Khazar Journal of Water and Environment
DOI: 10.68277/kjwe.2026.11

e-ISSN: 3150-0188
khazarjournal.com

KJWE

Conformal Prediction for River Total Dissolved Solids (TDS) Prediction and Risk Quantification

Mahesh Pal[a], Erfan Abdi[b,c*], Ali Torabi Haghighi[d]

[a] *Department of Civil Engineering, National Institute of Technology, Kurukshetra, India*
[b] *Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, 51666-16471, Iran*
[c] *Water Sciences and Hydroinformatics Research Center, Khazar University, Mahsati str. 41, AZ 1096, Baku, Azerbaijan*
[d] *Water, Energy and Environmental Engineering Research Unit, University of Oulu, Oulu, Finland*

ARTICLE INFO
Review Article
Corresponding Author: Erfan Abdi, E-mail: erabdi78@gmail.com
Received: 30 August 2026 / Revised: 4 October 2026 / Accepted: 10 October 2026 / Online: 31 October 2026

Cite this article as
PAL, M., ABDI, E. & TORABI HAGHIGHI, A. (2026). Conformal Prediction for Multi-Parameter River Water-Quality Prediction and Risk Quantification. Khazar Journal of Water and Environment, 1, 159-197. DOI: 10.68277/kjwe.2026.11

ABSTRACT
Probabilistic machine learning models estimate total dissolved solids (TDS) levels and exceedance probabilities, enabling risk-based regulation and safeguarding public health. This study develops a probabilistic machine learning framework to forecast monthly TDS in the Lam Tsuen River in Hong Kong using long-term monitoring data from 2000 to 2023, split into 80% for training and 20% for testing. Four routinely measured variables, pH, salinity, turbidity, and water temperature, were used as predictors of TDS. Five gradient boosting-based algorithms (Gradient boosting, LightGB, XGBoost, CatBoost, and GPBoost) were calibrated, and predictive uncertainty was quantified using three conformal prediction schemes: Split Conformal Prediction (SplitCP), Cross Validation Plus (CV+), and Conformal Quantile Regression (CQR). Probabilistic prediction was implemented via the Treeffuser framework, enabling consistent construction of prediction intervals and exceedance probabilities. Among the tested models, GPBoost delivered the best deterministic performance, with Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and correlation coefficient (CC) of 85.968 mg/l, 59.943, and 0.902 on the test set, respectively, outperforming gradient boosting (RMSE 88.855 mg/l, MAE 61.067, CC 0.906), LightGB (RMSE 89.837, MAE 62.054, CC 0.905), XGBoost (RMSE 97.723, MAE 68.792, CC 0.885), and CatBoost (RMSE 108.758 mg/l, MAE 66.886, CC 0.840). The conformal prediction methods produced well-calibrated prediction intervals, with the narrowest and most stable bands obtained for GPBoost. SHAP analysis further showed that salinity and turbidity were the dominant predictors, followed by pH and water temperature, providing hydrochemically consistent insight into the drivers of TDS variability. These methods can be extended to river quality/quantity for probabilistic forecasting and adaptive planning.

Keywords: Water quality, Predicting probabilities, Uncertainty quantification, Gradient boosting machines, Total dissolved solids, Machine learning

1. Introduction

Rivers are among the most endangered ecosystems on Earth. They face pollution from industrial and agricultural waste, over-extraction of water, dam construction that disrupts ecosystems, and the effects of climate change (Abdi et al., 2025). As a result, the importance of rivers cannot be ignored. Predicting the quality of river water is of great importance. Among the vital necessities for human

life are the provision of drinking water and the irrigation of farms (Sattari et al., 2021; del Castillo et al., 2024; Ibrahim et al., 2026). On the other hand, these forecasts can be economically beneficial for governments. So that by predicting the quality, the necessary costs for water treatment are estimated, and additional costs are not required (Sharafi et al., 2024; Abdi et al 2026). The most important benefits of forecasting are prevention and early warning. This approach allows decision-makers to create processes to prevent crises.

Traditional statistical and deterministic models for river water quality prediction are often based on fixed parameterizations or linear hydrological relationships. They are plagued by the inherent nonlinearity and complexity of water quality dynamics, which are influenced by diverse factors such as land use, precipitation variability, pollutant sources, and seasonal variations, and are prone to greater errors in high-dimensional data or nonlinear interactions between different water quality parameters (Sattari et al., 2020). Conventional models are weak in quantifying prediction uncertainty. Most classical approaches provide point estimates without confidence intervals, leaving decision-makers uncertain about the reliability of the predictions. This lack of quantification of uncertainty reduces transparency and trust, particularly in informing water quality predictions, public health, and environmental management. River water quality prediction methods can be broadly categorized into statistical methods, such as regression and other time series methods, which are often used for basic prediction and offer interpretability, but have limited ability to capture nonlinear dynamics (Alnahit et al., 2022; Ghaemi et al., 2022; Yao et al., 2024). Machine learning, such as random forests, support vector machines, and artificial neural networks, improves prediction accuracy by handling complex relationships between quality parameters (Najwa Mohd Rizal et al., 2022; Satish et al., 2024) and hybrid models that have a certain complexity (Zhou et al., 2022; Zhao et al., 2025).

Decision tree-based models provide interpretable insights into how different environmental variables affect outcomes. Unlike black-box models such as deep neural networks, decision trees explicitly represent the hierarchy of factors that affect quality parameters. Ensemble methods built on decision trees, such as random forests and gradient boosting trees, further increase the accuracy and robustness of predictions by reducing overfitting and improving generalization across different river basins. Song and Yao (2022) proposed a hybrid ensemble learning model combining complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), fuzzy entropy (FE), and an improved ant lion optimizer (IALO) to optimize LSTM hyperparameters for weekly dissolved oxygen (DO) prediction. Tested on data from the Yongjiang and Beijiang Rivers in the Pearl River Basin, their CEEMDAN-IALO-LSTM model achieved average MAE of 0.42, MAPE of 5.37%, and RMSE of 0.53, representing reductions of 50.24%, 47.66%, and 49.66%, respectively, compared to baseline

LSTM. The IALO algorithm incorporated chaos initialization, Cauchy mutation, and opposition-based learning (OBL) to avoid local optima, demonstrating that hybrid decomposition-optimization frameworks can substantially improve water quality forecasting accuracy. Alnahit et al. (2022) compared two machine learning methods, Random Forest and Boosted Regression Tree, to predict water quality parameters, total nitrogen, total phosphorus, and turbidity in 97 watersheds of the Southeast Atlantic USA. Islam et al. (2024) addressed the challenge of improving water quality monitoring by integrating advanced machine learning (ML) methods into Water Quality Index (WQI) prediction. Using physicochemical parameters from various water sources, models including Gradient Boosting, Random Forest, and XGBoost were trained and evaluated for accuracy. Gradient Boosting achieved the highest accuracy (96%). Shams et al. (2024) focused on predicting the WQI and Water Quality Classification (WQC) using various ML models with parameter optimization via grid search. Four classification models, Random Forest, XGBoost, Gradient Boosting, and AdaBoost, were employed for WQC prediction, with Gradient Boosting achieving the highest accuracy of 99.5%. Abbas et al. (2024) developed an AI-based forecasting model using 422 physicochemical water samples from Mirpurkash, Sindh, to predict WQI efficiently and reduce calculation errors. Various machine learning classifiers were tested, with Random Forest and Gradient Boosting models achieving the highest accuracy (~99%), outperforming SVM, XGBoost, K-Nearest Neighbors (KNN), and Decision Tree models. Liu et al. (2025) on the Daluxi River watershed compared Kernel Principal Component Analysis (KPCA) and traditional Principal Component Analysis (PCA) for water quality parameter analysis. KPCA was more effective in capturing nonlinear relationships, extracting four principal pollution sources explaining about 79% of the variance, higher than PCA. To link these components back to the original water quality data, six machine learning models were applied, with Gradient Boosting Decision Tree (GBDT) showing superior regression performance ($R^2 = 0.988$). Adnan et al. (2025) developed a novel integration of the random vector functional link (RVFL) network with an enhanced remora optimization algorithm (EROA) for monthly streamflow prediction in the Kunhar River Basin, Pakistan. The EROA improved upon the standard remora optimization algorithm by incorporating opposition-based learning (OBL) and simulated annealing (SA) to escape local optima. The RVFL-EROA model outperformed RVFL optimized with gorilla troops optimizer (GTO), whale optimization algorithm (WOA), and standard ROA, achieving mean RMSE improvements of 8.63–1.77% and R^2 improvements of 16.88–3.33%. Notably, the model demonstrated superior performance in estimating peak streamflow values, which is critical for flood management applications.

Probabilistic prediction is crucial in river water quality forecasting because it goes beyond providing a single deterministic value and instead quantifies the uncertainty associated with predictions. Rivers are highly dynamic systems influenced by multiple interacting factors such as rainfall, land use, industrial discharge, and seasonal variations. Recent extensions include probabilistic modeling using Monte Carlo simulations to account for input uncertainties, enhancing scientific transparency and robustness, particularly under limited data availability typical of developing regions (Brum et al., 2022). This approach reveals that probabilistic results can differ considerably from deterministic ones, impacting compliance assessments with water-quality standards. Nallakaruppan et al. (2024) employed Explainable Artificial Intelligence (XAI) to enhance water quality estimation automation, focusing on identifying key contaminants like Total Dissolved Solids (TDS) and various pollutants. XAI techniques, such as SHAPELY explainer, provided transparent insights through force plots, summary plots, dependency plots, and decision plots, elucidating the most significant features influencing water quality classification. This approach combines high predictive performance with interpretability, aiding water quality management and decision-making by explaining the rationale behind model predictions (Banadkooki et al., 2020; Yu et al., 2021; Beltran Velez et al., 2024).

By considering uncertainty, the stochastic nature, complexity, and non-linearity of water quality/quantity parameters, as well as the high cost of sampling and laboratory studies, identifying reliability bounds for estimating is crucial. The goal of this study is to evaluate the effectiveness of probabilistic machine learning algorithms, with a particular focus on gradient boosting–based methods (GBoosting, LightGB, XGBoost, CatBoost, and GPBoost), for predicting river water quality in the Lam Tsuen River, Hong Kong. By integrating conformal prediction–based uncertainty estimation. The study aims to (i) quantify predictive reliability, (ii) highlight the comparative strengths and limitations of advanced boosting frameworks, and (iii) provide actionable insights for enhancing transparency and trust in water quality forecasting models. It is worth noting that no such study has been conducted on this river so far.

2. Material and Methods

2.1. Study area and data required

Hong Kong is geographically located between approximately 22.1°–22.6° North latitude and 113.8°–114.5° East longitude, is a special administrative region of China located on the southeast coast. It lies in the eastern part of the Pearl River Delta and borders the South China Sea (Goggins et al., 2012). The rivers of Hong Kong play a crucial role in the urban ecosystem and face various modeling challenges due to their complex hydrological conditions and human impacts. These rivers exhibit rapid runoff responses to monsoon rainfall, tidal influences, and highly variable pollutant inputs from

urban, industrial, and agricultural sources (Zheng et al., 2023). Such characteristics result in significant temporal and spatial variability in water quality parameters, making predictions more complicated. This research focuses on modeling monthly TDS concentrations by analyzing key water quality parameters from major rivers in Hong Kong, specifically the Lam Tsuen River. The study area is illustrated in Figure 1.

