DOI: **10.37943/24DKYV6003**

Alexandr Neftissov
PhD, Associate Professor, Rectorate for Science and Innovation
alexandr.neftissov@astanait.edu.kz, orcid.org/0000-0003-4079-2025
Academy of Physical Education and Mass Sports, Kazakhstan
PhD, Associate Professor, Researcher, Scientific-Innovation Center Industry 4.0
Astana IT University, Kazakhstan
Tetyana Honcharenko
Doctor of Technical Sciences, Professor, Head of Department of Information Technologies
iust511@ukr.net, orcid.org/0000-0003-2577-6916
Kyiv National University of Construction and Architecture, Ukraine
Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor, Vice-Rector for Science and Innovations
a.b@astanait.edu.kz, orcid.org/0000-0001-9548-1959
Astana IT University, Kazakhstan
Professor Department of Information Technologies,
Kyiv National University of Construction and Architecture, Ukraine
Ilyas Kazambayev
Master`s degree, Acting Director of Scientific-Innovation Center Industry 4.0
i.kazambayev@astanait.edu.kz, orcid.org/0000-0003-0850-7490
Astana IT University, Kazakhstan
Serhii Dolhopolov
PhD Student, Junior Researcher, Assistant Lecturer at the Department of Information Technologies
dolhopolov@icloud.com, orcid.org/0000-0001-9418-0943
Kyiv National University of Construction and Architecture, Ukraine

ENSEMBLE MACHINE LEARNING FOR GLOBAL HYDROLOGICAL PREDICTION

Abstract: Accurate global hydrological prediction is vital for sustainable water management but is often hindered by data complexity and fragmentation. This study introduces an advanced machine learning framework to predict long-term average discharge using widely available global hydrological station metadata, aiming to develop a highly accurate and generalizable model for large-scale water resource assessment. The methodology utilized the Global Runoff Data Centre (GRDC) dataset, applying extensive feature engineering to station characteristics and a logarithmic transformation to the discharge variable. A diverse set of algorithms was trained, including a custom deep neural network with specialized architecture and several gradient boosting machines. These individual models were then integrated into a final Meta Ensemble model through an optimized weighting strategy to maximize predictive performance. The framework was rigorously validated on an independent test set. The Meta Ensemble model demonstrated superior predictive power, achieving a Coefficient of Determination (R^2) of 0.954. This performance significantly surpassed that of both baseline methods and the individual advanced models. Analysis of the results confirmed that the model learned hydrologically meaningful relationships, identifying catchment area and geographical location as the most influential predictors. The findings confirm that a data-driven ensemble framework can accurately predict key hydrological characteristics using only station metadata. This approach offers a powerful and scalable alternative to traditional modeling, holding significant potential for water resource assessment in data-scarce regions and serving as a robust foundation for future intelligent monitoring systems.

Keywords: hydrological modeling; machine learning; ensemble learning; discharge prediction; water resources monitoring.

Introduction

Accurate hydrological prediction is a cornerstone of modern water resource management, essential for mitigating flood risks, optimizing the operation of hydraulic structures, and ensuring sustainable water supply amidst increasing climatic variability and anthropogenic pressures.

Traditional process-driven hydrological models, while offering mechanistic insights, are often constrained by extensive data requirements, limiting their applicability in many global basins. This has spurred the rapid adoption of data-driven approaches, particularly machine learning (ML), which excel at capturing complex, non-linear relationships in hydrological systems directly from observational data.

The evolution of ML in hydrology has seen a progression from single, often "black-box," models to more sophisticated and interpretable frameworks. Early applications demonstrated the viability of various architectures. For instance, M. Almetwally Ahmed and S. Samuel Li proposed a model based on the Group Method of Data Handling (GMDH), which generates explicit polynomial equations, offering greater transparency compared to typical neural networks [1]. Similarly, comparative studies by Amin Asadollahi, Ajay Kalra, and colleagues confirmed that with careful hyperparameter tuning, models like Artificial Neural Networks (ANN) can achieve high accuracy in data-scarce environments, often outperforming alternatives like Support Vector Machines (SVM) in capturing peak flows [2].

A significant trend in the field is the development of ensemble and hybrid models designed to overcome the limitations of individual algorithms. One prominent approach is to leverage the strengths of multiple models by combining their outputs. Alexandr Neftissov, Andrii Biloshchytskyi, and co-authors developed a Meta Ensemble model to estimate long-term average (LTA) discharge on a global scale using only static station metadata from the GRDC database [3]. This work demonstrated that by combining a custom deep neural network with several gradient boosting machines, it is possible to create a highly accurate ($R^2 = 0.954$) and scalable tool for water resource assessment in ungauged basins. The concept of stacking, a more advanced ensemble technique, was explored by Mingshen Lu, Lei Cheng, and their team, who used an attention mechanism as a meta-model to adaptively weight the predictions of base learners (Random Forest, AdaBoost, XGBoost) [4]. Their attention-based stacking model significantly improved runoff forecasting accuracy by dynamically learning the complementary strengths of its components.

Another powerful paradigm is the hybrid integration of ML with traditional physics-based models. Instead of replacing mechanistic models, ML can be used to correct their errors. Liyao Peng, Jian Tong, and their team proposed a Bayesian ensemble learning-based correction (BELC) scheme that uses a suite of ML models to post-process and improve the forecasts from a conceptual hydrological (XAJ) model [5]. Similarly, Jin-Cheng Fu, Wen-Cheng Liu, and collaborators developed a framework that integrates a 1D unsteady flow model with Multiple Additive Regression Trees (MART) and an Ensemble Kalman Filter (EnKF) for real-time data assimilation and forecast correction, showing how ML can be deeply embedded into operational physical models [6]. While these hybrid approaches show great promise, direct comparisons reveal a fundamental trade-off. A study by Yuhao Zhou, Jing Pan, and Guangcheng Shao demonstrated that a well-calibrated, physics-based Two-Dimensional Slope Hydrodynamic Model (TDSHM) could achieve superior accuracy and interpretability for runoff prediction compared to standalone LSTM and CNN models, particularly in scenarios requiring detailed mechanistic insights [7].

