specifically XGB and RF exhibited superior performance. Both models effectively captured peak and low-flow conditions within complex streamflow dynamics, with predicted trajectories closely following observed river flow variations. Among them, XGB showed a slightly enhanced ability to reconstruct hydrological events, yielding more accurate representations of both the timing and magnitude of flow extremes.

In contrast, neural-network-based models (MLP and Transformer, TF) displayed weaker responsiveness to event-scale variability under interpolation. Their predictions tended to be smoother, with attenuated peak responses and limited ability to reproduce sharp peak valley structures, reflecting constrained sensitivity to rapid hydrological changes even when observations are relatively dense.

Figure 9 further illustrates model behavior under temporal extrapolation. When observations become temporally sparse, all evaluated models exhibit pronounced underestimation of river flow, particularly during high-flow events. This systematic bias indicates a reduced capacity to generalize hydrological relationships beyond the temporal domain covered by training data. Among the four models, RF demonstrates comparatively better robustness under data-scarce conditions, followed by the Transformer, whereas XGB and MLP show more severe performance degradation.

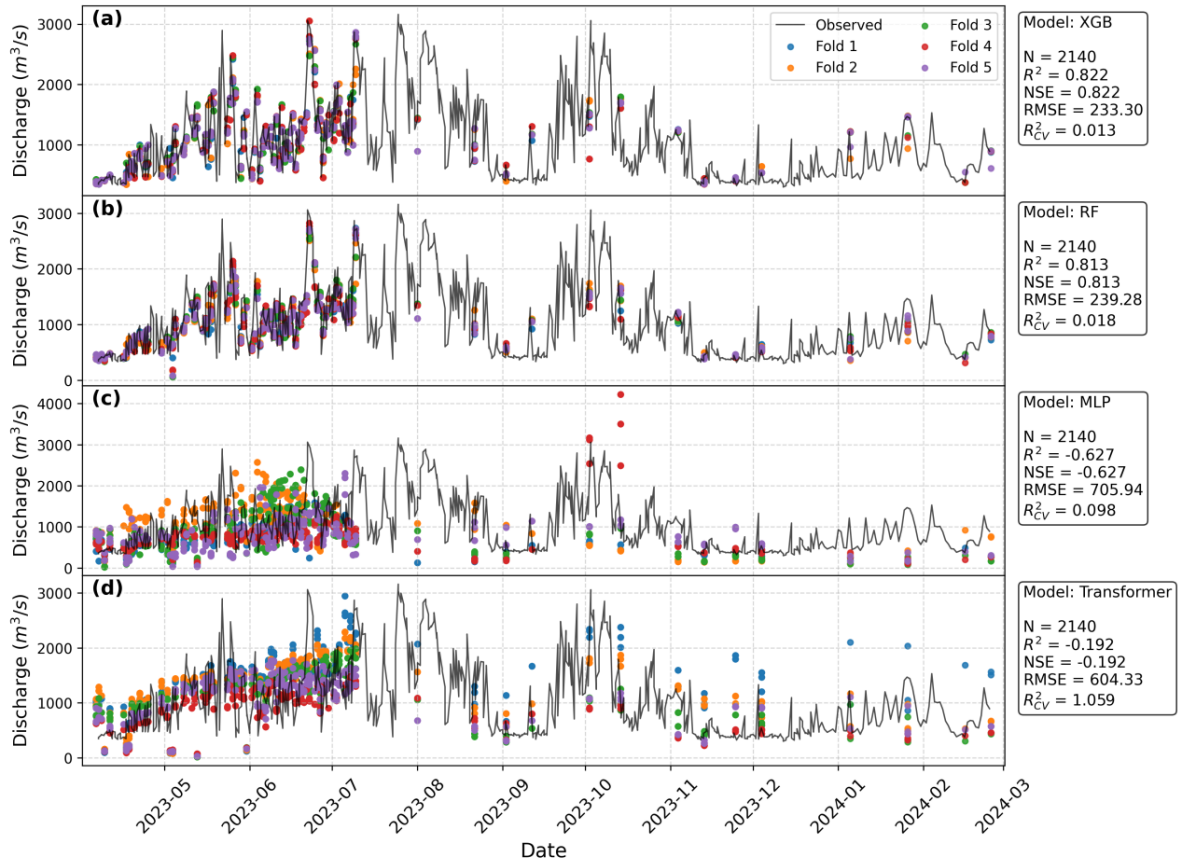

Figure 8. Hydrological process curves derived from machine learning predictions under the interpolating mode at Wulong Hydrological Station with dense observations. The gray–black broken line represents observed river flow, and colored points denote model predictions. Subplots (**a**–**d**) show results obtained using XGB, RF, MLP, and TF as regressors, respectively.