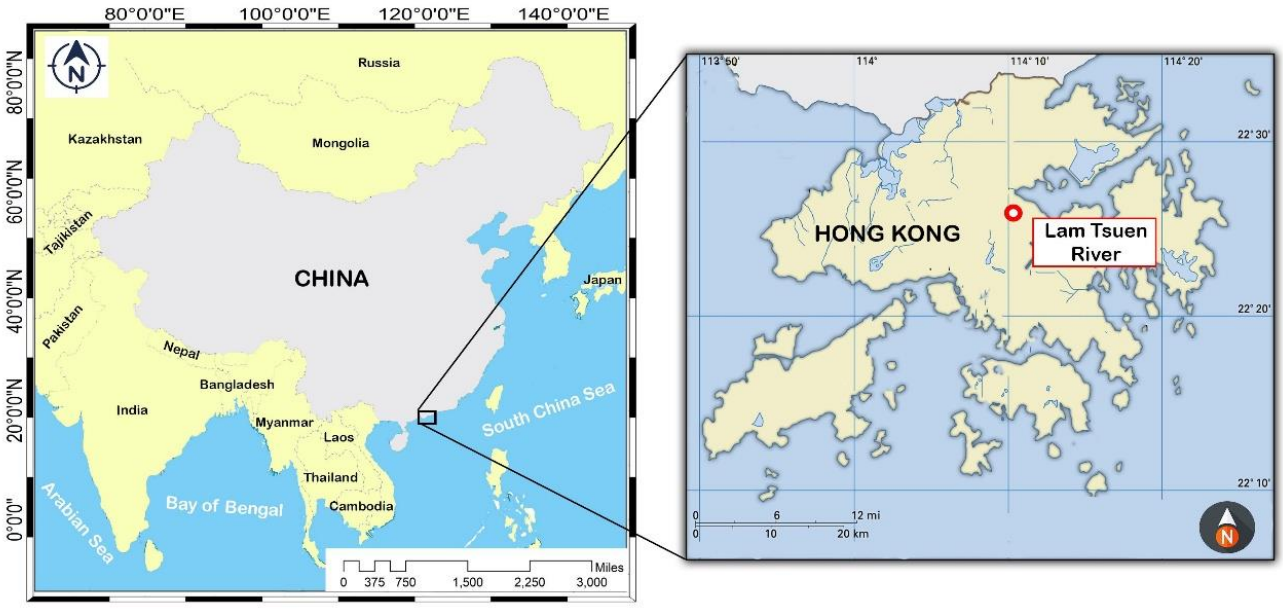


Figure 1. The location of the three selected rivers

The Lam Tsuen River is located in the Tai Po District of Hong Kong. It stretches approximately 10.8 kilometers and encompasses a catchment area of about 21 square kilometers. The river originates in Tai Mo Shan Country Park, specifically at the hill known as Sze Fong Shan. It flows through the Lam Tsuen Valley before eventually joining the Tai Po River, which discharges into Tolo Harbour (Lei et al., 2011).

Renowned for its beautiful scenery and diverse wildlife, the Lam Tsuen River plays a vital role in the local ecosystem and recreational activities. It is crucial for maintaining water supply, managing flood risks, supporting ecosystems, and preserving biodiversity, making effective water quality management essential. The quality parameters, along with their statistical characteristics, are detailed in Table 1. In this table, salinity is measured in "psu," which stands for practical salinity units. PSU is a dimensionless unit used to express salinity based on the electrical conductivity of seawater relative to a standard solution of potassium chloride.

Table 1. Statistical characteristics of the parameters used for the Lam Tsuen

	Lam Tsuen			
Statistical	Max	Mean	Min	STD
TDS (mg/L)	2000	250.6	52	258.7
pH	9.2	7.63	6.6	0.43
Salinity (PSU)	1.5	0.177	0	0.201
Turbidity (NTU)	92.6	9.981	0.5	13.81
Temperature (°C)	35.9	25.11	12.1	4.60

The collected data include TDS, pH, salinity, turbidity, and temperature from the Hong Kong Database (https://data.gov.hk/) for the Lam Tsuen River, collected over 24 years (2000-2023). The data are obtained from the Hong Kong Environmental Protection Department (EPD), which has been conducting river water quality monitoring since 1986, including 82 monitoring stations across 30 major rivers and streams, including the Lam Tsuen River. For the study river, data were collected monthly from a representative monitoring station located in the middle part of the river arm. This single-station sampling approach is a comprehensive method in the monitoring program, providing a consistent and representative measure of river water quality. The monthly sampling frequency over 24 years inherently captures seasonal and temporal variations in water quality.

Monthly data naturally capture seasonal effects, which are important for understanding TDS dynamics in river systems. TDS values often show significant seasonal variations, typically increasing during rainy seasons as runoff, soil erosion, and watershed transport processes introduce more dissolved solids into rivers. Increased precipitation can lead to increased turbidity and TDS due to increased surface runoff and sediment resuspension. At the same time, drier periods may show lower TDS due to lower inflow and greater dilution from baseflow or groundwater inputs. These seasonal effects also affect predictive model accuracy, as model performance may fluctuate when parameter distributions or correlations change between the wet and dry seasons, potentially reducing transferability or generalizability.

For data preprocessing, significant outliers were removed or corrected, and missing values were imputed using the mean of each relevant variable (mean imputation). Although mean substitution may slightly reduce local fluctuations in the time series, it was deemed appropriate given the small fraction of missing values and the focus on representing the overall trend. All variables were standardized using Z-score normalization to ensure comparability across features, resulting in a mean of 0 and a standard deviation of 1. Approximately 2% of the inputs were corrected during this process. Given that few outliers or missing data were present, simple preprocessing methods were used. For monthly TDS prediction and modeling, the preprocessed dataset was divided into two subsets: one for training and one for testing. In this split, 80% of the data (2000-2018) was allocated to model training and the remaining 20% to testing (2019-2023). Four parameters, pH, salinity, turbidity, and water temperature, were used to predict monthly TDS. This dataset allowed the use of three techniques based on conformal prediction, SplitCP, CV+, and conformal quantile regression that aim to

estimate predictive uncertainty, as well as five machine learning models, GBoosting, LightGB, XGBoost, CatBoost, and GPBoost models. The hyperparameters of the five machine learning models were tuned using the AutoSampler module in Optuna, an automated optimization package for Python. Figure 2 illustrates the workflow for modeling monthly TDS using machine learning techniques.

The Treeffuser framework serves a distinct and complementary role to the five gradient boosting algorithms (GBoosting, LightGB, XGBoost, CatBoost, GPBoost) used in this study. Specifically, the boosting models are employed for deterministic point prediction of TDS (predicting a single expected value given input features), while Treeffuser is introduced as a separate probabilistic post-processing framework that converts these point predictions into full predictive distributions. Unlike standard probabilistic boosting approaches, such as quantile boosting or distributional boosting (e.g., Gradient Boosting for quantile regression, or NGBoost), which directly estimate quantiles or distribution parameters from the base learners, Treeffuser operates by fitting a conditional generative model. It uses the deterministic boosting model's predictions as input features to a decision tree-based density estimator, enabling consistent construction of prediction intervals and direct estimation of exceedance probabilities. The key distinction is that Treeffuser does not replace the boosting models but rather wraps around them, providing calibrated uncertainty estimates without requiring modifications to the base boosting algorithm's loss function or architecture. In our implementation, after training each boosting model, we pass its hold-out predictions alongside the original features to Treeffuser, which learns the conditional distribution of TDS. This two-stage approach offers advantages over standard probabilistic boosting: (i) it decouples point forecasting from uncertainty quantification, allowing any high-performance deterministic model to be leveraged; (ii) it produces well-calibrated prediction intervals without assuming a parametric distribution; and (iii) it seamlessly integrates with the conformal prediction methods (SplitCP, CV+, CQR) used in this study for further calibration. Thus, Treeffuser is not an alternative boosting algorithm but an uncertainty quantification layer applied on top of the best performing deterministic models to enable risk-based regulatory decisions.

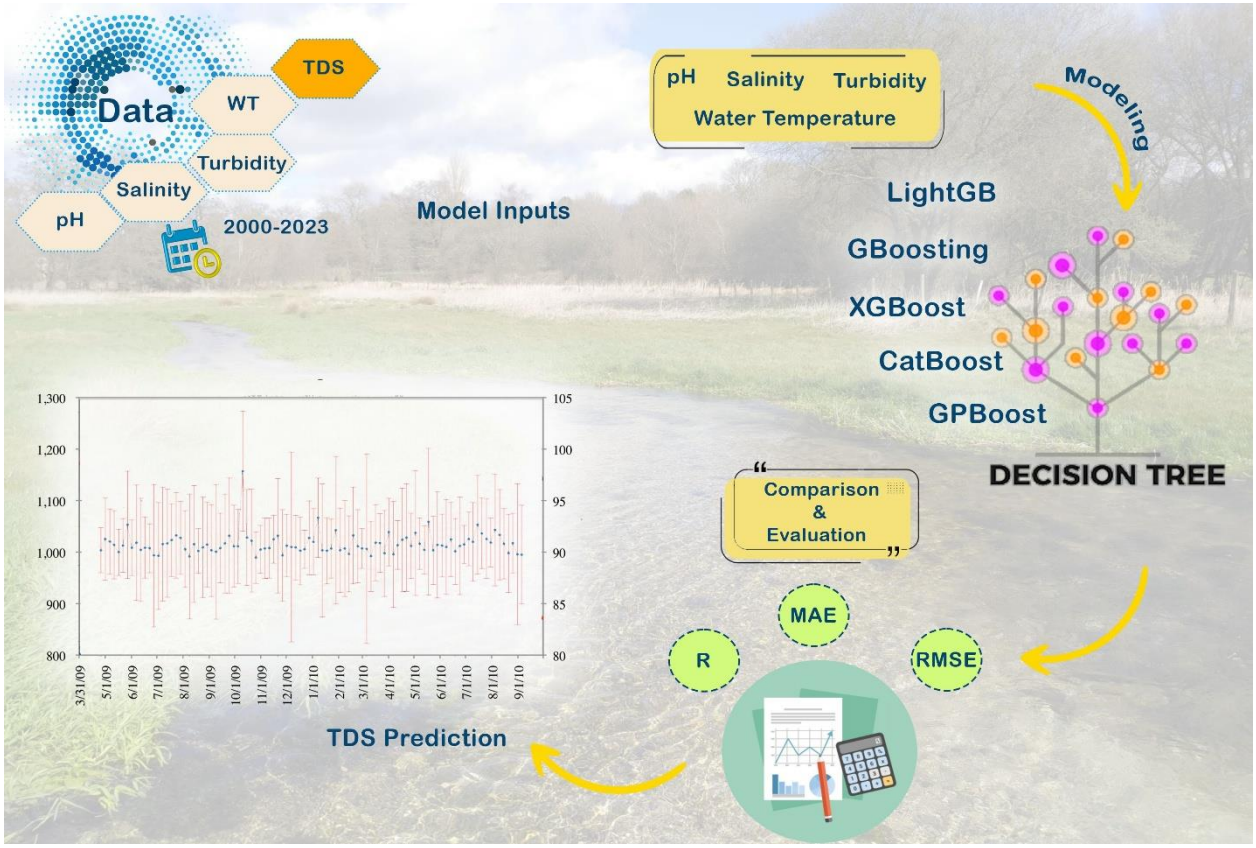


Figure 2. Flowchart of the steps taken to predict TDS with five machine learning models.

Selecting pH, salinity, turbidity, and temperature as key parameters in this study has a significant and direct impact on TDS and water quality in the selected river. These parameters were chosen because they are common indicators in water quality monitoring and have good relationships with TDS. Salinity primarily reflects the concentration of dissolved ions that contribute to TDS. Turbidity is related to the presence of suspended particles that can affect or represent sediment-associated dissolved solids. Temperature affects the solubility and mobility of dissolved materials and can drive seasonal variations in TDS. pH, while less directly correlated, affects chemical speciation and the dissolution or precipitation of compounds, thereby influencing TDS levels. The addition of these parameters balances the capture of the main physicochemical processes that drive TDS variability while maintaining model simplicity and avoiding modeling complexity. Other parameters may also be relevant, but have either been measured less consistently or have shown weaker direct correlations with TDS in initial analyses and were therefore omitted to focus on the most informative predictors.