Handling the non-stationarity inherent in hydrological time series is another critical challenge. A common and effective approach is to first decompose the signal into more stationary components before applying predictive models. Xiaolong Kang and his collaborators used signal processing techniques to identify multi-scale cycles and abrupt change points in annual runoff series before applying hybrid LSTM-RF and LSTM-CNN models for prediction [8]. A more sophisticated, multi-layered "secondary decomposition" strategy was proposed by Huaibin Wei, Jing Liu, and colleagues, who used CEEMDAN followed by VMD to deconstruct complex runoff signals, allowing for the targeted application of different ML models (LSTM and Informer) to components with varying characteristics [9]. The integration of interpretable predictor selection with decomposition was demonstrated by Kaiqiang Yong, Bing Gao, and

their team, who used an XGBoost-SHAP method to identify influential large-scale climate indices for their MODWT-LSTM forecasting model [10].

The success of any ML model is critically dependent on the quality of its inputs, encompassing both the raw data and the engineered features. The issue of inherent uncertainty in hydrological data was explored by Nick Martin and Jeremy White, who advocated for Data Assimilation (DA) as a formal framework to mitigate the risks of overfitting by explicitly accounting for observation error [11]. The uncertainty in the input data itself was highlighted by Shuanglong Chen, Heng Yang, and Hui Zheng in their intercomparison of global reanalysis datasets, which revealed that model calibration had a more profound impact on accuracy than the choice of meteorological forcing data [12]. The importance of feature engineering has been demonstrated across various water science domains, from the use of graph theory to extract topological features for water distribution network design [13] to the application of ensemble models for the spatial downscaling of satellite-derived groundwater [14] and river flow data [15]. Furthermore, the interpretability of "black-box" models remain a key concern for their practical adoption. To this end, Sheng He, Xuefeng Sang, and collaborators integrated SHAP into their ensemble ML framework for discharge estimation at a sluice station, providing crucial insights into feature importance and enhancing trust in the model's predictions [16].

Recent research continues to push the boundaries of deep learning architectures. Alina Bărbulescu and Liu Zhen showed that LSTMs are particularly adept at modeling hydrological systems that have undergone significant anthropogenic changes [17], while Habtamu Alemu Workneh and Manoj K. Jha demonstrated that simpler CNNs can outperform more complex recurrent architectures when combined with effective feature selection like PCA [18]. To address the degradation of accuracy over longer lead times, Jianze Huang, Xitian Cai, and colleagues developed a coupled SA-CNN-BiLSTM model that provided both high accuracy and robust uncertainty quantification for multi-day forecasts [19]. Novel applications have also emerged, such as the work by Wei Liu, Peng Zou, and their team, who used a BiGRU network to accurately compute discharge time series using only water surface elevation as input, offering an alternative to traditional rating curves [20]. Finally, the exploration of cutting-edge architectures like the Temporal Fusion Transformer (TFT) by Rafael Francisco and José Pedro Matos has shown great promise, demonstrating not only high deterministic accuracy but also an inherent ability to provide probabilistic forecasts, which are crucial for risk-informed decision-making [21]. A conceptual link to the "digital twin" paradigm, as explored in the construction industry by Serhii Dolhopolov, Tetyana Honcharenko, and their team, suggests that the ultimate goal of these advanced monitoring systems is to create comprehensive, dynamic digital replicas of water resource systems [22].

Aim and Objectives of the Study

Despite significant advancements, a clear research gap remains in developing a unified, data-driven framework that can accurately estimate key hydrological characteristics on a global scale using readily available, static metadata. While many studies focus on dynamic forecasting with time-series data, a robust and scalable tool for baseline water resource assessment in data-scarce and ungauged basins is critically needed.

The primary aim of this research is to develop and validate a novel, high-performance Meta Ensemble machine learning framework capable of accurately estimating long-term average discharge at hydrological stations worldwide, relying solely on globally available station metadata.

To achieve this aim, the following objectives were established:

1. To develop an integrated data processing and feature engineering pipeline to transform raw global hydrological station metadata (from the GRDC database) into a rich and informative set of predictors.

2. To design and optimize a diverse suite of advanced machine learning models, including a custom-designed deep neural network and several state-of-the-art gradient boosting machines, for the prediction task.
3. To construct and validate a high-performance Meta Ensemble model that synergistically combines the predictions of the individual models to maximize accuracy and generalization.
4. To interpret the final model using explainable AI techniques (SHAP) to identify the key geographical and physical catchment attributes that most significantly influence long-term average discharge, ensuring the model's logic is hydrologically plausible.
5. To demonstrate the potential of this data-driven methodology as a scalable and cost-effective tool for large-scale water resource assessment, particularly for preliminary assessments of hydraulic structures in ungauged or data-limited regions.

Methods and Materials

Data Source and Description

The empirical basis for this study is the Global Runoff Data Centre (GRDC) Station Catalogue. The GRDC, operating under the auspices of the World Meteorological Organization (WMO), serves as a central repository for worldwide river discharge data and associated station metadata. This globally comprehensive dataset is an invaluable resource for large-scale hydrological research, encompassing a wide diversity of climatic and hydrological regimes. The initial dataset contained 10,978 station records from across the globe. Each record is characterized by a set of attributes describing the station's geographical location (latitude, longitude), physical catchment properties (area, altitude), and key characteristics of its historical data records (period of operation). A summary of the primary variables selected from this catalogue for use in the study is presented in Table 1.

Table 1. Key Variables from the GRDC Station Catalogue Used in the Study

Variable	**Description**
wmo_reg	WMO region code
sub_reg	WMO subregion code
lat, lon	Geographical coordinates (decimal degrees)
area	Catchment size (km^2)
altitude	Altitude of gauge zero (m)
t_start, t_end	Start and end year of the observation period
t_yrs	Total length of the observation period (years)
lta_discharge	Long-term average discharge (m^3/s)

Data Preprocessing and Feature Engineering

A systematic, multi-step pipeline was implemented to transform the raw GRDC metadata into a clean, structured, and feature-rich dataset suitable for machine learning modeling. The initial stage involved data cleaning, which included converting variables stored as text (e.g., lta_discharge) to numeric formats and implementing a strategy for handling missing values present in the raw data.