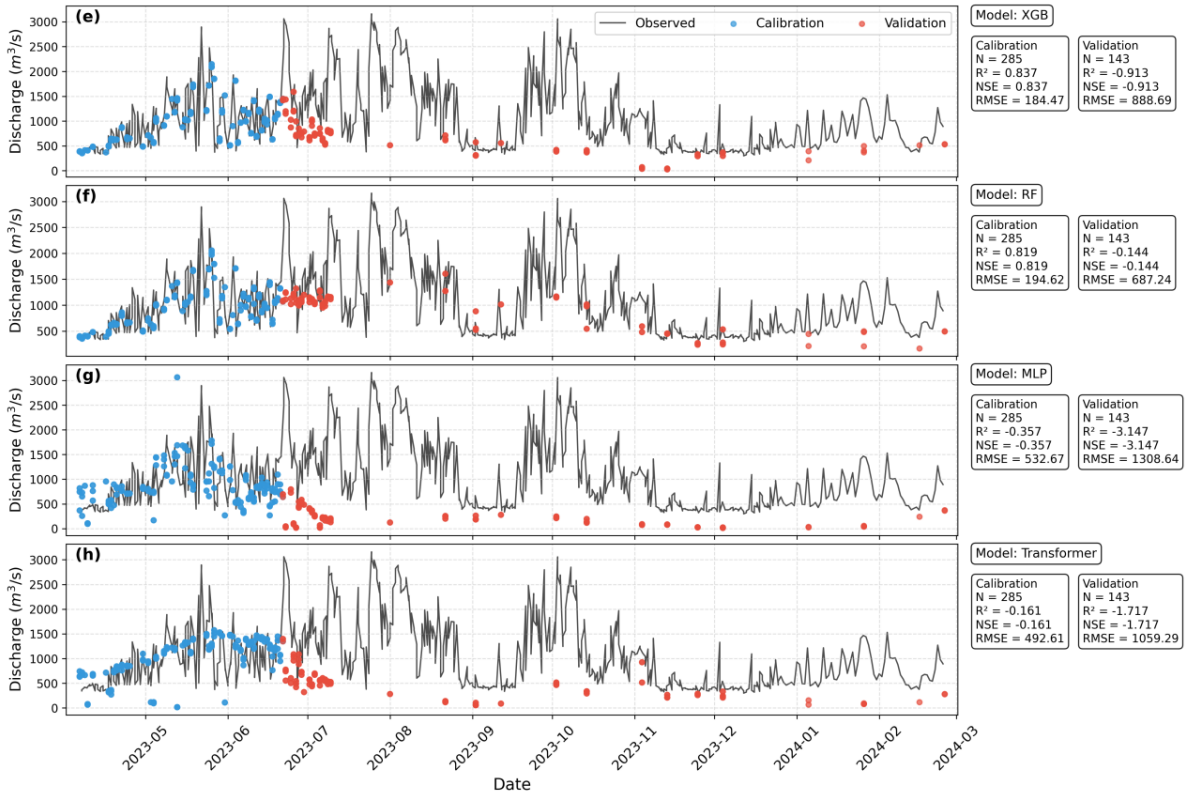

Figure 9. Hydrological process curves derived from machine learning predictions under the extrapolation mode at Wulong Hydrological Station with dense observations. The gray–black broken line represents observed river flow. Blue points represent predictions for the temporally ordered training period, while red points indicate predictions for the extrapolated validation period. Subplots (**e**–**h**) correspond to XGB, RF, MLP, and TF, respectively.

During periods with relatively denser observations, XGB and RF predictions exhibit a clear tendency toward mean convergence, resulting in suppressed variability and diminished representation of river flow extremes. In contrast, MLP and TF are able to partially reflect flow fluctuations and recover some event-scale variability; however, their predictions remain quantitatively inaccurate and fail to consistently match observed magnitudes.

Overall, the temporal extrapolation results reveal inherent limitations of data-driven machine learning models when training samples are limited and temporally incomplete. The observed underestimation during sparse periods and mean-convergent behavior during denser intervals collectively highlight the challenges of extrapolating river flow dynamics beyond the temporal range of available observations using purely data-driven approaches.

3.4. Sensitivity

Sensitivity analysis showed that random perturbations in channel slope produced the largest changes in predicted discharge, whereas water surface elevation (WSE) noise had only a minor influence, whereas channel width noise was in between (Figure 10). Hence, although discharge uncertainty is generally influenced by multiple observational variables, slope uncertainty can become a dominant contributor under certain conditions, particularly in low-gradient rivers or hydraulically mild reaches.

3.5. SHAP Analysis

Figure 11 presents the feature importance derived from SHAP values for the four ML models during both calibration and validation phases. Overall, spatial variables (latitude and longitude) consistently emerge as the most influential predictors across models, followed by hydrological variables such as WSE, width, and slope, while the seasonal indicators (DOY_sin and DOY_cos) generally contribute less to model predictions.