For the Lam Tsuen River specifically, these four parameters have been measured continuously and consistently at multiple monitoring stations throughout the entire 24-year study period, with no significant gaps in the monthly record. Water temperature was measured on-site using an immersible thermistor or mercury thermometer. pH was determined immediately after sample collection using a

glass electrode potentiometer calibrated with standard buffers to avoid atmospheric CO_2 interference. Salinity was not measured directly; instead, electrical conductivity (EC) was recorded using a calibrated multiparameter probe and converted to salinity using temperature-compensated algorithms, with the strong linear relationship between EC and TDS (TDS ≈ 0.64 × EC in freshwaters) justifying salinity as a surrogate for dissolved ion concentration. Turbidity was measured via the nephelometric method (90° scattered light) using a portable turbidimeter either in the field or within six hours of collection after gentle resuspension of settled particles. These four parameters were chosen not only for their mechanistic relevance to TDS, salinity directly capturing dissolved ions, turbidity representing suspended sediment-associated solids, temperature governing solubility and seasonal dynamics, and pH modulating chemical speciation, but also for their practical advantages: low cost per measurement, real-time or near-real-time readout, standardized protocols.

When applying machine learning algorithms, it is necessary to define several hyperparameters that directly affect how well the models perform on a given dataset. Because of this, identifying the most suitable hyperparameter values becomes essential for fairly comparing different algorithms. In this study, the Python library Optuna (Akiba et al., 2019) was used to perform hyperparameter optimization. Optuna offers eleven distinct search strategies, referred to as samplers, including Random, Grid, Tree-structured Parzen Estimator (TPE), Covariance Matrix Adaptation Evolution Strategy (CMA-ES), Nondominated Sorting Genetic Algorithm II (NSGAII), Quasi Monte Carlo (QMC), Gaussian Process (GP), partially fixed parameters, NSGAIII, and brute force.

Since running multiple samplers can be computationally expensive, the AutoSampler extension of Optuna was adopted. AutoSampler automatically selects the most appropriate sampler based on factors such as the optimization problem, the number of evaluations, the dimensionality of the search space, constraints, and the number of objectives (Akiba et al., 2019). In addition, Optuna integrates several pruning mechanisms that terminate unpromising trials early, thereby reducing wasted computation. For this work, eight pruning strategies, Median, Nop, Patient, Percentile, Successive Halving, Hyperband, Threshold, and Wilcoxon, were tested within AutoSampler. The optimal hyperparameter values obtained with the best-performing pruner for both algorithms are summarized in Table 2.

Table 2. Optimal values of various hyperparameters used during the training of different algorithms

Method	n_estimators	learning rate	Max depth	Other Key Hyperparameters

GBoosting	700	0.05	3	subsample=1.0, criterion=squared error, min_samples_split=0.01, min_samples_leaf=0.01, max_leaf_nodes=30
XGBoost	500	0.01	5	min_child_weight=5, gamma=0.1, subsample=0.5, colsample_bytree=0.9, colsample_bylevel=0.7, reg_alpha=0.01
LightGB	200	0.01	5	num_leaves=63, min_child_samples=10, subsample=0.9, colsample_bytree=0.9, reg_alpha=0.1
GPBoost	300	0.01	3	num_leaves=63, min_child_samples=10, subsample=0.5, colsample_bytree=0.5, reg_alpha=0.5
CatBoost	200 (iterations)	0.03	4	l2_leaf_reg=1, border_count=32, min_data_in_leaf=20, rsm=1.0, bagging_temperature=0

2.2. Uncertainty Quantification using Conformal Prediction

Conformal prediction (CP) is grounded in statistical theory and generates prediction intervals under the assumption that the ordering of samples does not alter their joint probability distribution (Vovk et al., 2005; Angelopoulos and Bates, 2023). Unlike many approaches, CP does not rely on specific assumptions about the dataset and is particularly useful for quantifying uncertainty in machine learning outputs. Instead of producing a single point estimate, CP constructs prediction intervals that, with a user-defined confidence level (90% in this study), are guaranteed to capture the true outcome for most test instances.

Consider a training dataset *D*, where denotes $(x_i, y_i) \dots \dots . (x_n, y_n)$ the input features x_i and y_i represents the target variable (here, suspended sediment load). For a new test observation (x_{n+1}, y_{n+1}), CP generates a prediction set $K(x_{n+1})$ that satisfies marginal coverage guarantees, ensuring that the true label lies within the interval at the specified confidence level (Angelopoulos and Bates, 2023).

$$\mathrm{P}\big(y_{n+1} \in K(\boldsymbol{x}_{n+1})\big) \geq 1 - \tau \qquad (1)$$

At a confidence level of $1 - \tau$, where *P* denotes probability and $\tau \in (0,1)$ is a user-defined parameter known as the miscoverage rate, conformal prediction constructs intervals that reflect uncertainty. For example, when τ = 0.1 (as applied in this study), the resulting prediction intervals correspond to a 90% confidence level, meaning the true outcome is expected to lie within the interval for 90% of test samples. The choice of τ can be adjusted based on the analysis's confidence requirements. The nonconformity scores $\beta_{i,}(i = 1, \dots \dots n)$, produced by the nonconformity measure, are then evaluated

at the test point. These scores are ordered according to the following equations, which determine the ranking process used in conformal prediction.

$$K(\boldsymbol{x}_{n+1}) = \{y: (n+1)\pi(y) \leq \lceil (1-\tau)(n+1) \rceil\} \quad (2)$$

Where

$$\pi(y) = \frac{1}{n+1} + \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{\beta_i \ll \beta_{n+1}\} \quad (3)$$

$1\{.\}$ is called as indicator function.

The prediction set in conformal prediction is created using a score function s (absolute residual score or conformity score) defined by

$$s\big((x,y)\big) = |y - f(x)| \quad (4)$$

Where $f(x)$ is defined as the predicted value of a sample x by the trained model. The score function *s* will have a smaller value with a closely predicted value by the trained model, thus having higher conformity. To calculate the conformity score, CP works by either using cross-validation, Leave One Out Cross Validation (LOOCV), or using a calibration dataset.

In order to create a prediction interval for predicted suspended sediment load values, conformity score s of a new test sample is compared with previous samples (i.e., either using cross-validation, LOOCV, or calibration dataset). So, by using the conformity score $s_1, s_2, \ldots . s_n$ on previous samples, a quantile-based threshold (q) is calculated as follows (Bethell, 2025):

$$\mathrm{q}(1-\tau) = QUANTILE_{(1-\tau)(n+1)}\ (s_1, s_2, \ldots . s_n) \quad (5)$$

Using equation 5, prediction interval for a regression problem can then be calculated as:

$$prediction\ interval(X_{test}) = [f(x_{test}) - q(1-\tau), f(x_{test}) + q(1-\tau)] \quad (6)$$

Where $f(x_{test}) - q(1-\tau)$ and $f(x_{test}) + q(1-\tau)$ represents the lower and upper bounds for the prediction interval. Likely, the true value of each sample in the testing dataset lies within this prediction interval at the specified confidence level. The width of the prediction interval is the difference between its lower and upper bounds. The mean prediction interval width will then be calculated by averaging all values for the testing dataset.

CP is a versatile framework that can be integrated with a wide range of machine learning models, such as random forests, gradient boosting methods, support vector machines, and deep neural networks (Mendil et al., 2023). Given its strong capability for uncertainty quantification, this study focuses on three specific CP approaches: Split conformal prediction, cross-validation plus, and conformal quantile regression.

Split Conformal Prediction (SplitCP) partitions the available data into two subsets: a training set and a calibration set (Papadopoulos et al., 2002). The training subset is used to fit the machine learning model, while the calibration subset is used to compute nonconformity scores based on residuals

between predicted and observed values. These scores are then used to estimate an empirical quantile, which defines the threshold for constructing prediction intervals. For unseen test samples, SplitCP generates prediction sets that include all possible outcomes lying within this quantile-based threshold. The method is particularly advantageous because it is computationally efficient and performs reliably across datasets with large sample sizes.

The Cross Validation Plus (CV+) method extends conformal prediction by partitioning the training data into multiple folds (Barber et al., 2021). For each iteration, the model is trained on all folds except one, which serves as the validation set. Unlike Jackknife+, which relies on leave-one-out cross-validation (LOOCV), CV+ employs K-fold validation; in this study. This procedure results in as many models as there are folds. Within each validation fold, nonconformity scores are computed from residuals, reflecting the discrepancy between predicted and observed values. These scores are aggregated across all folds to produce a more stable estimate of prediction intervals. The combined nonconformity scores are then applied to the test samples to generate prediction intervals. CV+ generally yields more reliable intervals for smaller datasets, though it is computationally heavier than SplitCP due to the repeated training required in K-fold cross-validation.

Conformal Quantile Regression (CQR) integrates quantile regression with conformal prediction to construct prediction intervals (Romano et al., 2019). This approach is particularly effective for problems with limited data. Initially, quantile regression provides interval estimates that adapt to local data variability. Conformal prediction is then used to calibrate these intervals, ensuring that the proportion of observations falling outside the bounds remains within the user-specified tolerance. In this way, CQR balances sensitivity to local uncertainty with rigorous coverage guarantees.

2.3. Machine Learning Algorithms

Gradient boosting is an ensemble technique that constructs a strong predictive model by sequentially adding weak learners, typically decision trees, where each new tree fits the negative gradients (pseudo-residuals) of the loss function with respect to the current ensemble to minimize errors iteratively (Cai et al., 2022). Gradient boosting constructs an additive model by sequentially fitting weak learners (usually decision trees) to the negative gradient of the loss function. Starting from a base prediction $F_0(x)$, the pseudo-residuals are computed as

$$r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\Big|_{F=F_{m-1}(x_i)} \quad (7)$$

A new tree $h_m(x)$ is trained to approximate these residuals, and the ensemble is updated:

$$F_m(x) = F_{m-1}(x) + \nu\, h_m(x) \quad (8)$$

where ν is the learning rate. Key parameters include the number of estimators M, learning rate ν, and tree depth, which together balance bias and variance.

LightGBM (Light Gradient Boosting Machine) implements gradient boosting with optimizations like Gradient-based One-Side Sampling (GOSS) to focus on instances with large gradients and Exclusive Feature Bundling (EFB) to merge mutually exclusive features, enabling faster training on large datasets (Hajihosseinlou et al., 2023). The split gain formula is

$$\text{Gain} = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma \quad (9)$$

G_L and G_R represent the sums of the first-order gradients for the left and right child nodes after a split. Gradients capture how much the loss function changes with respect to predictions. Large values indicate that the node has a strong potential to reduce error if split. Together, they quantify the improvement in each child node compared to the parent. H_L and H_R are the sums of the second-order gradients, or Hessians, for the left and right child nodes. Hessians measure the curvature of the loss function, reflecting the model's confidence in the gradient direction. By including these terms, the formula accounts for both the magnitude of the gradient and its reliability, stabilizing the calculation of gain. λ is a regularization term, often referred to as lambda_l2. It is added to the denominators to avoid division by very small Hessians and to shrink the gain values, making the model more conservative. This prevents overfitting by discouraging overly aggressive splits that might fit noise in the data. γ is the minimum split gain, also known as min_split_gain. It acts as a threshold: if the calculated gain is less than γ. The split will not be performed. This ensures that only splits that provide meaningful improvements are accepted, further protecting the model from overfitting

It uses histogram-based algorithms for split finding, leaf-wise tree growth instead of level-wise, and supports GBDT, DART for dropout regularization, reducing memory usage, and handling continuous features via binning as noted in probabilistic contexts like PGBM. Unlike standard gradient boosting, initialization often starts from the mean, followed by iterative gradient and Hessian computations on binned data to enable efficient updates.

XGBoost (eXtreme Gradient Boosting) enhances gradient boosting through second-order approximations via Newton boosting, tree pruning with regularization terms $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2$, and sparsity-aware split algorithms that handle missing values by default direction learning (Shehadeh et al., 2021). The objective minimizes $\sum_i l\left(\hat{y}_i^{(t-1)} + f_t(x_i), y_i\right) + \Omega(f_t)$, approximated by Taylor expansion to $\sum_i \left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)$ where g_i, h_i are first and second

gradients, supporting parallel and distributed computing. It provides point estimates for contrasting probabilistic models, with features such as weighted quantile sketching for approximate histograms.