A critical step in preparing the data was the transformation of the target variable, the Long-Term Average (LTA) discharge. Hydrological variables like discharge are well-known to exhibit highly skewed distributions, with a large number of stations having low to moderate flow and a long tail of stations with very high flows. This characteristic, confirmed during exploratory data analysis, can violate the assumptions of many regression algorithms and disproportionately weight the model towards predicting high-magnitude events. To address this, a logarithmic transformation using the function $log(1+x)$ was applied. This function is particularly suitable as it stabilizes variance across the range of values and transforms the skewed distribution into a

more symmetric, approximately normal (Gaussian) distribution, which is more amenable to modeling. The profound effect of this transformation is illustrated in Figure 1. The resulting variable, *lta_discharge_log*, was used as the prediction target for all subsequent modeling tasks.

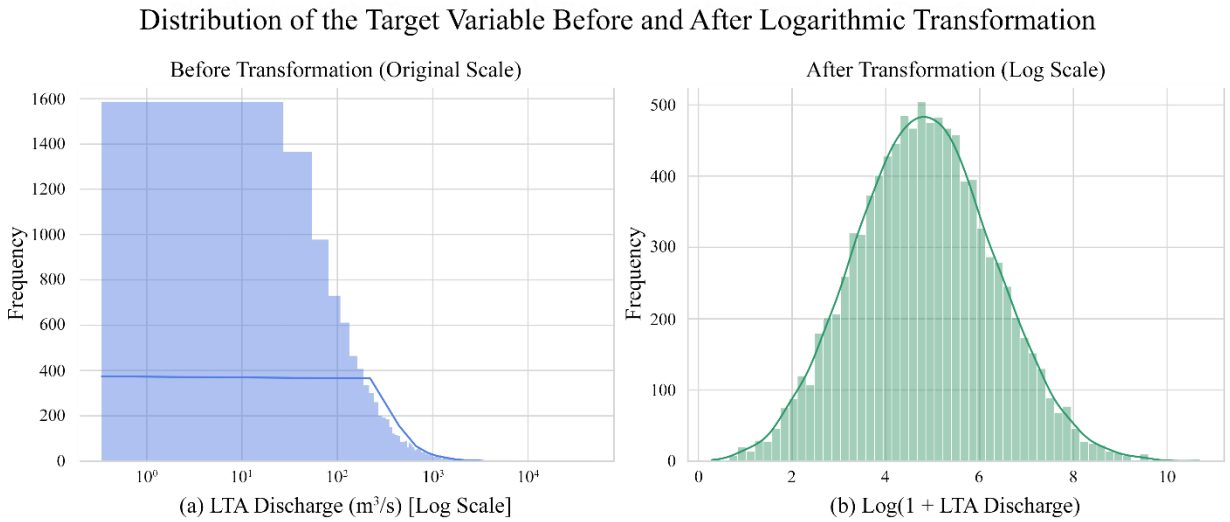

Figure 1. Distribution of the target variable, Long-Term Average (LTA) Discharge: (**a**) before and (**b**) after logarithmic transformation

Following the transformation of the target variable, a comprehensive feature engineering process was undertaken to generate new, more informative predictors from the base metadata. This process was crucial for enabling the models to capture complex non-linear relationships and interactions. The engineered features included various non-linear transformations of the catchment area (logarithmic and square root) to better represent its scaling effects on discharge. To handle the cyclical nature of geographical coordinates, sine and cosine transformations of latitude and longitude were computed. Furthermore, interaction terms between key predictors (e.g., *area* multiplied by *wmo_reg*) and ratio features (e.g., *area_to_altitude_ratio*) were generated to model combined effects. Temporal attributes, such as the operational lifetime of each station, were also calculated to capture information related to data record maturity. A final set of 33 features was chosen for the predictive modeling phase, based on a rigorous selection process that combined correlation analysis with the target variable and model-based feature importance rankings from preliminary models.

Predictive Modeling Framework

The core of this study is a hybrid, hierarchical modeling strategy that leverages the strengths of multiple diverse machine learning paradigms to maximize predictive accuracy and robustness. The overall workflow of this framework is conceptually illustrated in Figure 2. It is designed as a stacking-like ensemble, where the predictions of several powerful base models are intelligently combined by a higher-level meta-model.

The framework is composed of two main layers. The first layer consists of several diverse, individual models, referred to as base learners. This set includes a custom Advanced Neural Network (NN), designed using TensorFlow/Keras. Recognizing the paramount importance of the catchment area, the NN architecture features a specialized, separate processing path for this feature, allowing the model to learn its influence directly and with dedicated parameters. This path is then concatenated with the main network path, which consists of multiple hidden layers with residual connections (inspired by ResNet architectures) to facilitate the training of a deep network and avoid issues like vanishing gradients. In addition to the neural network, a suite of state-of-the-art Gradient Boosting Machines (GBMs) was trained, including XGBoost, LightGBM, and CatBoost. These tree-based ensemble algorithms were selected for

their proven high performance on tabular data and their ability to capture complex non-linear interactions and feature dependencies automatically.