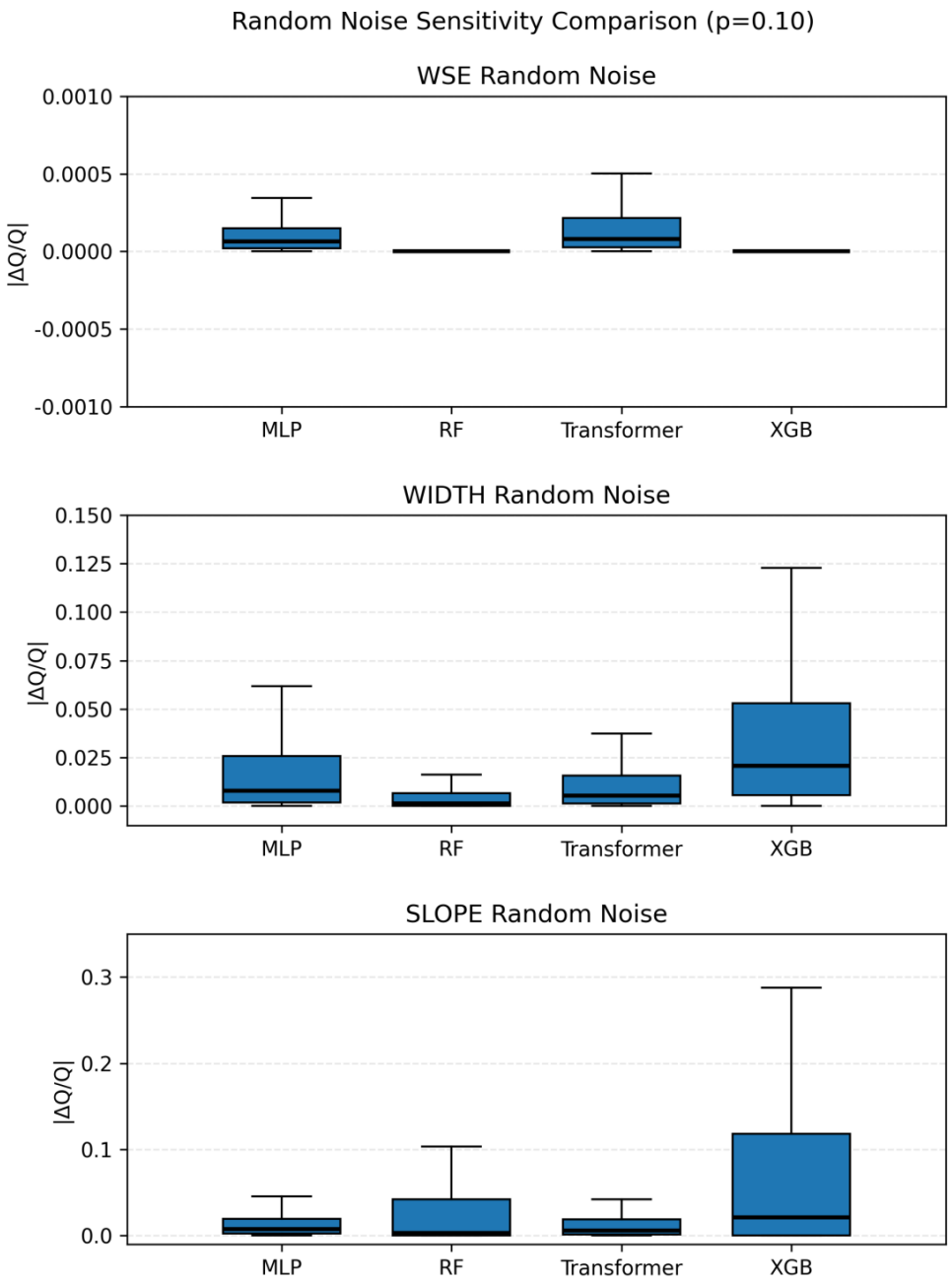

Figure 10. A comparison chart of relative traffic deviations after adding random errors. From top to bottom, the chart shows the relative changes in traffic after adding errors to WSE, width, and slope.

For the tree-based models (XGB and RF), the ranking of feature importance remains largely consistent between calibration and validation. Latitude was the most dominant feature in both models, indicating that spatial gradients across river basins play a critical role in determining discharge variability. Longitude and WSE followed as secondary predictors, while width and slope showed moderate importance. The relatively stable ranking between calibration and validation suggests that tree-based ensemble methods capture robust spatial partitioning structures in the feature space, allowing the models to maintain similar decision boundaries when applied to unseen data. In contrast, the neural network models (MLP and Transformer) exhibit more noticeable changes in feature importance between calibration and validation. During calibration, spatial variables (Lat and Lon) dominate feature contributions, reflecting models' tendency to learn strong spatial correlations within the training data. However, in the validation phase, hydrological variables—particularly width and seasonal signals (DOY_sin)—become more prominent. For example, in MLP validation results, width and DOY_sin emerge as the two most influential predictors, indicating that the neural network adjusts its reliance toward dynamic hydrological variables when applied to unseen observations. This shift suggests neural networks distribute predictive influence across multiple correlated inputs and may adapt their internal representations depending on the data subset. Across all

models, WSE remains a consistently important variable, while width and slope also contribute substantially, reflecting their physical relevance to channel geometry and hydraulic gradient. In comparison, seasonal harmonic terms (DOY_sin and DOY_cos) show lower importance, implying that seasonal variability plays a secondary role relative to spatial heterogeneity and instantaneous hydraulic conditions.

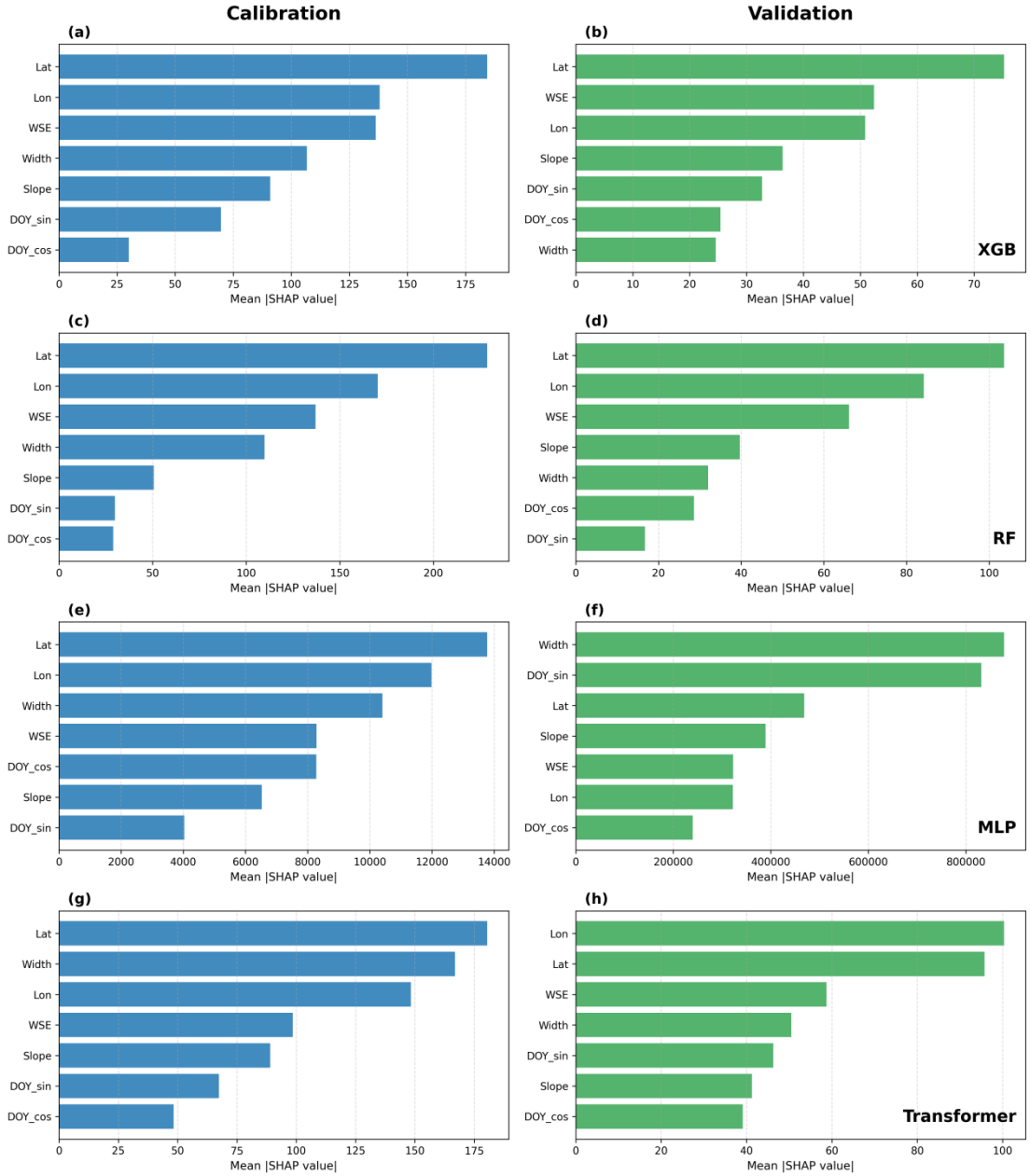

Figure 11. SHAP analysis results of different models on the test (calibration) and validation sets. Subplots are arranged by model and dataset: (**a**) XGB—Training set, (**b**) XGB—Validation set, (**c**) RF—Calibration set, (**d**) RF—Validation set, (**e**) MLP—Calibration set, (**f**) MLP—Validation set, (**g**) Transformer—Calibration set, (**h**) Transformer—Validation set. In each subplot, the horizontal axis represents the absolute value of the average SHAP, and the vertical axis represents the six input features (latitude, longitude, WSE, river width, slope, and DOY sine/cosine). Blue bars correspond to the calibration (test) set results, and green bars correspond to the validation set results.