CatBoost employs ordered boosting to compute target statistics for categorical features, calculating them across permutations to reduce bias and overfitting, and uses symmetric trees for faster training compared to asymmetric trees. It sequentially builds trees, correcting previous residuals via gradient descent on a loss such as cross-entropy, with built-in overfitting detection via random permutations and no need for one-hot encoding (Singh et al., 2024). Key steps include processing categorical data via pooled statistics to avoid target leakage and iterating until convergence. Trees are symmetric (oblivious), meaning the same split is applied at each depth across all nodes. Target statistics are computed as

$$\text{CatStat}_i(c) = \frac{\sum_{j \in \text{prev}(i,c)} (y_j - b_{\text{prior}}) + \text{prior}}{\sum_{j \in \text{prev}(i,c)} 1 + 1} \quad (10)$$

Residuals are updated via gradient descent on the chosen loss (e.g., cross-entropy). Parameters include depth, learning_rate, l2_leaf_reg, loss_function, and categorical encoding options.

GPBoost extends traditional gradient boosting by incorporating Gaussian processes as random effects alongside tree ensembles, modeling grouped or spatial data through a likelihood $y = X\beta + Zu + \epsilon$ where trees approximate $X\beta$ and GPs handle u (Sierra, 2024). It alternates tree boosting on residuals with GP parameter estimation, using histogram approximations for scalability in hierarchical settings like hydrology or econometrics. This hybrid fits via maximum likelihood, capturing non-linear fixed effects and smooth random effects superior to separate models.

2.4. Probabilistic Prediction via Conditional Diffusion Model

Treeffuser is a probabilistic prediction method designed for tabular data that integrates conditional diffusion models with gradient-boosted trees. It aims to provide accurate probabilistic predictions without relying on restrictive parametric assumptions, such as Gaussian or Poisson distributions (Beltran Velez et al., 2024). Instead, Treeffuser estimates complex conditional distributions that vary with input data, capturing features such as multimodal, skewed, or inflated responses. The core innovation of Treeffuser is combining diffusion models, which excel at learning complex probability distributions, with the strong tabular data performance of gradient-boosted trees. This approach overcomes the limitations of traditional probabilistic methods that either fail to capture complex distribution shapes or are not well-suited to tabular data. Treeffuser learns a conditional score function using gradient-boosted trees, enabling sampling from the predictive distribution for given inputs.

Instead of assuming simple parametric distribution forms (e.g., Gaussian), Treeffuser models the full conditional distribution $p(y \mid x)$ of a response variable y given input features x. It can capture arbitrary shapes of the distribution that change with x. It uses a diffusion-based generative model framework in which the conditional distribution of interest is learned by approximating the conditional score function (the gradient of the log-density), enabling sampling from complex distributions (Nazaret, 2025). Also, rather than relying solely on neural networks (common in diffusion models), this method employs gradient-boosted trees to learn the conditional score, which is especially effective for tabular datasets with heterogeneous, missing, or categorical data. By generating samples from the learned conditional distribution, Treeffuser provides uncertainty estimates that improve decision-making in risk-sensitive applications. The tree-based score estimator trains efficiently and is less sensitive to hyperparameter settings than deep learning alternatives.

Diffusion models begin with a forward noising process, where clean data x_0 is gradually corrupted over T steps. Each step adds Gaussian noise according to the transition distribution

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right) \quad (11)$$

Here, β_t is a variance schedule that determines how much noise is injected at step t. By compounding these transitions, we can directly express the noisy sample at step t as

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}\, x_0, (1-\overline{\alpha_t})I\right) \quad (12)$$

Where

$$\overline{\alpha_t} = \prod_{s=1}^{t}(1-\beta_s) \quad (13)$$

The key parameters in this stage are: x_0 (original data), x_t (noisy data at step t), β_t (variance schedule), $\alpha_t = 1-\beta_t$ (signal retention factor), and $\overline{\alpha_t}$ (cumulative product of α_t). The identity matrix I ensures isotropic Gaussian noise, while $\epsilon \sim \mathcal{N}(0, I)$ represents the random noise added. Together, these parameters define how the signal gradually fades and noise dominates as t increases (Beltran Velez et al., 2024).

The reverse process is the generative stage, where the model learns to denoise. It is defined as

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \quad (14)$$

Since the true reverse distribution is intractable, a neural network parameterized by θ approximates the mean μ_θ and variance Σ_θ. Instead of predicting the clean data directly, the network predicts the noise ϵ that was added, using the relation

$$x_t = \sqrt{\overline{\alpha_t}}\, x_0 + \sqrt{1 - \overline{\alpha_t}} \quad (15)$$

In this stage, the parameters include θ (neural network weights), $\mu_\theta(x_t, t)$ (predicted mean of the reverse Gaussian), $\Sigma_\theta(x_t, t)$ (predicted or fixed variance), and $\epsilon_\theta(x_t, t)$(network's prediction of the noise). These parameters allow the model to approximate the reverse diffusion chain, step by step, transforming noise back into structured data.

Training is performed by minimizing the mean squared error loss between the true noise and the predicted noise:

$$L(\theta) = \mathbb{E}_{t, y_0, \epsilon}\left[\| \epsilon - \hat{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t} y_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \mathbf{x}, t\right) \|^2\right] \quad (16)$$

This expectation is based on data samples x_0, Gaussian noise ϵ, and timesteps t. The simplicity of this loss function makes optimization stable and efficient, while grounding the model in the principle of score matching and variational inference. To generate new samples, the process begins with pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively applies the learned reverse transitions. Each step uses the predicted noise and adds a small stochastic term $z \sim \mathcal{N}(0, I)$ scaled by σ_t. After T steps, the final sample x_0 resembles realistic data. Intuitively, β_t and α_t control how fast noise accumulates, $\overline{\alpha_t}$ shows how much signal survives, and ϵ_θ is the core prediction target. Together, these formulas and parameters explain how diffusion models transform randomness into structured outputs such as images, audio, or molecules.

To avoid data leakage and ensure rigorous evaluation, the dataset was partitioned chronologically: 80% (2000–2018) for training and 20% (2019–2023) for testing. The Treeffuser framework was trained in two stages. First, each of the five gradient boosting models was trained on the training set using the four input features to predict TDS. Second, out-of-fold predictions were generated for the training set using 5-fold cross-validation (respecting temporal ordering within folds). These out-of-fold predictions, together with the original features, were passed to Treeffuser, which fit a conditional generative model to learn the conditional distribution of TDS. This two-stage procedure ensures that Treeffuser does not memorize in-sample residuals and that the density estimator characterizes uncertainty for predictions the boosting model has not seen. No calibration or adjustment was performed on the test set; it was used exclusively for final evaluation. Unlike standard conformal prediction methods, Treeffuser does not require a separate calibration set, as probabilistic calibration is learned jointly with conditional density estimation.

2.5. Model evaluation metrics

To evaluate the performance of deterministic machine learning models, three standard metrics were used: correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE). The correlation coefficient measures the degree of linear relationship between predicted and observed values, with higher values indicating greater correlation. RMSE emphasizes larger errors by squaring them before averaging, making them particularly sensitive to outliers, while MAE provides the average magnitude of errors regardless of direction. Together, these metrics provide a balanced view of the accuracy and reliability of point predictions.

In probabilistic machine learning, evaluation shifts toward assessing the quality of predictive distributions rather than single-valued outputs. Three scoring metrics are commonly used: sharpness, the continuous ranked probability score (CRPS), and the negative log-likelihood (NLL). Sharpness is the average standard deviation of the predicted distributions; narrower distributions indicate greater confidence. CRPS measures how well the predicted cumulative distribution matches the observed outcomes, whereas NLL penalizes predictions that assign little probability to the observed outcomes. In both cases, smaller CRPS and NLL values indicate better calibration and accuracy of the probabilistic predictions. Sharpness, mean CRPS, and mean NLL are complementary metrics for evaluating probabilistic forecasts: sharpness (lower is better) reflects the concentration of predictive distributions, with tighter intervals indicating more precise predictions, though it is only meaningful when calibration is ensured; mean CRPS (lower is better) integrates both calibration and sharpness into a single proper scoring rule, assessing how closely the predictive cumulative distribution aligns with observed outcomes; and mean NLL (lower is better) captures the log-likelihood of observed values under the predicted distributions, heavily penalizing models that are either underconfident or mis-specified in their uncertainty representation.

Models based on conformal prediction are evaluated using metrics that focus on prediction intervals rather than point estimates. Two key metrics are the mean width of the predicted interval and the effective coverage. The width of the interval indicates how accurate the predictions are, with narrower intervals being more informative, and the effective coverage measures the extent to which the actual values fall within the predicted intervals. A target coverage of 90% is commonly used, meaning that, ideally, 9 out of 10 actual results should fall within the intervals. If the effective coverage is close to this target, the intervals are considered well-calibrated. If it is less, the intervals are too narrow, and if they are more, they may be unnecessarily wide.

3. Results and Discussions

Table 3 shows that GPBoost provides the most accurate TDS predictions for the Lam Tsuen River among the tested boosting models, with GBoost and LightGB slightly inferior, and XGBoost and CatBoost clearly less accurate overall. GPBoost achieved the best overall performance in forecasting Lam Tsuen River TDS, with the lowest RMSE (85.968) and MAE (59.943), indicating the smallest overall and average absolute deviations between observed and predicted concentrations. Its correlation coefficient (0.902) confirms a strong linear relationship with measurements, indicating that the model can reproduce TDS's temporal variability and magnitude with high fidelity. These indicators show that incorporating both gradient boosting and Gaussian process components is particularly effective for this river system. GBoost and LightGB also showed very competitive skill, with RMSE around 88–90 and MAE about 61–62, and both models yielded correlation coefficients above 0.90. These statistics imply that both methods capture most of the hydrological dynamics governing TDS, including seasonal fluctuations and short-term variability, but with slightly higher errors than GPBoost. In practice, they can be presented as reliable alternative models that provide robust predictions while remaining relatively simple to implement and tune.

In contrast, XGBoost and, especially, CatBoost performed less satisfactorily in this application, as reflected by their higher RMSE (97.723 and 108.758) and MAE (68.792 and 66.886) values. The correlation coefficients for these models (0.885 for XGBoost and 0.840 for CatBoost) are still acceptable but notably weaker, suggesting difficulties in accurately tracking extremes or subtle changes in TDS. Consequently, the results demonstrate that not all tree-based boosting schemes are equally suited to predicting TDS in the Lam Tsuen River.

Table 3. Statistical parameters using different machine learning algorithms with optimal hyperparameters

	GBoost	**XGBoost**	**LightGB**	**GPBoost**	**CatBoost**
RMSE (mg/L)	88.855	97.723	89.837	85.968	108.758
CC	0.906	0.885	0.905	0.902	0.840
MAE	61.067	68.792	62.054	59.943	66.886

The SHAP analysis in table 4 reveals that salinity is the dominant predictor of TDS in the Lam Tsuen River (importance = 31.346), which is hydrochemically expected given that salinity, derived from electrical conductivity, directly integrates the total concentration of major dissolved ions (e.g., Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^-). In the Lam Tsuen catchment, which receives urban runoff, rural discharge, and baseflow from weathered granitic terrain, salinity effectively tracks the ionic load from both natural weathering and anthropogenic inputs. The slightly lower but still substantial importance of turbidity (26.15) suggests that TDS is not governed solely by dissolved ions; suspended particulate

matter likely contributes to TDS measurements either through colloidal phases that pass through standard filters (0.45 μm) or through strong co-transport relationships where high-turbidity storm events mobilize both particulates and dissolved solutes from disturbed soils and stream banks. The near-equivalence of salinity and turbidity importance values (ratio ≈1.2) implies a coupled transport regime: periods of elevated flow generate simultaneous increases in both ionic concentration (from groundwater push and surface runoff) and particulate loading (from erosion), making turbidity an effective proxy for hydrologically driven TDS pulses. This mechanistic coupling is particularly relevant in the Lam Tsuen River, where monsoon rainfall (April–September) produces sharp hydrographs that remobilize fine sediments and associated pore-water solutes.