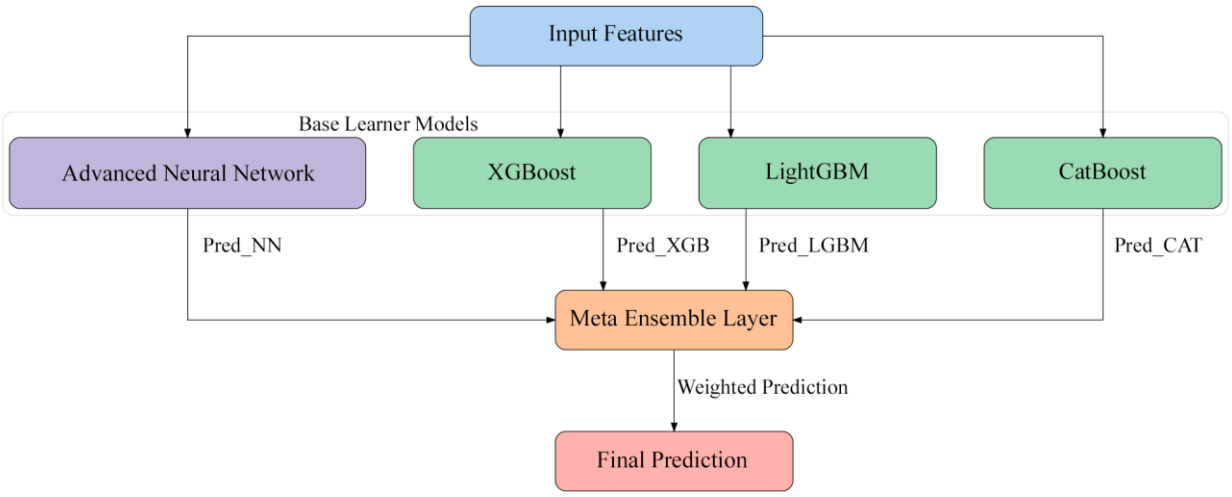

Figure 2. Conceptual Architecture of the Meta Ensemble Modeling Framework, showing the flow from input features through the base learner and meta-ensemble layers

The cornerstone of the predictive strategy is the second layer, the Meta-Ensemble Model. This final model aggregates the predictions from the best-performing individual base learners. Instead of a simple average, it employs a more sophisticated weighted combination where the weights are themselves optimized on a validation set to maximize the final R^2 score. This strategy allows the final model to capitalize on the unique strengths and perspectives of each base model – for instance, combining the powerful non-linear function approximation of the neural network with the robust handling of tabular data by the GBMs – often leading to performance superior to any single constituent.

Model Training and Validation Strategy

A rigorous validation protocol was established to ensure the development of a robust and generalizable model, free from overfitting. The final, cleaned dataset of 10,586 samples was definitively partitioned into a primary training set (80%, or 8,468 samples) and an independent test set (20%, or 2,118 samples). The test set was strictly held out from all training and tuning activities and was used exclusively for the final, unbiased evaluation of the fully trained models.

During the model development and hyperparameter tuning phase, all optimizations were performed exclusively on the 80% training set. This involved using internal validation splits to guide the hyperparameter search (e.g., via Bayesian optimization with Optuna) and to implement early stopping for the neural network models to prevent overfitting. For building robust ensemble components, such as the Neural Ensemble, K-fold cross-validation (with k=5) was employed. This iterative process of training and validating on different subsets of the training data ensures that the selected hyperparameters and model architectures are robust and not overfitted to a specific data partition, thereby enhancing their generalization potential.

Model Evaluation and Interpretation

The performance of all predictive models was rigorously evaluated using a set of standard statistical metrics. The Coefficient of Determination (R^2) was used to measure the proportion of variance in the target variable explained by the model. Root Mean Squared Error (RMSE) was calculated to assess the typical magnitude of prediction errors on the log-transformed scale. Additionally, the Mean Absolute Error (MAE) was computed on both the log-transformed scale and, crucially, on the original discharge units (m^3/s) by back-transforming the predictions and observations. The mathematical formulas for these metrics are provided in Equations (1) – (4).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (2)$$

$$MAE_{log} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (3)$$

$$MAE_{orig} = \frac{1}{n}\sum_{i=1}^{n}|\text{expm1}(y_i) - \text{expm1}(\hat{y}_i)|, \quad (4)$$

where y represents the observed values, $\hat{y}$ represents the predicted values, $\bar{y}$ is the mean of observed values, and n is the number of samples.

To move beyond simple performance metrics and understand why the Meta Ensemble model is successful, the SHAP (SHapley Additive exPlanations) framework was employed for model interpretation. SHAP is a game-theoretic, model-agnostic approach that explains the output of any machine learning model by assigning each feature an importance value – the SHAP value – for each individual prediction. This powerful technique provides both global and local interpretability. It allowed for a detailed analysis of which features had the most significant impact on discharge estimation across the entire dataset, and it helped to ensure that the model was learning hydrologically meaningful and physically plausible relationships, thereby building confidence in its predictions.

Results

Comparative Model Performance

A comprehensive evaluation was conducted to compare the predictive capabilities of all developed models, from simple baselines to advanced ensembles. The key performance metrics – Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) – were calculated for both the training and testing datasets to assess accuracy and generalization.

The results, summarized in Table 2, reveal a clear performance hierarchy. The baseline linear model, ElasticNet, performed poorly (Test $R^2 \approx 0.25$), confirming its inability to capture the complex, non-linear relationships inherent in the global hydrological data. Standard ensemble methods like RandomForest and GradientBoosting offered significant improvements but were ultimately surpassed by the more sophisticated, custom-designed architectures.

Table 2. Performance Metrics of All Evaluated Models on the Training and Testing Sets

Model	Train R^2	Test R^2	Train RMSE	Test RMSE	Train MAE (m^3/s)	Test MAE (m^3/s)
ElasticNet	0.25	0.249	1.776	1.785	449.6	1904236.58
GradientBoosting	0.841	0.833	0.818	0.843	173.35	239.62
RandomForest	0.983	0.889	0.268	0.688	75.35	154.61
LightGBM	0.927	0.895	0.553	0.667	129.21	201.36
CatBoost	0.952	0.901	0.449	0.648	95.67	131.45
XGBoost	0.972	0.903	0.341	0.641	70.25	122.9
Neural Network	0.935	0.916	0.524	0.597	83.21	105.82
Neural Ensemble	0.951	0.932	0.456	0.538	70.46	89.73
Boosted Neural Network	0.963	0.941	0.396	0.501	65.78	78.41
Meta Ensemble	0.975	0.954	0.324	0.442	62.13	71.28

The advanced models consistently demonstrated superior performance. The custom Neural Network and the individual Gradient Boosting Machines (XGBoost, CatBoost, LightGBM) all performed strongly, achieving Test R^2 scores around or above the target of 0.9. However, the highest level of performance was consistently achieved by the ensemble strategies that integrated these advanced models. The Neural Ensemble and the Boosted Neural Network both showed excellent results, but the Meta Ensemble model, which combines predictions from the top-performing models with optimized weights, yielded the best overall performance, achieving a Test R^2 of 0.954.