Taken together, these results indicate that tree-based and neural network models both rely on similar hydrological variables, but the former maintain more stable feature attribution across datasets, whereas the later is flexible and context-dependent. This difference likely reflects the structural distinctions between decision-tree ensembles, which partition the feature space through hierarchical splits, and neural networks, which learn distributed nonlinear representations across multiple inputs.

4. Discussion

4.1. Suitability of SWOT for River Morphology Mapping

The results of this study highlight the distinctive advantages of SWOT observations over previous satellite missions such as Landsat and Sentinel-1 which primarily provide information on river extent or width for river flow reconstruction. Owing to its advanced Ka-band Radar Interferometer (KaRIn), SWOT provides substantially higher spatial resolution and two-dimensional measurements of river width, WSE, and slope than earlier nadir altimeters (e.g., Jason-2/3 and Sentinel-3), which are limited to one-dimensional along-track observations [4,8,48]. Together with its relatively short effective revisit cycle at mid-latitudes, SWOT enables a more detailed characterization of river dynamics that can be directly exploited by data-driven models. This enhanced observational capability helps explain the strong interpolation performance explains the strong interpolating performance observed in Figure 3 and the steep right-skewed CDFs during calibration, particularly for tree-based models.

Compared to the temporal interpolation results, the extrapolation results showed a systematic leftward shift of the CDFs (Figure 6) and reduced validation R^2 and NSE (Figure 4). This likely reflects the combined effects of data sparsity, observation noise, and the statistical structure of the learning problem. Although the extrapolation samples often fall within the nominal range of the observed WSE–width–slope feature space, extreme and transitional hydrological states remain sparsely represented in the training set.

To further evaluate the impact of logarithmic transformation on model performance, we compared the results obtained using the original (non-transformed) discharge data under both interpolation and extrapolation scenarios (Figures 12 and 13).

The results show that, without logarithmic transformation, the interquartile ranges (IQRs) of all four evaluation metrics (R^2, NSE, RMSE, and MAE) during the calibration phase increase substantially across all models. This indicates a significant rise in model instability and dispersion of predictive performance. Although the median or mean values of certain metrics may appear slightly improved compared to those obtained using log-transformed data, the overall variability is markedly larger, suggesting reduced robustness and generalization capability.

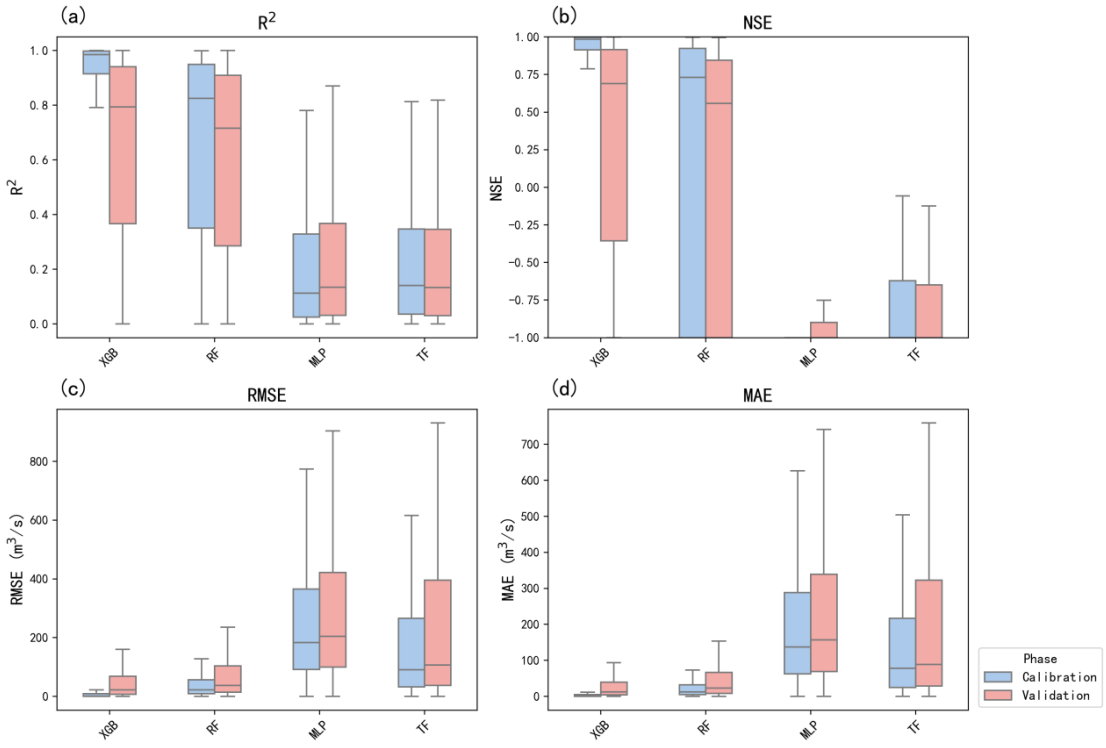

Figure 12. Machine learning model performance for river flow prediction under the interpolation mode using non-log-transformed discharge data. Panels (**a**–**d**) show station-level distributions of R^2, NSE, RMSE, and MAE, respectively, for the calibration and validation phases. Here, R^2 denotes the coefficient of determination, NSE the Nash–Sutcliffe efficiency, RMSE the root mean squared error, and MAE the mean absolute error. For each station, performance metrics are computed independently. The horizontal line within each box indicates the median value across stations, the box spans the interquartile range (25–75th percentiles), and the whiskers represent the full range, characterizing inter-station variability. Compared with results based on log-transformed data, a noticeable expansion of interquartile ranges is observed, indicating increased variability and reduced model stability across stations.