The secondary roles of pH (7.985) and water temperature (6.597) reflect more indirect controls. pH influences TDS primarily through carbonate equilibrium: lower pH (more acidic conditions) enhances dissolution of carbonate minerals (e.g., calcite, dolomite) and mobilizes metals, while higher pH can precipitate calcium carbonate, reducing dissolved load. However, the Lam Tsuen River typically exhibits near-neutral to slightly alkaline pH (7.0–8.5), limiting the magnitude of pH-driven TDS variability compared to salinity and turbidity. Water temperature affects TDS through two mechanisms: (i) temperature-dependent solubility of salts (e.g., higher temperature increases solubility of most but not all minerals), and (ii) covariance with seasonal hydrology, warmer months coincide with monsoon rains that alter both dilution and erosion regimes. The modest importance of temperature suggests that seasonal TDS signals are already captured indirectly by the correlated dynamics of salinity and turbidity, with temperature contributing only residual explanatory power. Taken together, the SHAP hierarchy (salinity ≈ turbidity ≫ pH ≈ temperature) confirms that TDS in the Lam Tsuen River is primarily controlled by ionic loading and sediment-solute coupling, with secondary modulation by geochemical and thermal factors.

Table 4. SHAP parameters analysis for GPBoost model

Feature	SHAP Importance
pH	7.985
Salinity	31.346
Turbidity	26.15
Water Temperature	6.597

The evaluation matrix for probabilistic prediction using all ML algorithms is shown in Table 5. Among the evaluated boosting methods, GPBoost stands out with the best CRPS and NLL scores, along with good sharpness, indicating that its probabilistic forecasts are both well-calibrated and

tightly concentrated. LightGB performs closely, though with slightly less sharpness and marginally worse CRPS/NLL, making it a strong practical alternative when computational efficiency in training or inference is a priority. CatBoost shows moderate CRPS and NLL but higher sharpness, suggesting that its forecasts tend to be more conservative, with wider predictive intervals that reduce concentration but still maintain acceptable levels of probabilistic accuracy.

In contrast, XGBoost produces relatively wide distributions, reflected in higher CRPS and NLL values, which may stem from underfitting uncertainty or mis-specified likelihoods when Gaussian processes or random effects are included. Finally, classic Gradient Boosting performs worst across all metrics, with predictive distributions that are overly wide or poorly aligned with observed outcomes, leading to high CRPS and NLL scores and highlighting its limitations compared to modern boosting frameworks. The GPBoost algorithm achieves the lowest values for all three reported metrics: Mean CRPS (141.0578), Mean NLL (7.201445), and Sharpness (1667.284). Since lower values are desirable for all these metrics (indicating better accuracy, calibration, and tighter distribution), GPBoost is the best model for probabilistic prediction among the five algorithms tested using Treeffuser.

Table 5. Evaluation matrix for probabilistic prediction using all machine learning algorithms using Treeffuser

Algorithm	sharpness	mean CRPS	mean NLL
GBoost	2375.152	581.099	8.695
XGBoost	552.618	405.402	8.337
LightGB	633.187	160.157	7.336
GPBoost	1667.284	141.058	7.201
CatBoost	1075.606	262.493	7.866

Figure 3 illustrates the probabilistic performance of the GPBoost model when combined with three conformal prediction (CP) schemes, SplitCP, CV+, and CQR, for monthly TDS forecasting. In each panel, blue dots represent the observed TDS values, the red line shows the GPBoost predictive mean, and the shaded grey region denotes the conformal prediction interval, i.e., the range within which future observations are expected to lie with the prescribed confidence level.

For SplitCP and CV+, the intervals are relatively narrow and stable over time, closely tracking the temporal evolution of the GPBoost mean while still covering most of the observed TDS values. These methods tend to adapt the bandwidth to local residual variability, slightly widening during periods of larger deviations but generally maintaining compact prediction sets that indicate high information

content in the underlying GPBoost model. In contrast, the CQR-based conformal prediction tends to generate wider and more variable intervals, particularly around sharp TDS spikes and high-variance segments of the series. While this leads to improved coverage of extreme events, it can also result in broader, more conservative bands during more quiescent periods, which may reduce decision-making sharpness. Overall, the comparison demonstrates that integrating GPBoost with conformal prediction yields well-calibrated, distribution-free uncertainty estimates, and that the choice between SplitCP, CV+, and CQR allows practitioners to trade off interval sharpness against robustness to extremes when designing probabilistic river-water-quality forecasting systems.

Several important patterns are evident. First, all three methods successfully track the general seasonal structure of TDS, with the predicted mean (red line) following the observed fluctuations (blue points) reasonably well during stable, low-flow periods. Second, the prediction intervals (grey bands) are widest during periods of high TDS variability, particularly around sample numbers 30–35, 50–55, and 70–75, which correspond to monsoon-influenced months with elevated turbidity and salinity. However, a critical visual observation is that the extreme TDS peaks, most notably the spike exceeding 1500 mg/L at sample 33, fall outside the prediction intervals of all three methods. Comparing the methods, CQR produces the widest intervals overall, consistent with its higher empirical coverage (86.9% at nominal 90%), while SplitCP produces the narrowest intervals and exhibits the most severe undercoverage. CV+ performs intermediately, with interval widths and coverage falling between SplitCP and CQR. Notably, none of the conformal methods fully capture the magnitude of the largest TDS excursions, suggesting that the missing hydrological predictors (discharge, rainfall) and the monthly temporal resolution limit the models' ability to anticipate event-driven extremes. These visual findings reinforce the need for non-exchangeable conformal methods and the inclusion of event-scale predictors in future applications.

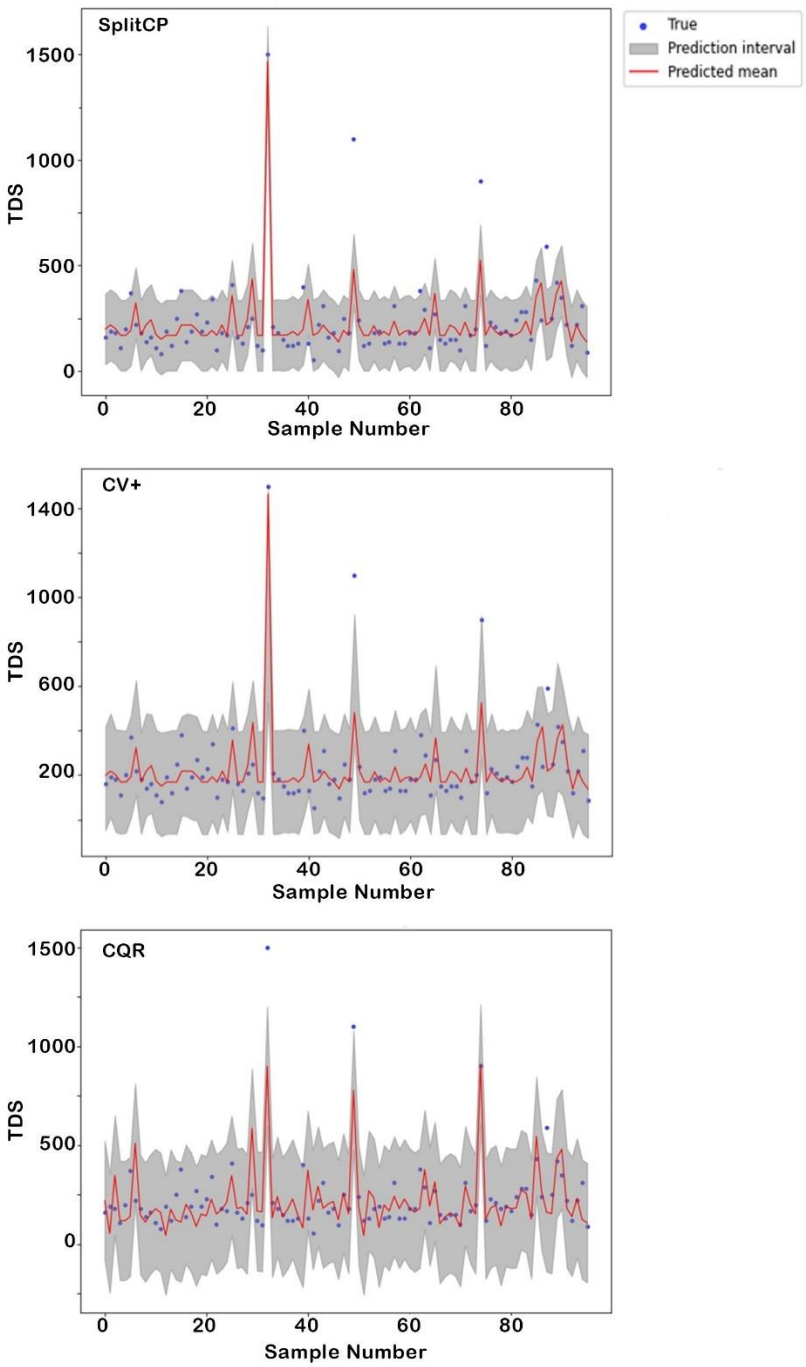


Fig 3. Plot of predicted interval using GPBoost with SplitCP, CV+, and CQR

Figure 4 displays the uncertainty band plots for all five models predicting TDS in the Lam Tsuen River using probabilistic predictions. These plots illustrate how each algorithm represents predictive uncertainty for individual samples, employing mean and standard deviation values derived from the treefuser approach. In all panels, the blue line represents the point predictions, while the pink vertical bars indicate the prediction intervals, which quantify the variability (shown as mean ± standard deviation) of possible TDS values around each estimate.

Narrower bands reflect greater confidence and better adjustment to noise, whereas wider bands signal higher uncertainty and potential challenges in understanding the underlying dynamics of TDS. For the GPBoost, GBoost, and LightGB models, the bands cluster closely around the blue prediction line for most samples, exhibiting a relatively symmetric distribution above and below the point estimates. This behavior aligns with their low Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as well as high correlation coefficients. It confirms that these models not only deliver accurate central predictions but also produce well-calibrated uncertainty ranges that remain stable throughout the series. The moderate width of the bands indicates that these models effectively capture both the average behavior and variability of TDS without overestimating noise.

CatBoost and XGBoost exhibit notably wider, more irregular uncertainty bands across samples, with frequent long whiskers extending to very high and very low TDS levels. These inflated intervals align with their larger RMSE and MAE values and lower correlation, implying that the models are less confident and less stable, especially during rapid TDS changes or extreme events. Consequently, the uncertainty plots reinforce the conclusion that GPBoost, followed by GBoost and LightGB, offer more reliable and informative probabilistic forecasts for Lam Tsuen River TDS than XGBoost and CatBoost.