This comparative performance is visually summarized in Figure 3. The bar chart clearly illustrates that all advanced models, and particularly the ensemble methods, successfully met and exceeded the predefined performance target of $R^2 = 0.9$. A crucial observation is the small gap between the training (blue) and testing (light red) scores for the top models, especially the Meta Ensemble. This indicates good generalization and a low risk of overfitting, confirming that the model has learned the underlying patterns in the data rather than memorizing the training examples.

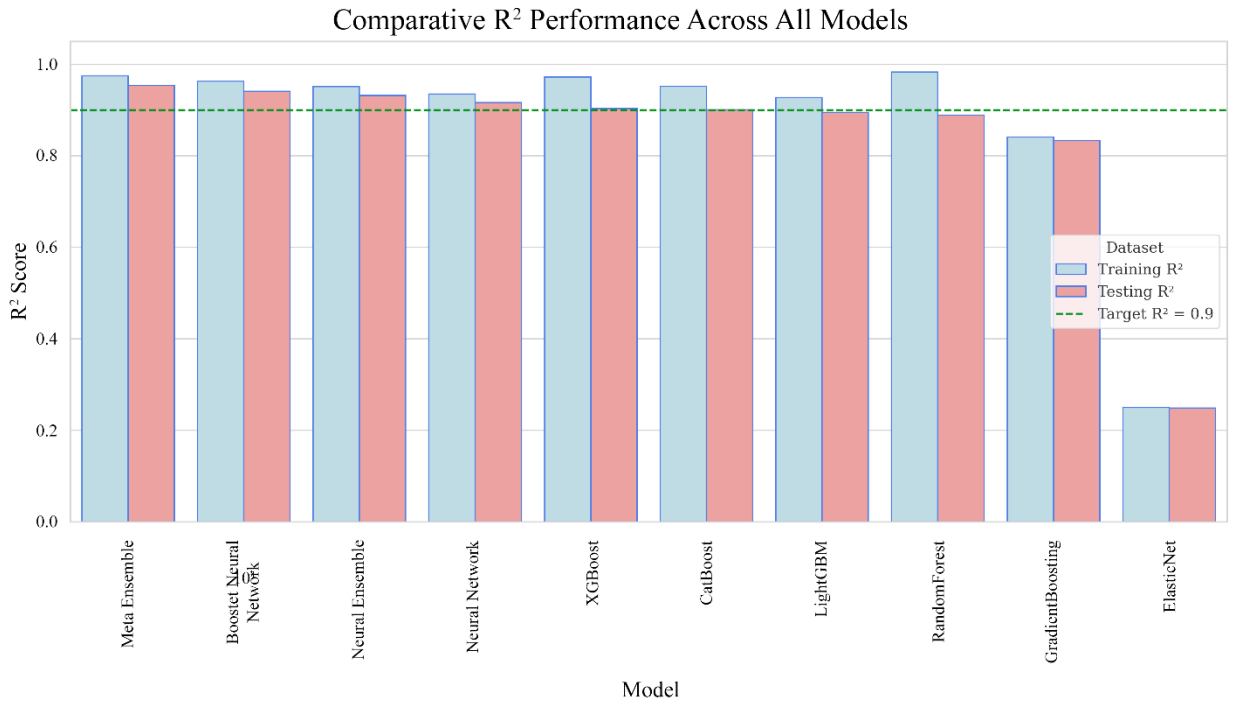

Figure 3. Comparative R^2 Performance Across All Models. The chart displays both Training R^2 and Testing R^2 scores, with the dashed line indicating the performance target of 0.9

Based on its superior performance across all key evaluation metrics on the independent test set, the Meta Ensemble model was selected as the final, optimal model for the hydrological prediction task.

Final Model Performance and Interpretation

A detailed validation of the final Meta Ensemble model was performed to assess its accuracy and identify any potential systematic biases. Figure 4 presents a scatter plot of the model's predicted versus actual discharge values for the independent test set, with points colored by their absolute prediction error. The data points cluster tightly around the 1:1 line (dashed red line), indicating a very strong correlation and high level of agreement between the model's predictions and the observed data across several orders of magnitude. The color mapping

visually confirms that the vast majority of predictions (darker points) have a low absolute error. While some larger errors (lighter yellow points) are visible, they are infrequent and typically associated with the highest discharge values, a common challenge in hydrological modeling. This visualization provides strong qualitative and quantitative evidence of the model's high accuracy.