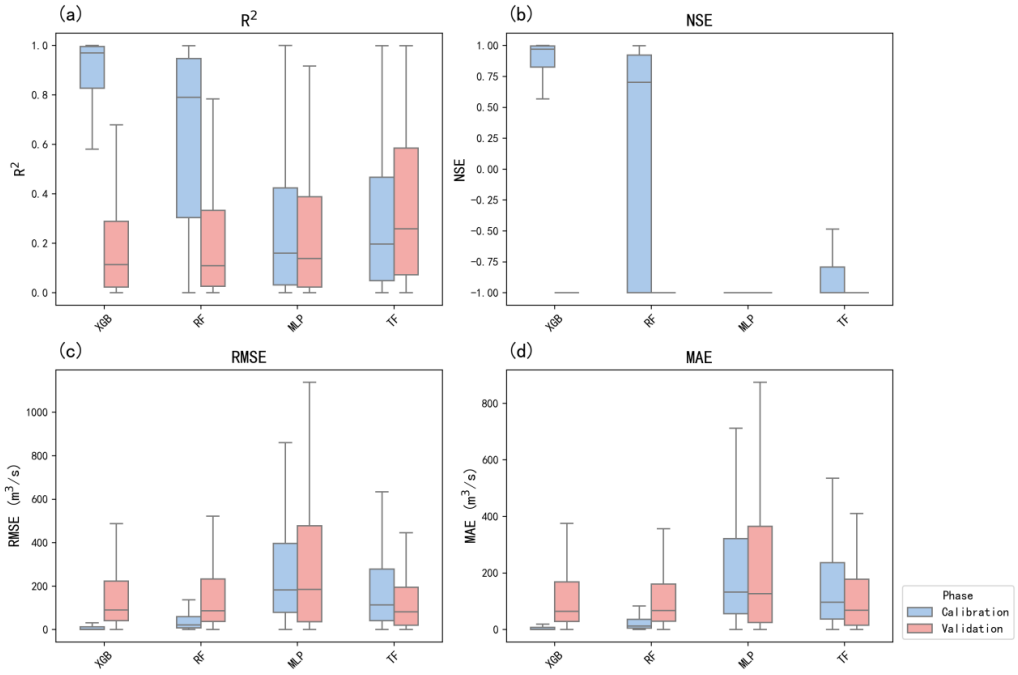

Figure 13. Machine learning model performance for river flow prediction under the extrapolation mode using non-log-transformed discharge data, where validation samples fall outside the range of the training data. Panels (**a**–**d**) present station-level distributions of R^2, NSE, RMSE, and MAE, respectively, for the calibration and validation phases. Here, R^2 denotes the coefficient of determination, NSE the Nash–Sutcliffe efficiency, RMSE the root mean squared error, and MAE the mean absolute error. The median value across stations is shown by the horizontal line within each box, the interquartile range by the box length, and the full range of inter-station variability by the whiskers. Relative to the log-transformed case, the distributions exhibit greater dispersion and instability, particularly during the calibration phase, reflecting the amplified influence of extreme discharge values under extrapolation conditions.

This phenomenon can be attributed to the extremely wide dynamic range of discharge (Q) in the original data. In the absence of transformation, high-flow samples dominate the optimization process under MSE-type loss functions, leading the model to preferentially fit large rivers with high discharge values. Consequently, the model tends to overfit these high-magnitude observations while underrepresenting the more numerous small- and medium-sized rivers, which are characterized by greater heterogeneity in geomorphology, hydrological conditions, and observational uncertainty.

In contrast, logarithmic transformation effectively compresses the scale of high-flow values, reducing their disproportionate influence during training and enabling a more balanced representation of different flow regimes. This not only stabilizes the optimization process but also improves the consistency of model performance across stations, as evidenced by the reduced dispersion in evaluation metrics.

Therefore, although logarithmic transformation may theoretically introduce a tendency to underestimate peak flows, the experimental results demonstrate that, in this study, its benefits in enhancing model stability and mitigating scale imbalance outweigh its potential drawbacks. In particular, the improved robustness across diverse hydrological conditions suggests that logarithmic transformation is a necessary preprocessing step for large-scale, multi-station discharge modeling.

Second, zero-flow conditions play a disproportionately important role in shaping the observed degradation. In these cases, *in situ* river flow is zero, while SWOT may still report nonzero river width or water surface elevation due to residual water or retrieval uncertainties. This introduces highly nonlinear and sometimes non-unique relationships between SWOT observables and river discharge, particularly under intermittent or low-flow conditions [17,48]. Consequently, all models exhibit a systematic low-flow bias during validation (Figures 7 and 8), consistent with the widespread underestimation and the expansion of RMSE and MAE tails observed in the extrapolation statistics.

The importance of slope in the sensitivity of the results (Figure 10) has also been reported previously. Because many discharge estimation approaches are derived from Manning-type hydraulic relationships, the nonlinear dependence between discharge and slope can amplify slope measurement errors and propagate them into substantial discharge uncertainty. In addition, satellite observation geometry can further increase slope-related errors. For example, radar interferometric effects such as layover may degrade the accuracy of retrieved water

surface elevation and slope, thereby introducing additional uncertainty into discharge estimates [36,52]. Nevertheless, slope error is not always the sole dominant source of uncertainty; discharge accuracy is typically influenced by the combined effects of WSE, river width, slope, and hydraulic parameterization [34]. Overall, the sensitivity pattern observed in this study—where slope perturbations produce the largest impact on discharge predictions—is consistent with the general understanding of uncertainty propagation in SWOT-based river discharge estimation.

4.2. Impact of Machine Learning Classifier on the Results

The contrasting behaviors between model classes in terms of their CDF showed different impact on the results. The tree-based models (RF and XGB) achieved near-perfect calibration skill, as indicated by their highly right-skewed calibration CDFs (Figures 5 and 6) for both inter- and extrapolation, but their validation distributions was toward low R^2 values under extrapolation. This behavior suggests that these models strongly exploit station-specific seasonal and regime-dependent patterns that are not transferable beyond the training period. In contrast, the neural-network-based models (MLP and Transformer) displayed much smaller CDF shifts between calibration and validation, indicating improved temporal robustness. Although their absolute accuracy remains moderate, the stability of their CDF shapes implies that these models learn more time-transferable hydrological representations rather than merely memorizing specific seasonal patterns. Expanding the effective training data volume and diversity is of interest in future studies.