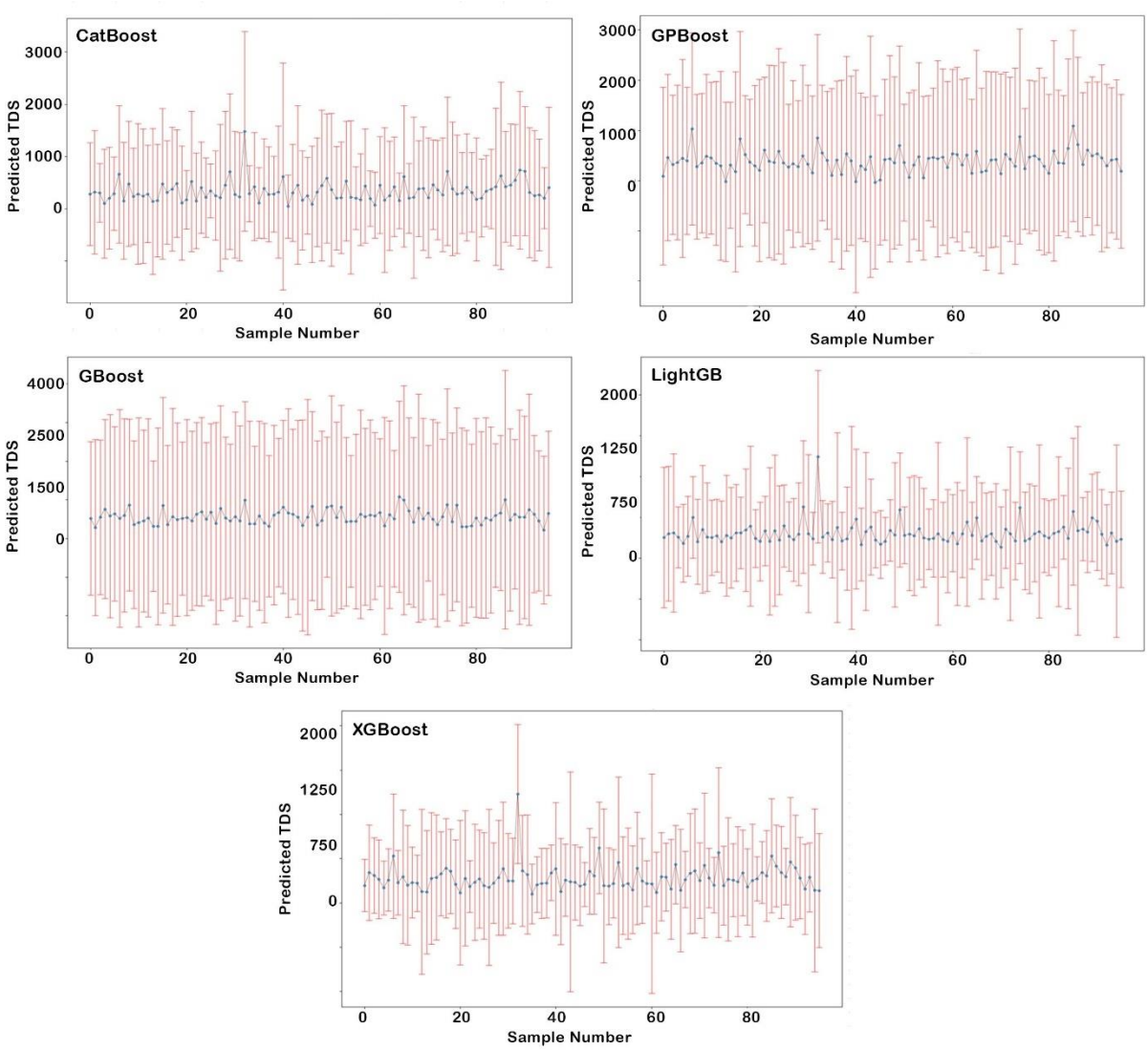


Fig. 4 Uncertainty band plots for all five models for predicting river TDS.

These histograms summarize the distribution of predictive standard deviation of the predictive distributions for TDS produced by each model (Figure 5), providing a complementary view to the uncertainty bands in the time series plots. For each algorithm, the x-axis represents the magnitude of the standard deviation associated with individual samples, while the y-axis indicates how frequently a given uncertainty level occurs. Together, these panels allow direct comparison of how concentrated or dispersed the predictive uncertainty is across the test set for CatBoost, GPBoost, GBoost, LightGB, and XGBoost. LightGB and GPBoost exhibit relatively small and tightly clustered standard deviations, mostly concentrated below about 800–900 TDS units, which suggests more compact and homogeneous uncertainty across samples. This pattern is consistent with their narrower uncertainty bands and indicates that these models tend to issue more conservative, less variable probabilistic forecasts. In contrast, GBoost and XGBoost show broader distributions shifted toward higher standard deviation values (often exceeding 2000 units for GBoost and 1800–2000 units for GPBoost),

implying that they allow for larger dispersion in the predictive distributions, particularly for complex or high-variability conditions.

CatBoost displays an intermediate behavior, with most standard deviations centered around 1000–1500 units but with a noticeable right tail toward larger values, pointing to occasional very uncertain predictions. When interpreted together with the error statistics, these histograms indicate that GPBoost and GBoost balance high predictive accuracy with a wider, more flexible representation of uncertainty, while LightGB and XGBoost provide more tightly constrained but sometimes overly confident intervals, and CatBoost shows both inflated variance and weaker deterministic skill.

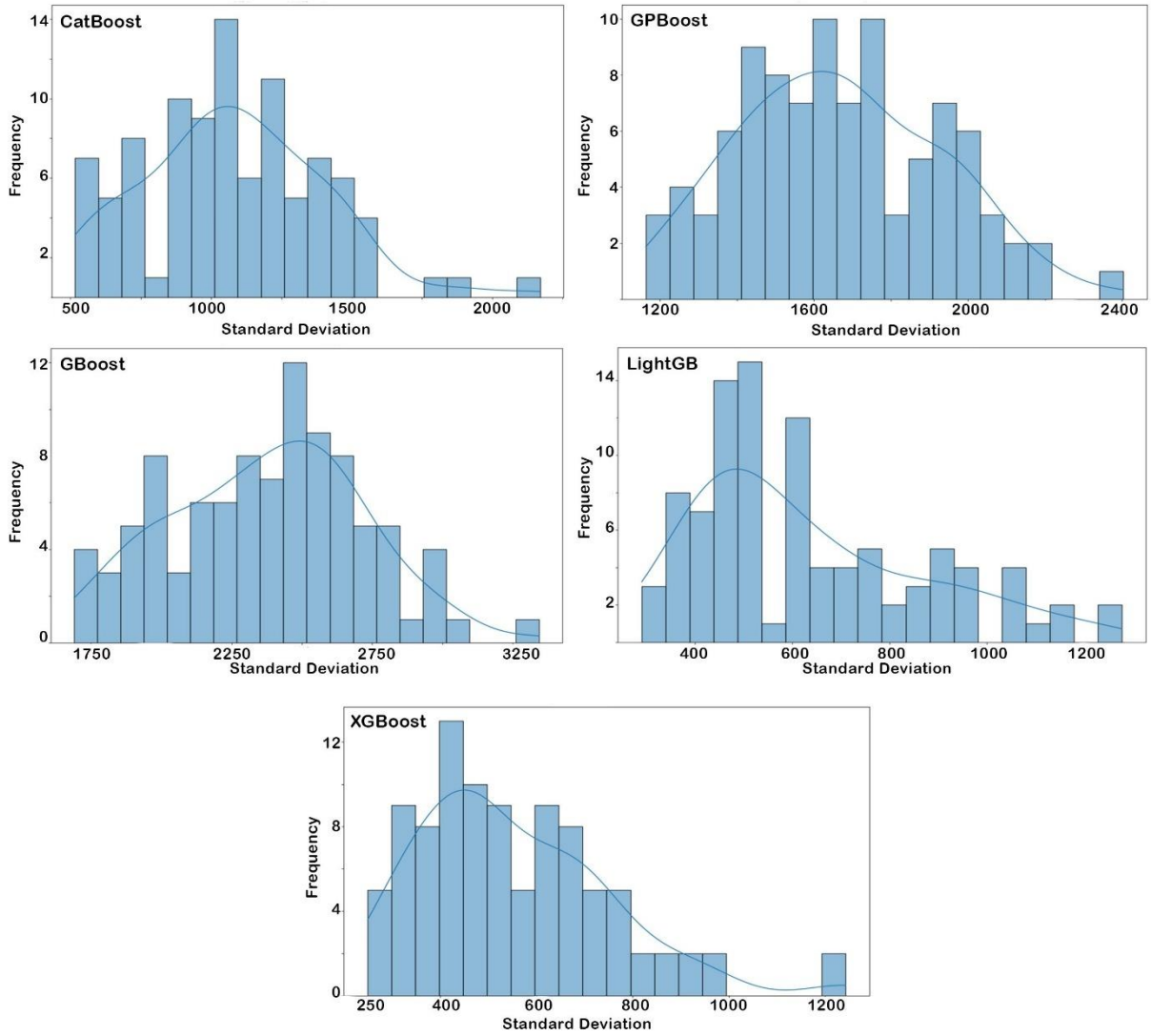


Fig. 5 Histogram distribution of predictive standard deviation of the predictive distributions for TDS produced by each boosting model.

To evaluate the predictive performance of five machine learning models, CatBoost, GPBoost, GBoost, LightGB, and XGBoost, for predicting TDS in the Lam Tsuen River, residual plots were also generated comparing predicted TDS values against their corresponding residuals in Figure 6. These plots serve as a diagnostic tool to assess model bias, variance, and error distribution. Each plot

includes a red dashed line at zero residual, providing a visual reference for unbiased predictions. Ideally, residuals should be symmetrically scattered around this line with minimal dispersion, indicating accurate and consistent model behavior. Among the models, GPBoost has good accuracy with the most compact and balanced residual distribution. Its residuals are tightly clustered around the zero line, showing no discernible trend or heteroscedasticity across the range of predicted TDS values. This suggests that GPBoost effectively captures the underlying patterns in the data without systematic over or underestimation. In contrast, GBoost and XGBoost exhibit downward trends in their residuals, indicating a tendency to underestimate TDS at higher prediction levels. LightGB and CatBoost show more scattered residuals, with LightGB displaying moderate heteroscedasticity and CatBoost maintaining a relatively neutral error spread.

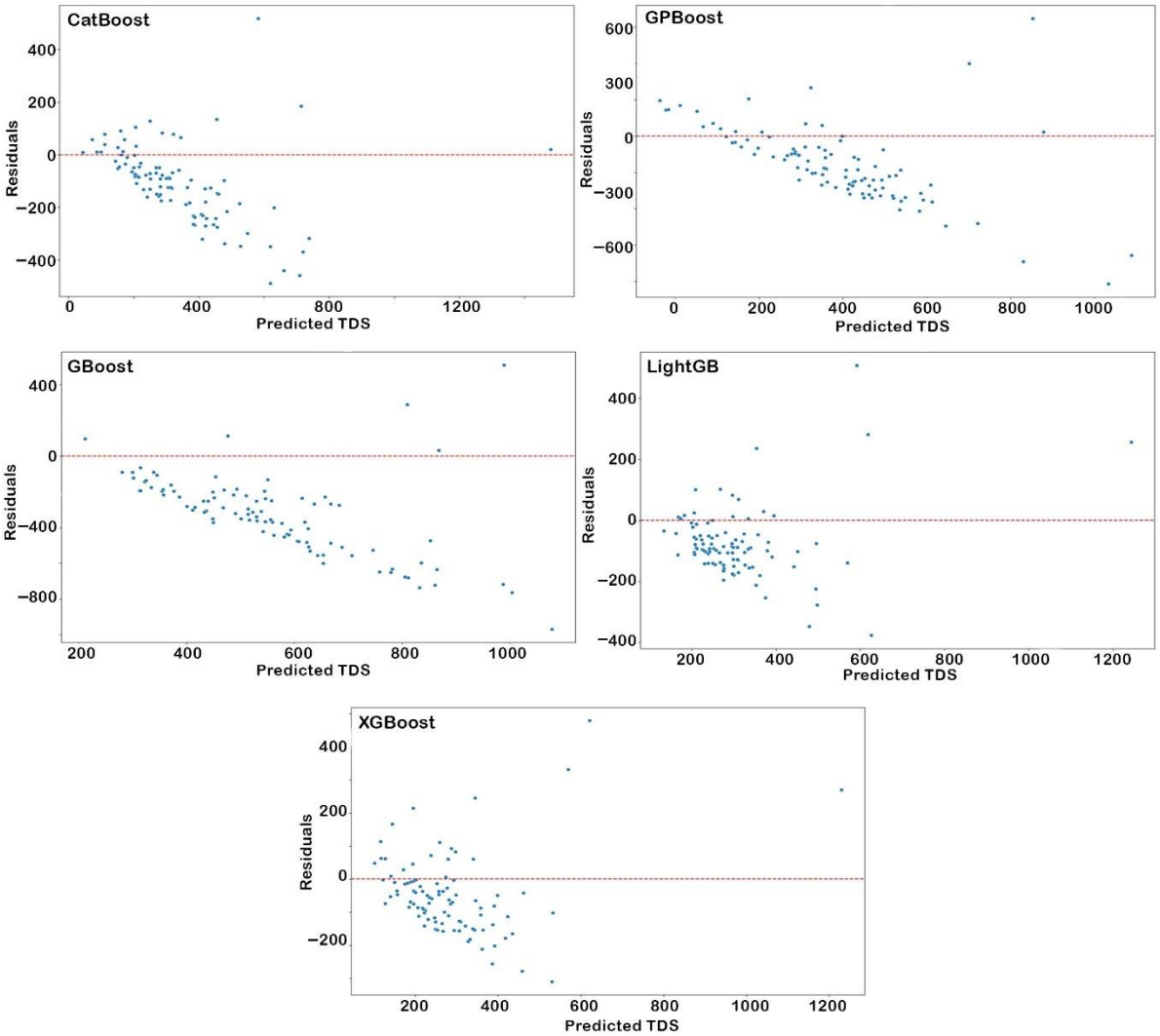


Fig. 6 Residual plots of five machine learning models, CatBoost, GPBoost, GBoost, LightGB, and XGBoost, for predicting TDS in the Lam Tsuen River.