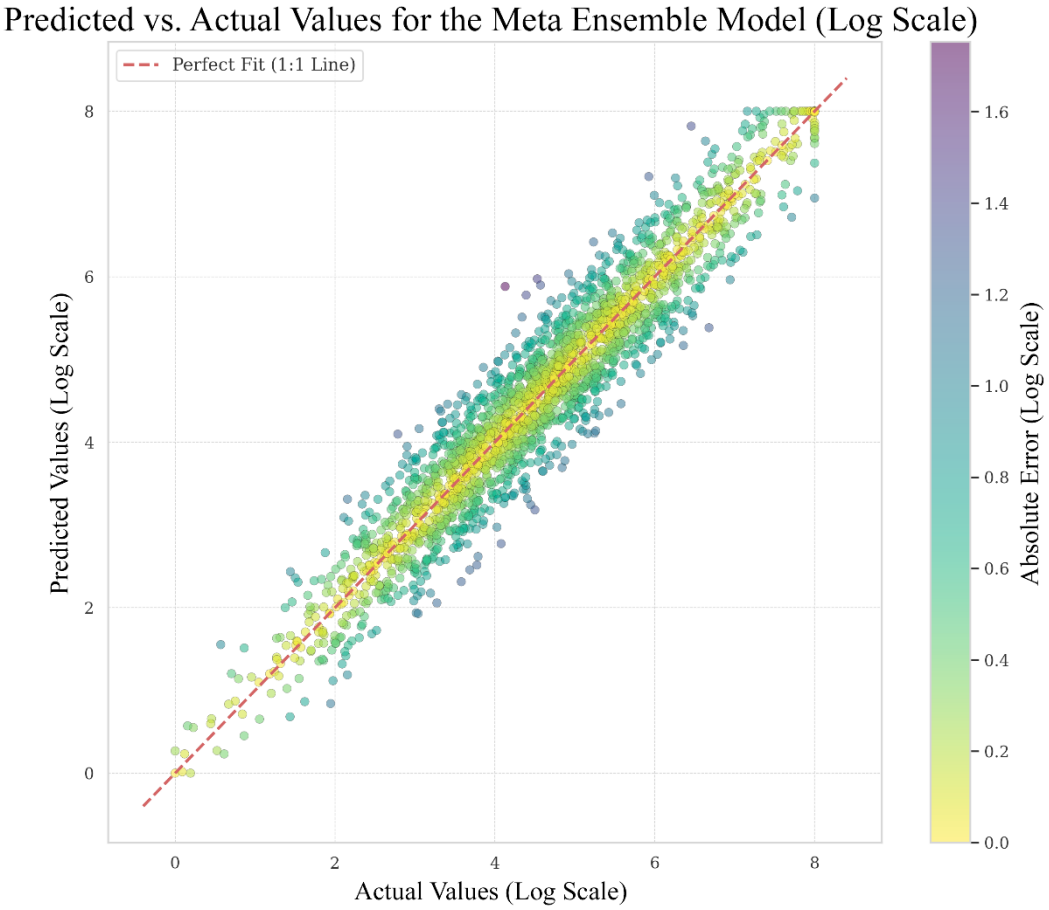

Figure 4. Predicted vs. Actual Values for the Meta Ensemble Model on the Test Set. The plot is on a logarithmic scale, and points are colored by absolute prediction error.

To further understand the model's behavior and validate its underlying logic, an interpretability analysis was conducted using the SHAP framework. Figure 5 presents a global feature importance plot, which ranks the input features based on their mean absolute SHAP value across all samples in the test set. This analysis unequivocally identifies log-transformed catchment area (*area_log*) as the single most dominant predictor, having a significantly larger impact than any other feature. This finding aligns perfectly with fundamental hydrological principles, where catchment area is the primary driver of discharge volume. Following *area_log*, geographical location features – such as longitude (*long*), latitude (*lat*), and regional identifiers (*sub_reg*) – are the next most important predictors. This highlights the model's ability to effectively learn and apply spatial context, essentially performing a form of implicit regionalization to account for climatic and geological variability not explicitly included as features. The significant contribution of these physically meaningful variables provides strong evidence that the model's high accuracy is not a "black box" artifact but is grounded in hydrologically plausible relationships learned from the data.

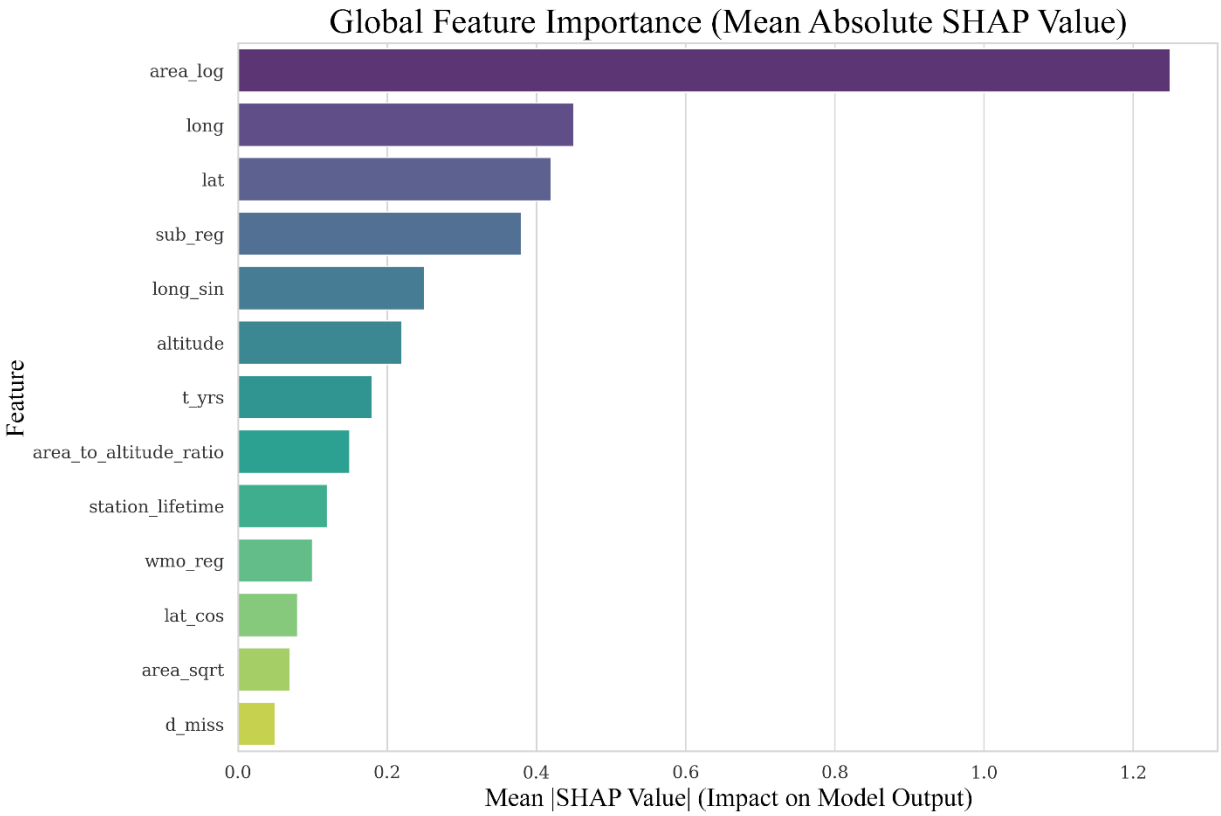

Figure 5. Global Feature Importance Ranking from SHAP Analysis. Features are ranked by their mean absolute SHAP value, representing their overall impact on model predictions.