While tree-based models appear to showed superior accuracy under interpolation, their temporal extrapolation capability is fundamentally constrained by the limited sampling of extreme and transitional hydrological states (such as rapid flow increases or recession phases between low- and high-flow regimes) and by the tendency of error-minimizing regressors to revert toward conditional means in poorly constrained regions. Deep learning models, particularly Transformers, are better suited to exploit longer temporal context and complex nonlinear dependencies, but require substantially larger dataset, which for large regions like China is difficult to obtain, and also the dataset should ideally be more diverse. Integrating SWOT with additional satellite missions, longer station records, meteorological forcing, and basin-scale physiographic information would not only improve the CDF distributions and reduce extrapolation bias, but also enable models to learn more stable and transferable representations of river dynamics across time.

4.3. Insights from Differences in Feature Importance

The SHAP-based feature-importance reveal clear differences between tree-based models and neural network models in terms of stability and feature utilization. For the tree-based models (RF and XGB), the relative importance of predictors remains largely consistent between the calibration and validation phases. Spatial variables (Lat and Lon) and key hydraulic variables (WSE and width) consistently dominate the prediction process. This stability likely arises from the structural characteristics of decision-tree ensembles, which partition the feature space through hierarchical split rules. By splitting on spatial coordinates, tree models effectively perform an implicit regionalization of hydrological behavior, allowing different relationships between hydraulic variables and discharge to emerge in different geographic regions. Given the strong spatial heterogeneity of river morphology and hydrological regimes across China, this piecewise modeling strategy may partly explain the relatively robust performance of tree-based models.

In contrast, the neural network models (MLP and Transformer) exhibit more noticeable variations in feature importance between calibration and validation. The importance of predictors becomes more distributed, and the dominant variables shift across datasets. This behavior reflects the different learning mechanism of neural networks, which attempt to learn a single global nonlinear mapping across all basins rather than region-specific relationships. When the training data are limited or unevenly distributed, such models may struggle to generalize spatially heterogeneous hydrological processes. These results suggest that improving neural network performance may require expanding the training dataset to better capture the diversity of hydrological conditions. In addition, incorporating additional physically meaningful predictors—such as catchment characteristics, climate variables, or upstream hydrological information—may help neural networks learn more stable and transferable representations of river discharge dynamics.

4.4. Limitations of Time-Series Models under Sparse and Irregular Observations

Time-series models such as ARIMA and recurrent neural networks (e.g., LSTM) are widely used in hydrological prediction because they can capture temporal dependencies in continuous hydrological records [53,54].

However, their effectiveness generally relies on relatively dense and regularly spaced observations that adequately resolve the temporal evolution of hydrological processes.

In SWOT-based river flow reconstruction, these conditions are rarely satisfied. SWOT observations must be matched with *in situ* station data, resulting in datasets with long and irregular temporal intervals between observations. Due to orbital coverage and data filtering during the matching process, the effective sampling interval for a given river reach may range from several days to multiple weeks [4,55]. Such sparse and irregular sampling limits the applicability of conventional time-series models. Methods such as ARIMA assume regularly spaced observations, while LSTM-type models rely on dense sequential information to learn temporal dependencies. When large temporal gaps exist between observations, key hydrological processes—such as flood peaks and rapid rising limbs—may occur between observations and remain unrecorded. As a result, time-series models may struggle to reconstruct hydrological dynamics and correctly identify flow variability. Under these conditions, models that exploit instantaneous hydraulic relationships among variables such as water surface elevation, width, and slope may provide more robust performance.

5. Conclusions

This study investigated the potential of using SWOT satellite observations with different machine learning (ML) models for reconstructing streamflow at gauged locations across China. One of the characteristics of the dataset is temporally sparse gauge observations combined with irregular satellite sampling, providing a suitable testbed for evaluating temporal interpolation and extrapolation capabilities. Using SWOT-derived river width, water surface elevation, and slope as predictors, all ML models were able to accurately recover hydrological variability when predictions were made within the temporal domain covered by observations. Under such interpolating conditions, the tree-based models RF and XGB consistently achieved the highest predictive skill, characterized by high median R^2 and NSE values, low RMSE and MAE, and relatively narrow inter-station variability. Under temporal extrapolation, all models experienced substantial degradation in validation performance relative to calibration. This degradation is most pronounced for tree-based models, which show a sharp decline from near-perfect calibration skill to low median validation $R^2$and frequently negative NSE values. The strong contrast between calibration and validation indicates that these models are highly sensitive to temporal shifts and rely heavily on patterns specific to the training period. Despite this decline, feature-importance analysis suggests that tree-based models maintain relatively stable predictor usage, likely because their hierarchical splitting structure effectively partitions the feature space and captures spatial heterogeneity in hydrological regimes.

The results not only demonstrate the potential of using SWOT satellite observation for streamflow reconstruction, but also highlight the necessity of expanding both the volume and diversity of training data to achieve robust and transferable discharge prediction at large spatial scales. In particular, integrating SWOT with complementary satellite missions, longer *in situ* discharge records, meteorological forcing, and basin-scale physiographic variables may help models better capture the complex controls on river discharge. Such data-enriched frameworks are expected to improve temporal generalization and enable machine learning models—especially neural networks—to learn more stable and physically transferable representations of river dynamics. Such data-enriched frameworks provide a critical pathway toward scalable, reliable, and long-term satellite-based river flow monitoring in data-limited regions.

Author Contributions

S.Z.: methodology, investigation, software, validation, data curation, visualization, writing—original draft; K.M.: investigation, supervision, writing—review and editing; J.T. (Jing Tian): conceptualization, supervision, writing—review and editing, funding acquisition; J.T. (Jingyi Tian): writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Inner Mongolia Autonomous Region Water Conservancy Research Special Project (Grant No. NSK202301) and the National Natural Science Foundation of China (Grant No. 42361144709). The APC was funded by the Y.Q.Z. research group at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

We used the following datasets: podaac/hydrocron: 1.6.4 (https://doi.org/10.5281/zenodo.15831782); PO.DAAC Cookbook v2024.01 (https://doi.org/10.5281/zenodo.10530664); SWOT River Database (SWORD) v17b (https://doi.org/10.5281/zenodo.15299138). The code used in this article is stored in: https://github.com/ZSY-IGSNRR/SWOT-in-China.