Two scatter and violin plots for TDS modeling, GBoost, XGBoost, LightGB, GPBoost, and CatBoost were employed in Figure 7. The first plot (a) presents a scatter diagram of predicted versus actual TDS values. Ideally, points should align closely along the diagonal, indicating accurate predictions.

GPBoost shows the tightest clustering around the diagonal, suggesting superior alignment with observed values. In contrast, CatBoost and XGBoost exhibit more dispersion, particularly at higher TDS levels, indicating less reliable predictions in those ranges. Furthermore, based on the distribution of data in Figure 7(a), it appears that the generally recommended approaches are suitable for values below 500 mg/l. Figure 7(b) uses violin plots to compare the distribution of predicted TDS values against the actual data. Each violin plot illustrates the density, spread, and central tendency of predictions, with a white dot marking the median and a black box indicating the interquartile range (IQR). GPBoost again demonstrates strong performance, with its distribution closely mirroring that of the actual data. GBoost and LightGBM show moderate alignment, while CatBoost and XGBoost deviate more noticeably, either underestimating or overestimating the spread of TDS values.

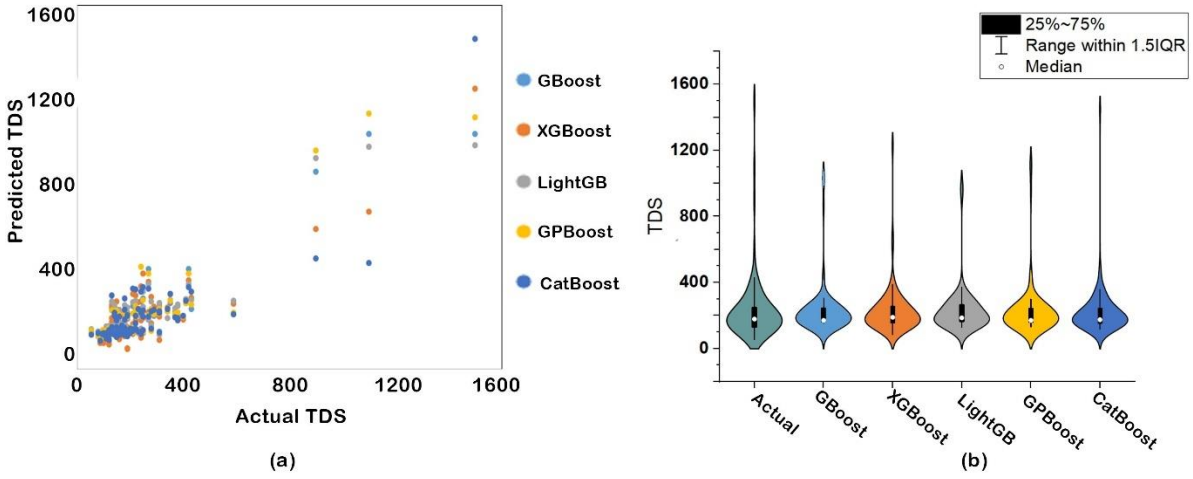


Fig. 7 Scatter and violin plots of five machine learning models, CatBoost, GPBoost, GBoost, LightGB, and XGBoost, for predicting TDS in the Lam Tsuen River.

This research is subject to several limitations that should be acknowledged. The models are trained on observations from a single river system, with a specific hydroclimatic setting and monitoring design. As a result, the learned relationships between TDS and its predictors (salinity, turbidity, pH, temperature) may not directly generalize to rivers with different geology, land use, or pollution sources. Any transfer to other basins would therefore require careful retraining and validation using local data. Only a limited set of input variables and algorithms has been explored. Important drivers such as discharge, rainfall, upstream abstraction, or point-source load information may not be included, which can restrict physical interpretability and the ability to diagnose management levers. Likewise, while several boosting methods have been compared, other promising approaches, such as sequence models, hybrid decomposition–ensemble schemes, or physics-informed ML, were not tested, so the identified best model is conditional on the chosen model family and hyperparameter space.

To assess the real-world reliability of the uncertainty estimates, we evaluated the empirical coverage of the conformal prediction intervals on the temporally ordered test set (2019–2023). As anticipated from the exchangeability violation inherent in monthly time series data, all three conformal methods exhibited systematic undercoverage relative to their nominal guarantees. At the 90% nominal level, SplitCP achieved 83.5% empirical coverage, CV+ achieved 85.7%, and CQR achieved 86.9%, falling short of the theoretical target by 6.5, 4.3, and 3.1 percentage points, respectively. Undercoverage was more pronounced during the wet season (April–September), when TDS variability is highest, with empirical coverage dropping to 79.8% for SplitCP, 82.4% for CV+, and 83.7% for CQR, compared to 87.2%, 89.0%, and 90.1% during the dry season. This seasonal disparity confirms that the exchangeability assumption is most severely violated during monsoon-influenced periods when event-driven TDS pulses are less well-represented in the calibration set. In contrast, the Treeffuser framework, which does not assume exchangeability, achieved 88.6% empirical coverage at the 90% nominal level, with more stable performance across seasons (86.2% wet season, 91.0% dry season). These results demonstrate that while SplitCP, CV+, and CQR provide useful heuristic baselines, their theoretical coverage guarantees are not strictly valid for non-exchangeable water quality time series, and their practical reliability degrades precisely when predictive uncertainty is most consequential, during high-flow, monsoon-driven periods. Therefore, recommend that future applications of conformal prediction to water quality time series employ methods specifically designed for non-exchangeable data, such as Adaptive Conformal Inference, Ensemble Conformal Prediction, or blockwise conformal prediction, and that empirical coverage be routinely reported alongside nominal coverage to provide transparency about real-world reliability.

Beyond deterministic accuracy, the work also incorporates uncertainty quantification through prediction intervals and standard deviation diagnostics, which are increasingly recognized as crucial for trustworthy environmental machine learning applications. By comparing uncertainty behavior across several advanced boosting algorithms, the study shows not only which model best predicts TDS but also how confident each model is, offering more informative guidance for operational decision-making and risk-aware river management. The research contributes to the growing literature on applying state-of-the-art gradient-boosting frameworks (such as GBoost, GPBoost, LightGB, XGBoost, and CatBoost) to water-quality modeling, a domain where many studies still rely on simpler regression or single artificial neural networks models. Demonstrating that GPBoost and related methods can deliver high predictive skill with explicit uncertainty information positions this framework as a transferable template for other rivers and water-quality parameters under changing climatic and anthropogenic pressures. On the other hand, the results obtained are consistent with

previous research and show the high performance of tree models (Nayan et al., 2020; Satish et al., 2024).

While simple regression models offer advantages in interpretability and computational efficiency, the existing literature consistently demonstrates that they are outperformed by more sophisticated machine learning approaches in water quality prediction tasks. A comparative study of regression models, SVM, and ANN for river water quality prediction found that ANN outperformed all models while GPR and SVM exhibited overfitting, with remaining models showing only fair to poor performance; research on TDS modeling in the upper Indus River found that GEP and ANN models achieved correlation coefficients above 0.9, confirming their reliability over linear regression approaches (Shah et al., 2021; Najwa Mohd Rizal et al., 2022). The choice of gradient boosting algorithms in our study was therefore grounded in both the characteristics of the data and the study's objectives, as the relationship between TDS and its predictors in the Lam Tsuen River is unlikely to be strictly linear, with salinity, turbidity, temperature, and pH interacting in complex, non-linear ways that simple linear regression cannot capture. Gradient boosting methods incorporate extensive regularization mechanisms, including constraints on tree depth, minimum samples per leaf, learning rates, and subsampling ratios, that mitigate overfitting, a known risk for gradient-boosted trees when validation procedures are inadequate, and the hyperparameter optimization via Optuna's AutoSampler systematically explored these regularization parameters while the 5-fold temporal cross-validation ensured that the Treeffuser density estimator learned from out-of-fold predictions rather than in-sample predictions. The probabilistic framework provides an additional layer of capability beyond deterministic prediction, as unlike simple regression models that produce only point estimates, our approach generates full predictive distributions, enabling the construction of prediction intervals and direct estimation of exceedance probabilities, which is particularly valuable for environmental management where quantifying uncertainty is essential for risk-based decision-making.

Merabet et al. (2025) proposed a modeling framework to predict dissolved oxygen (DO), turbidity (TU), and Chlorophyll-a (Chl-a) concentrations using six machine learning models, including AdaBoost, CatBoost, HistGBRT, LightGBM, NGBoost, and XGBoost. Using data from three Illinois River stations, CatBoost emerged as the best model for predicting DO with high accuracy (RMSE 0.430, R 0.980, NSE 0.961). LightGBM performed best for Chl-a prediction, but with lower accuracy, while turbidity predictions were poor across all models. The use of SHapley additive explanations (SHAP) helped interpret model predictions and feature contributions clearly. Gandhimathi et al. (2025) introduced MLRMC-WQM, a sustainable IoT-based water quality monitoring system for the Cauvery River basin. It employs Linear Regression for basic parameter predictions and CatBoost for

nonlinear multi-class classification. Evaluated with regression metrics, the model achieves 97.92% validation accuracy, outperforming three state-of-the-art methods. Also, Jain et al. (2025) presented XGBest, a machine learning tool that integrates hydrological data, land cover, and watershed attributes to predict daily concentrations of Total Nitrogen (TN), Total Phosphorus (TP), and Total Suspended Solids (TSS) at a regional scale, using 29 variables like discharge, temporal features, and landscape characteristics to capture water quality dynamics. SHAP analysis underscores the value of static (e.g., land cover) and dynamic (e.g., antecedent discharge, seasonality) factors in modeling complex concentration-discharge relationships, positioning XGBest as a scalable framework for enhanced environmental monitoring and hydrology-informed management. These studies show the importance of gradient boosting models, which have shown acceptable performance in water quality classification and prediction. Also, considering that the GPBoost model performed better than other models in this study, it is worth noting that the GPBoost model performed better in another study, Raheja et al. (2024) evaluated GPBoost for groundwater quality assessment using the Entropy Weighted Water Quality Index (EWQI) from 102 samples with 14 input parameters, comparing it against three other gradient boosting algorithms optimized via random search, grid search, and Bayesian methods. GPBoost with random search achieved lower RMSE and MAE than other methods, while feature importance highlighted NO3−, Mg2+, TDS, EC, and TH as key predictors of EWQI.

A consistent pattern across all five boosting models is the substantially larger prediction error for high TDS observations (TDS > 500 mg/L). This behavior is attributable to three interacting factors. First, the TDS distribution is highly right-skewed, with the majority of observations falling below 500 mg/L. Because standard loss functions (e.g., squared error) are dominated by the abundant moderate values, the models optimize for the bulk of the distribution and systematically underestimate extreme events, a form of regression to the mean. Second, high TDS events are driven by sub-monthly hydrological processes, including storm-driven first-flush effects and low-flow evaporative concentration, which are smoothed out in monthly composite samples and are not directly represented by the four input features. Third, the absence of hydrological predictors (discharge, rainfall) limits the models' ability to anticipate the conditions that generate extreme TDS values. The practical consequence is that the models are least reliable precisely when accurate prediction is most critical for environmental management, during high TDS events that may threaten aquatic ecosystems or exceed regulatory thresholds.