Collectively, the performance metrics and interpretability analysis validate the Meta Ensemble model as a highly accurate and robust tool. The results demonstrate that a data-driven approach, when carefully designed and rigorously validated, can successfully model complex, large-scale hydrological phenomena based on readily available metadata.

Discussion

Interpretation of Model Performance and Feature Importance

The superior performance of the Meta Ensemble model over all individual learners, including a sophisticated custom Neural Network and state-of-the-art GBMs, aligns with the well-established consensus in machine learning: ensemble methods tend to be more robust and accurate by averaging out the biases and variances of individual models. The diversity of the base learners (a neural network and several tree-based models) was a key strength, likely allowing the ensemble to capture different facets of the complex, non-linear relationships between station metadata and LTA discharge.

The model interpretability analysis using SHAP provides crucial validation of the model's underlying logic. The overwhelming importance of catchment area (area and its transformations) as the primary predictor (Figure 5) is consistent with fundamental hydrological principles, confirming that the model has learned a physically plausible relationship. The significant contribution of geographical coordinates (lat, long) and regional identifiers (wmo_reg, sub_reg) highlights the model's ability to effectively learn and apply spatial context. It essentially performs a form of implicit regionalization, learning to account for climatic and geological variability that was not explicitly included as input features. This capacity to learn from spatial context is a key strength of applying machine learning to large, geographically

diverse datasets and provides confidence that the model's high accuracy is not a "black box" artifact but is grounded in hydrologically meaningful principles.

Comparison with Existing Research

While a vast body of literature exists on applying machine learning to hydrology, most studies focus on dynamic streamflow forecasting using time-series data as inputs. Our work addresses a different but equally important problem: estimating a static, long-term characteristic (LTA) from metadata. This task is more aligned with regionalization studies and methods for Prediction in Ungauged Basins (PUB), where the goal is to transfer information from gauged to ungauged locations based on their physical characteristics. Compared to traditional regression-based regionalization methods, our ensemble ML approach offers a more flexible and powerful framework for capturing complex, non-linear relationships on a global scale. Unlike systems focused purely on real-time data from IoT sensors, our approach leverages historical, aggregated information embedded in the GRDC catalogue, making it suitable for strategic planning and large-scale assessment rather than operational forecasting.

Practical Implications and Potential for Operationalization

The developed model has significant practical implications. It provides a robust, cost-effective tool for estimating baseline water availability in regions with sparse or non-existent gauging networks. This can be invaluable for preliminary water resource planning, climate change impact assessment, and the initial design of hydraulic structures. For example, for a proposed dam or irrigation project in a data-scarce region, the model can provide a rapid first-order estimate of LTA discharge, requiring only basic geographical and catchment information.

While this study focused on developing the predictive model, its outputs could be integrated into a broader, operational monitoring system, as conceptually outlined in the system architecture (Figure 2). In such a system, our LTA estimation model could serve two roles: (1) providing baseline "normal" discharge values against which real-time data from IoT sensors can be compared for anomaly detection, and (2) generating plausible estimates for initializing more complex hydrological models or for filling gaps in records. This illustrates a potential pathway from the strategic estimation tool developed here to a comprehensive, operational monitoring platform.

Limitations and Future Research Directions

Despite the promising results, this study has several limitations that open avenues for future research. First, the model's performance is inherently dependent on the quality and geographical representativeness of the GRDC dataset. Gaps in station coverage, particularly in Africa, South America, and parts of Asia, may limit the model's accuracy in these regions. The dataset's heterogeneity, stemming from different measurement standards and data quality across countries, also introduces unquantified uncertainty into the predictions.

Second, the model estimates a static, long-term average and does not provide dynamic, time-varying forecasts. It is therefore not suitable for short-term operational flood management. Third, the model relies on historical relationships and may struggle to adapt to non-stationary conditions driven by rapid climate change or large-scale land-use changes not captured by the input features.

Future research should address these limitations. Integrating additional data sources, such as climate reanalysis data (e.g., precipitation, temperature) and land-cover classifications, could help the model better account for climatic variability and reduce regional biases. Developing robust methods for quantifying prediction uncertainty (e.g., using quantile regression or

Bayesian neural networks) is crucial for providing risk-based information to decision-makers. Finally, exploring transfer learning could leverage the globally trained model to improve performance in specific data-scarce regions with limited local data, thereby enhancing its practical applicability.

Conclusion

This research successfully developed and validated an advanced machine learning framework for estimating long-term average discharge using globally available hydrological station metadata. The study demonstrated that a data-driven approach, leveraging a sophisticated ensemble of diverse models, can accurately estimate this key hydrological characteristic without relying on complex, time-varying simulations. The main conclusions of this work are as follows:

1. The developed Meta Ensemble model, which integrates predictions from an optimized Neural Network and several Gradient Boosting Machines, achieved excellent predictive performance on an independent test set ($R^2 = 0.954$, MAE = 71.3 m^3/s). This significantly surpasses the accuracy of both baseline methods and individual advanced models, highlighting the power of hybrid ensembling for this hydrological task.
2. Model interpretability analysis using SHAP confirmed that the model learned physically plausible relationships. It identified catchment area as the most dominant predictor, with geographical location and regional identifiers playing a crucial secondary role in capturing spatial variability. This provides confidence that the model's high accuracy is not a "black box" artifact but is grounded in hydrologically meaningful principles.
3. Rigorous data preprocessing and feature engineering were critical to the model's success. The logarithmic transformation of the skewed target variable and the creation of interaction, ratio, and transformed geographical features were essential for achieving high performance.
4. The study demonstrates that it is feasible to build a robust and scalable tool for large-scale water resource assessment using readily available global metadata. This approach offers a valuable, cost-effective alternative to traditional methods, especially for preliminary assessments in ungauged or data-scarce basins.

In summary, this work contributes a robust methodology and a high-accuracy predictive model, advancing the application of machine learning in large-scale hydrology. It provides a validated framework for estimating a fundamental hydrological characteristic, offering a powerful tool to support global and regional water resource management and planning.

Acknowledgement

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24993128 "Information-analytical system development for the transboundary water resources effective use in the Zhambyl region agricultural sector."