Acknowledgments

I would like to thank IGSNRR for providing the research environment and paper repository, and thank the teachers and students in Y.Q.Z.'s research group for their invaluable assistance to my research.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT to investigate the background and current status of the research to assist in code generation. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Nearing, G.; Cohen, D.; Dube, V.; et al. Global prediction of extreme floods in ungauged watersheds. *Nature* **2024**, *627*, 559–563. https://doi.org/10.1038/s41586-024-07145-1.
2. Newman, A.J.; Clark, M.P.; Sampson, K.; et al. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.* **2015**, *19*, 209–223. https://doi.org/10.5194/hess-19-209-2015.
3. Oubanas, H.; Gejadze, I.; Malaterre, P.-O.; et al. Discharge Estimation in Ungauged Basins Through Variational Data Assimilation: The Potential of the SWOT Mission. *Water Resour. Res.* **2018**, *54*, 2405–2423. https://doi.org/10.1002/2017WR021735.
4. Biancamaria, S.; Lettenmaier, D.P.; Pavelsky, T.M. The SWOT Mission and Its Capabilities for Land Hydrology. *Surv. Geophys.* **2016**, *37*, 307–337. https://doi.org/10.1007/s10712-015-9346-y.
5. Bjerklie, D.M.; Moller, D.; Smith, L.C.; et al. Estimating discharge in rivers using remotely sensed hydraulic information. *J. Hydrol.* **2005**, *309*, 191–209. https://doi.org/10.1016/j.jhydrol.2004.11.022.
6. Ke, L.; Xu, J.; Fan, C.; et al. Remote sensing reconstruction of long-term water level and storage variations of a poorly-gauged river in the Tibetan Plateau. *J. Hydrol. Reg. Stud.* **2022**, *40*, 101020. https://doi.org/10.1016/j.ejrh.2022.101020.
7. Scherer, D.; Schwatke, C.; Dettmering, D.; et al. Monitoring river discharge from space: An optimization approach with uncertainty quantification for small ungauged rivers. *Remote Sens. Environ.* **2024**, *315*, 114434. https://doi.org/10.1016/j.rse.2024.114434.
8. Alsdorf, D.E.; Rodríguez, E.; Lettenmaier, D.P. Measuring surface water from space. *Rev. Geophys.* **2007**, *45*, 2006RG000197. https://doi.org/10.1029/2006RG000197.
9. Tang, Q.; Gao, H.; Lu, H.; et al. Remote sensing: Hydrology. *Prog. Phys. Geogr. Earth Environ.* **2009**, *33*, 490–509. https://doi.org/10.1177/0309133309346650.
10. Ahmad, W.; Kim, D. Estimation of flow in various sizes of streams using the Sentinel-1 Synthetic Aperture Radar (SAR) data in Han River Basin, Korea. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *83*, 101930. https://doi.org/10.1016/j.jag.2019.101930.
11. Mengen, D.; Ottinger, M.; Leinenkugel, P.; et al. Modeling River Discharge Using Automated River Width Measurements Derived from Sentinel-1 Time Series. *Remote Sens.* **2020**, *12*, 3236. https://doi.org/10.3390/rs12193236.
12. McCabe, M.F.; Rodell, M.; Alsdorf, D.E.; et al. The future of Earth observation in hydrology. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 3879–3914. https://doi.org/10.5194/hess-21-3879-2017.
13. Bjerklie, D.M.; Lawrence Dingman, S.; Vorosmarty, C.J.; et al. Evaluating the potential for measuring river discharge from space. *J. Hydrol.* **2003**, *278*, 17–38. https://doi.org/10.1016/S0022-1694(03)00129-X.