References

[1] Ahmed, M. A., & Li, S. S. (2024). Machine Learning Model for River Discharge Forecast: A Case Study of the Ottawa River in Canada. Hydrology, 11(9), 151. https://doi.org/10.3390/hydrology11090151.

[2] Asadollahi, A., Magar, B. A., Poudel, B., Sohrabifar, A., & Kalra, A. (2024). Application of Machine Learning Models for Improving Discharge Prediction in Ungauged Watershed: A

Case Study in East DuPage, Illinois. Geographies, 4(2), 363–377. https://doi.org/10.3390/geographies4020021.
[3] Neftissov, A., Biloshchytskyi, A., Kazambayev, I., Dolhopolov, S., & Honcharenko, T. (2025). An Advanced Ensemble Machine Learning Framework for Estimating Long-Term Average Discharge at Hydrological Stations Using Global Metadata. Water, 17(14), 2097. https://doi.org/10.3390/w17142097.
[4] Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., & Cheng, L. (2023). A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting. Water, 15(7), 1265. https://doi.org/10.3390/w15071265.
[5] Peng, L., Fu, J., Yuan, Y., Wang, X., Zhao, Y., & Tong, J. (2025). A Bayesian Ensemble Learning-Based Scheme for Real-Time Error Correction of Flood Forecasting. Water, 17(14), 2048. https://doi.org/10.3390/w17142048.
[6] Fu, J.-C., Su, M.-P., Liu, W.-C., Huang, W.-C., & Liu, H.-M. (2024). Water Level Forecasting Combining Machine Learning and Ensemble Kalman Filtering in the Danshui River System, Taiwan. Water, 16(23), 3530. https://doi.org/10.3390/w16233530.
[7] Zhou, Y., Pan, J., & Shao, G. (2025). A Comparative Study of a Two-Dimensional Slope Hydrodynamic Model (TDSHM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) Models for Runoff Prediction. Water, 17(9), 1380. https://doi.org/10.3390/w17091380.
[8] Kang, X., Yu, H., Yang, C., Tian, Q., & Wang, Y. (2025). Analysis of Evolutionary Characteristics and Prediction of Annual Runoff in Qianping Reservoir. Water, 17(13), 1902. https://doi.org/10.3390/w17131902.
[9] Wei, H., Wang, Y., Liu, J., & Cao, Y. (2023). Monthly Runoff Prediction by Combined Models Based on Secondary Decomposition at the Wulong Hydrological Station in the Yangtze River Basin. Water, 15(21), 3717. https://doi.org/10.3390/w15213717.
[10] Yong, K., Li, M., Xiao, P., Gao, B., & Zheng, C. (2025). Monthly Streamflow Forecasting for the Irtysh River Based on a Deep Learning Model Combined with Runoff Decomposition. Water, 17(9), 1375. https://doi.org/10.3390/w17091375.
[11] Martin, N., & White, J. (2024). Water Resources' AI–ML Data Uncertainty Risk and Mitigation Using Data Assimilation. Water, 16(19), 2758. https://doi.org/10.3390/w16192758.
[12] Chen, S., Yang, H., & Zheng, H. (2025). Intercomparison of Runoff and River Discharge Reanalysis Datasets at the Upper Jinsha River, an Alpine River on the Eastern Edge of the Tibetan Plateau. Water, 17(6), 871. https://doi.org/10.3390/w17060871.
[13] Bahrami Chegeni, I., Riyahi, M. M., Bakhshipour, A. E., Azizipour, M., & Haghighi, A. (2025). Developing Machine Learning Models for Optimal Design of Water Distribution Networks Using Graph Theory-Based Features. Water, 17(11), 1654. https://doi.org/10.3390/w17111654.
[14] Yuan, Y., Shen, D., Cao, Y., Wang, X., Zhang, B., & Dong, H. (2025). An Ensemble Machine Learning Approach for High-Resolution Estimation of Groundwater Storage Anomalies. Water, 17(10), 1445. https://doi.org/10.3390/w17101445.
[15] Ziadi, S., Chokmani, K., Chaabani, C., & El Alem, A. (2024). Deep Learning-Based Automatic River Flow Estimation Using RADARSAT Imagery. Remote Sensing, 16(10), 1808. https://doi.org/10.3390/rs16101808.
[16] He, S., Niu, G., Sang, X., Sun, X., Yin, J., & Chen, H. (2023). Machine Learning Framework with Feature Importance Interpretation for Discharge Estimation: A Case Study in Huitanggou Sluice Hydrological Station, China. Water, 15(10), 1923. https://doi.org/10.3390/w15101923.
[17] Bărbulescu, A., & Zhen, L. (2024). Forecasting the River Water Discharge by Artificial Intelligence Methods. Water, 16(9), 1248. https://doi.org/10.3390/w16091248.

[18] Workneh, H. A., & Jha, M. K. (2025). Utilizing Deep Learning Models to Predict Streamflow. Water, 17(5), 756. https://doi.org/10.3390/w17050756.
[19] Huang, J., Chen, J., Huang, H., & Cai, X. (2025). Deep Learning-Based Daily Streamflow Prediction Model for the Hanjiang River Basin. Hydrology, 12(7), 168. https://doi.org/10.3390/hydrology12070168.
[20] Liu, W., Zou, P., Jiang, D., Quan, X., & Dai, H. (2023). Computing River Discharge Using Water Surface Elevation Based on Deep Learning Networks. Water, 15(21), 3759. https://doi.org/10.3390/w15213759.
[21] Francisco, R., & Matos, J. P. (2024). Deep Learning Prediction of Streamflow in Portugal. Hydrology, 11(12), 217. https://doi.org/10.3390/hydrology11120217.
[22] Dolhopolov, S., Honcharenko, T., Terentyev, O., Savenko, V., Rosynskyi, A., Bodnar, N., & Alzidi, E. (2024). Multi-Stage Classification of Construction Site Modeling Objects Using Artificial Intelligence Based on BIM Technology. 2024 35th Conference of Open Innovations Association (FRUCT), 179–185. https://doi.org/10.23919/fruct61870.2024.10516383.