14. Altenau, E.H.; Pavelsky, T.M.; Durand, M.T.; et al. The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A Global River Network for Satellite Data Products. *Water Resour. Res.* **2021**, *57*, e2021WR030054. https://doi.org/10.1029/2021WR030054.
15. Andreadis, K.M.; Coss, S.P.; Durand, M.; et al. A first look at river discharge estimation from SWOT satellite observations. *Geophys. Res. Lett.* **2025**, *52*, e2024GL114185.
16. Eggleston, J.; Mason, C.; Bjerklie, D.; et al. Siting Considerations for Satellite Observation of River Discharge. *Water Resour. Res.* **2024**, *60*, e2023WR034583. https://doi.org/10.1029/2023WR034583.
17. Gleason, C.J.; Smith, L.C. Toward global mapping of river discharge using satellite images and at-many-stations hydraulic geometry. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4788–4791. https://doi.org/10.1073/pnas.1317606111.
18. Hagemann, M.W.; Gleason, C.J.; Durand, M.T. BAM: Bayesian AMHG-Manning Inference of Discharge Using Remotely Sensed Stream Width, Slope, and Height. *Water Resour. Res.* **2017**, *53*, 9692–9707. https://doi.org/10.1002/2017WR021626.
19. Durand, A.; Baron, Y.; Redjem, W.; et al. Broad Diversity of Near-Infrared Single-Photon Emitters in Silicon. *Phys. Rev. Lett.* **2021**, *126*, 083602. https://doi.org/10.1103/PhysRevLett.126.083602.
20. Gleason, C.J.; Durand, M.T. Remote Sensing of River Discharge: A Review and a Framing for the Discipline. *Remote Sens.* **2020**, *12*, 1107. https://doi.org/10.3390/rs12071107.
21. Kim, D.; Yu, H.; Lee, H.; et al. Ensemble learning regression for estimating river discharges using satellite altimetry data: Central Congo River as a Test-bed. *Remote Sens. Environ.* **2019**, *221*, 741–755. https://doi.org/10.1016/j.rse.2018.12.010.
22. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. https://doi.org/10.1029/2018WR022643.
23. Kratzert, F.; Klotz, D.; Herrnegger, M.; et al. Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resour. Res.* **2019**, *55*, 11344–11354. https://doi.org/10.1029/2019WR026065.
24. Mosavi, A.; Ozturk, P.; Chau, K.-W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. https://doi.org/10.3390/w10111536.
25. Reichstein, M.; Camps-Valls, G.; Stevens, B.; et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. https://doi.org/10.1038/s41586-019-0912-1.
26. Braca, G. *Stage-Discharge Relationships in Open Channels: Practices and Problems*; Dipartimento di Ingegneria Civile e Ambientale, Università di Trento: Trento, Italy, 2008.
27. Chen, Z.; Jiang, F.; Cheng, Y.; et al. XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud. In Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 15–17 January 2018; pp. 251–256.
28. Cheng, M.; Fang, F.; Kinouchi, T.; et al. Long lead-time daily and monthly streamflow forecasting using machine learning methods. *J. Hydrol.* **2020**, *590*, 125376. https://doi.org/10.1016/j.jhydrol.2020.125376.
29. Fang, W.; Ren, K.; Liu, T.; et al. An evaluation of random forest based input variable selection methods for one month ahead streamflow forecasting. *Sci. Rep.* **2024**, *14*, 29766. https://doi.org/10.1038/s41598-024-81502-y.
30. Kratzert, F.; Klotz, D.; Brenner, C.; et al. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.
31. Nearing, G.S.; Kratzert, F.; Sampson, A.K.; et al. What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resour. Res.* **2021**, *57*, e2020WR028091. https://doi.org/10.1029/2020WR028091.
32. Du, B.; Jin, T.; Liu, D.; et al. Accurate Discharge Estimation Based on River Widths of SWOT and Constrained At-Many-Stations Hydraulic Geometry. *Remote Sens.* **2023**, *15*, 1672. https://doi.org/10.3390/rs15061672.
33. Yoon, Y.; Durand, M.; Merry, C.J.; et al. Estimating river bathymetry from data assimilation of synthetic SWOT measurements. *J. Hydrol.* **2012**, *464–465*, 363–375. https://doi.org/10.1016/j.jhydrol.2012.07.028.
34. Frasson, R.P.D.M.; Durand, M.T.; Larnier, K.; et al. Exploring the Factors Controlling the Error Characteristics of the Surface Water and Ocean Topography Mission Discharge Estimates. *Water Resour. Res.* **2021**, *57*, e2020WR028519. https://doi.org/10.1029/2020WR028519.
35. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
36. Durand, M.; Neal, J.; Rodríguez, E.; et al. Estimating reach-averaged discharge for the River Severn from measurements of river water surface elevation and slope. *J. Hydrol.* **2014**, *511*, 92–104. https://doi.org/10.1016/j.jhydrol.2013.12.050.
37. Tarpanelli, A.; Mondini, A.C.; Camici, S. Effectiveness of Sentinel-1 and Sentinel-2 for flood detection assessment in Europe. *Nat. Hazards Earth Syst. Sci.* **2022**, *22*, 2473–2489. https://doi.org/10.5194/nhess-22-2473-2022.
38. Solomatine, D.P.; Ostfeld, A. Data-driven modelling: Some past experiences and new approaches. *J. Hydroinformatics* **2008**, *10*, 3–22. https://doi.org/10.2166/hydro.2008.015.
39. McMillan, H.K.; Westerberg, I.K.; Krueger, T. Hydrological data uncertainty and its implications. *Wiley Interdiscip. Rev. Water* **2018**, *5*, e1319. https://doi.org/10.1002/wat2.1319.

40. Bargam, B.; Boudhar, A.; Kinnard, C.; et al. Evaluation of the support vector regression (SVR) and the random forest (RF) models accuracy for streamflow prediction under a data-scarce basin in Morocco. *Discov. Appl. Sci.* **2024**, *6*, 306. https://doi.org/10.1007/s42452-024-05994-z.
41. He, M.; Xu, X.; Wu, S.; et al. Multi-step ahead forecasting of daily streamflow based on the transform-based deep learning model under different scenarios. *Sci. Rep.* **2025**, *15*, 5451. https://doi.org/10.1038/s41598-025-89837-w.
42. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
43. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; et al. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. https://doi.org/10.13031/2013.23153.
44. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
45. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE). *Clim. Res.* **2005**, *30*, 79–82. https://doi.org/10.3354/cr030079.
46. Karpatne, A.; Atluri, G.; Faghmous, J.H.; et al. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168.
47. Durand, M.; Gleason, C.J.; Garambois, P.A.; et al. An intercomparison of remote sensing river discharge estimation algorithms from measurements of river height, width, and slope. *Water Resour. Res.* **2016**, *52*, 4527–4549. https://doi.org/10.1002/2015WR018434.
48. Pavelsky, T.M.; Durand, M.T.; Andreadis, K.M.; et al. Assessing the potential global extent of SWOT river discharge observations. *J. Hydrol.* **2014**, *519*, 1516–1525. https://doi.org/10.1016/j.jhydrol.2014.08.044.
49. NASA Jet Propulsion Laboratory. *Surface Water and Ocean Topography (SWOT) science requirements document (JPL D-61923, Rev. B)*; California Institute of Technology: Pasadena, CA, USA, 2018.
50. Shapley, L.S. A Value for n-Person Games. In *Contributions to the Theory of Games II*; Kuhn, H.W., Tucker, A.W., Eds.; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–317.
51. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Von Luxburg, U., Bengio, S., et al., Eds.; NIPS: La Jolla, CA, USA, 2017; pp. 4765–4774.
52. Durand, M.; Andreadis, K.M.; Alsdorf, D.E.; et al. Estimation of bathymetric depth and slope from data assimilation of swath altimetry into a hydrodynamic model. *Geophys. Res. Lett.* **2008**, *35*, 2008GL034150. https://doi.org/10.1029/2008GL034150.
53. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; et al. *Time Series Analysis: Forecasting and Control*, 5th ed.; Wiley: Hoboken, NJ, USA, 2015.
54. Kratzert, F.; Klotz, D.; Shalev, G.; et al. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.
55. Fu, L.; Pavelsky, T.; Cretaux, J.; et al. The Surface Water and Ocean Topography Mission: A Breakthrough in Radar Remote Sensing of the Ocean and Land Surface Water. *Geophys. Res. Lett.* **2024**, *51*, e2023GL107652. https://doi.org/10.1029/2023GL107652.