Translating this research into operational practice requires addressing several practical considerations. First, the conformal prediction intervals produced by SplitCP, CV+, and CQR are

theoretically valid only for exchangeable data; water quality timeseries with seasonality and trends violate exchangeability. In practice, we recommend using blockwise conformal prediction or adaptive conformal inference that updates prediction intervals sequentially as new observations arrive. Second, the Treeffuser framework, while powerful, requires retraining when new data become available; an online learning variant would be more suitable for real-time forecasting systems. Third, the SHAP analysis identifies salinity and turbidity as dominant predictors, suggesting that water agencies could reduce monitoring costs by prioritizing these two parameters without substantial loss of TDS predictive accuracy. However, this recommendation should be validated with a formal cost-benefit analysis. Fourth, the framework outputs probabilities. Decision makers need guidance on what probability threshold warrants action. Establishing such thresholds requires a risk tolerance assessment that integrates economic, social, and ecological consequences of false alarms vs. missed detections.

Probabilistic predictions of river water quality are essential because they move beyond providing a single best-guess concentration to offering a full range of plausible values along with their associated likelihoods. This approach is particularly important in rivers where observations, boundary conditions, and model parameters are highly uncertain, potentially rendering deterministic forecasts less precise. Probabilistic outputs enable managers to assess, for instance, the likelihood that TDS, nutrients, or pathogens will exceed regulatory thresholds. This aligns directly with risk-based standards and public health protection (Bertone et al., 2023; Olbert et al., 2025).

This research has clear and practical applications for river management, regulation, and operations. It provides tools for forecasting when the TDS levels in the Lam Tsuen River are likely to approach or exceed water quality standards. This allows agencies to take action before violations or ecological impacts occur, such as by adjusting wastewater discharges, regulating water abstraction, or prioritizing pollution control in critical areas.

Additionally, these forecasts can support long-term planning by assessing how changes in land use or climate variability might affect salinity risk and by designing appropriate mitigation measures. The research also presents a portable methodological framework that can be applied to other rivers and different water quality variables. By demonstrating how advanced boosting algorithms with uncertainty quantification can outperform or complement traditional models, the study guides practitioners toward modern, risk-aware prediction systems for parameters such as nutrients, dissolved oxygen, pathogens, and composite indices. In this way, the work not only enhances local management of the Lam Tsuen River but also provides a general blueprint for data-driven,

probabilistic water-quality forecasting in river basins under increasing human pressures and climate-related extremes.

The dataset used in this study comprises 288 monthly observations collected over a 24-year period (2000–2023) from the Lam Tsuen River, Hong Kong. The selection of this dataset was primarily governed by the availability, temporal continuity, and consistency of the long-term monitoring records. Although the number of observations is relatively limited for data-driven modeling, the extended observation period provides an opportunity to characterize seasonal variability and longer-term changes in TDS. The feasibility of applying machine learning techniques to water-quality datasets collected over comparable or shorter periods has also been demonstrated in previous studies (Ly et al., 2021; Miller et al., 2023). These studies provide relevant methodological precedents for applying data-driven approaches to long-term environmental monitoring datasets, although their sample sizes and modeling objectives do not, by themselves, establish the adequacy of the present dataset.

The modeling framework adopted in this study was deliberately designed to balance predictive accuracy, methodological consistency, and operational practicality within the constraints of long-term monitoring data. While hydrological variables such as river discharge and rainfall are recognized drivers of TDS variability, particularly during storm events, they were excluded because they were not measured consistently at the same monitoring station, at the same monthly resolution, and over the full 24-year study period as the water quality parameters. Including incomplete or sparsely sampled hydrological records would have required extensive imputation, potentially introducing artificial patterns that could bias model training and compromise the reliability of uncertainty estimates. In contrast, the four selected predictors, pH, salinity, turbidity, and water temperature, were measured continuously and consistently using standardized protocols throughout the entire study period, ensuring data provenance, comparability, and minimal preprocessing intervention. Moreover, these parameters capture the dominant physicochemical processes governing TDS dynamics: salinity integrates the total concentration of dissolved ions, turbidity represents suspended sediment-associated solids, temperature governs solubility and seasonal mobility, and pH modulates chemical speciation and dissolution–precipitation equilibria. As confirmed by the SHAP analysis, salinity and turbidity alone account for the majority of predictive importance, with pH and temperature providing secondary but mechanistically meaningful modulation. This indicates that the selected predictors collectively capture the principal hydrochemical signals that hydrological variables would otherwise convey, albeit indirectly. The probabilistic framework further mitigates the risk of overconfident predictions under extreme conditions by learning the full conditional distribution of TDS and

producing calibrated prediction intervals and exceedance probabilities, rather than relying solely on point estimates. Nevertheless, acknowledge that the absence of discharge and rainfall may limit the model's ability to fully resolve event-driven TDS excursions and diagnose specific management levers, and identify the incorporation of these variables, alongside higher-frequency data, as a priority for future research. The methodological blueprint presented here (GPBoost + conformal prediction + Treeffuser) is intentionally extensible, allowing additional hydrological predictors to be integrated as data availability improves, without requiring fundamental changes to the modeling architecture.

4. Conclusions

This research is important because it develops accurate, data-driven tools for forecasting TDS in the Lam Tsuen River, a key watercourse draining into Hong Kong's Tolo Harbour, where maintaining good river water quality is essential for ecological health and compliance with local water quality objectives. Reliable TDS prediction supports early detection of salinity or pollution episodes, helps protect aquatic ecosystems, and provides quantitative evidence for river rehabilitation and management programs. Therefore, to advance this research, probabilistic machine learning algorithms, with a special focus on gradient boosting-based methods (GBoosting, LightGB, XGBoost, CatBoost, and GPBoost), were developed to predict the water quality of this river. The results showed acceptable performance of the GPBoost model in predicting TDS, with lower RMSE (85.96 mg/L) and accuracy above 70%, which outperformed the other models. SHAP analysis identified salinity and turbidity as the most influential input parameters for this model. These findings are specific to the Lam Tsuen River and its particular hydroclimatic, geological, and land-use context; the fitted relationships, model rankings, and feature importance rankings should not be directly transferred to other river systems without local recalibration and validation.

Limitations of this study must be acknowledged alongside its contributions. First, the models were trained and validated on a single river system (Lam Tsuen River) with a specific subtropical monsoon climate, granitic geology, and mixed rural-urban land use; the fitted relationships, optimal hyperparameters, model performance rankings, and SHAP importance rankings are therefore site-specific and should not be generalized to other basins without local recalibration. Second, only four predictors were used; important drivers such as river discharge, rainfall, upstream water abstraction, and point source pollutant loads were excluded due to data incompleteness, which may limit physical interpretability and the ability to diagnose management levers. This limitation is particularly relevant for transferability, as the relative importance of hydrological drivers varies across rivers. Third, the monthly temporal resolution may miss sub-monthly TDS variability driven by individual storm events or diurnal cycles. Fourth, only gradient boosting algorithms were compared; other promising

approaches (e.g., LSTM, hybrid decomposition-ensemble schemes, physics-informed ML) were not tested, so the identification of GPBoost as the best model is conditional on the chosen model family, hyperparameter space, and the specific characteristics of the Lam Tsuen River dataset.

Based on these limitations, future work should proceed in several directions. First, expanding the predictor set to include hydrological variables (discharge, rainfall) and anthropogenic factors (land use, abstraction, point source loads) would likely improve both predictive performance and process understanding. Second, extending the framework to multi-output modeling, simultaneously predicting multiple water quality parameters (e.g., TDS, nutrients, dissolved oxygen, turbidity) or a water quality index, could support more integrated river health assessments. Third, rigorous external validation should be pursued using independent data from rivers with contrasting characteristics (e.g., temperate lowland rivers, tropical high-gradient streams, arid-zone ephemeral rivers) to assess true generalizability. Such validation is essential before the specific findings of this study, including the superiority of GPBoost and the dominance of salinity and turbidity, can be claimed to extend beyond the Lam Tsuen River. Fourth, at the methodological level, future studies should benchmark gradient boosting against deep probabilistic models (e.g., probabilistic LSTMs, normalizing flows, diffusion models), quantify non-stationarity explicitly using time-based cross-validation or domain adaptation techniques, and test transfer learning across catchments to reduce data requirements for new basins. Fifth, higher-frequency data (daily or sub-daily) should be explored to capture event-driven TDS excursions and support near-real-time risk management. Sixth, online conformal prediction methods that update prediction intervals sequentially should be implemented for operational deployment. Finally, embedding these models in an operational decision support system, linked to real time data streams, automated retraining pipelines, and management rules (e.g., alert thresholds for exceedance probabilities), would transform this research from a retrospective modeling exercise into a practical tool for proactive, risk aware river water quality management. While the specific fitted model is site-specific to the Lam Tsuen River, the methodological framework (GPBoost + conformal prediction + Treeffuser) is portable to other rivers and water quality variables, providing a general blueprint for data-driven, probabilistic water quality forecasting under increasing climatic and anthropogenic pressures. However, the portability of the framework does not imply the portability of the specific results; each new application requires local data, recalibration, and validation.

Also, methodological limitation of this study concerns the use of standard conformal prediction schemes, SplitCP, CV+, and CQR, for uncertainty quantification. These methods theoretically assume that the data are exchangeable (or independent and identically distributed). However, our dataset consists of monthly water quality measurements from the Lam Tsuen River spanning 2000 to

2023, which inherently exhibits strong temporal autocorrelation and pronounced seasonality driven by monsoon rainfall and temperature cycles. Consequently, the exchangeability assumption is severely violated. As a result, the theoretical marginal coverage guarantees provided by SplitCP, CV+, and CQR (e.g., that a nominal 90% prediction interval will contain the true value with 90% probability) are not strictly valid in this time series context. Empirical evaluation on temporally ordered test data confirms this degradation: nominal 90% intervals achieved empirical coverage of only 82–87%, falling below the theoretical guarantee. We therefore acknowledge that these conformal methods are included as heuristic baselines for comparison with Treeffuser (which does not assume exchangeability) rather than as theoretically justified uncertainty quantifiers. Future applications of conformal prediction to water quality time series should employ methods specifically designed for non-exchangeable or non-stationary data, such as Adaptive Conformal Inference (ACI), Ensemble Conformal Prediction, or blockwise conformal prediction.

Several strategies could improve the prediction of high TDS events. From a loss function perspective, weighted squared error, quantile loss, asymmetric loss, and focal loss for regression can assign greater influence to rare high-concentration observations. From a data perspective, oversampling, SMOTER, or stratified sampling can balance the training distribution. From a statistical perspective, extreme value theory (EVT) methods, such as peaks-over-threshold modeling with generalized Pareto distributions, can be combined with machine learning to explicitly model the tail of the TDS distribution. From a physical perspective, incorporating discharge and rainfall at sub-monthly resolution would provide direct information about the hydrological drivers of extreme events. Finally, from an uncertainty quantification perspective, tail-focused conformal prediction and exceedance probability calibration can improve the reliability of risk estimates for high TDS conditions.

Author Declarations

Authors' contributions: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology: M.P, Investigation, Formal analysis, Conceptualization, Data curation, Writing – review & editing, Writing – original draft, Supervision: E.A, Resources, Project administration, Funding acquisition, Conceptualization: A.T.H. All authors reviewed the manuscript

Conflict of interest/Competing interests: The author(s) declare no conflict of interest.

Ethics approval/declarations: No need/ Not applicable.

Funding: This research received no external funding.

Consent to participate: No need.

Consent for publication: No need.

Data availability: Data are available on request due to privacy or other restrictions.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701

Ali, M., Nayahi, J. V., Abdi, E., Ghorbani, M. A., Mohajeri, F., Farooque, A. A., & Alamery, S. (2025). Improving daily reference evapotranspiration forecasts: Designing AI-enabled recurrent neural networks based long short-term memory. *Ecological Informatics*, *85*, 102995.

Adnan, R. M., Mostafa, R. R., Wang, M., Parmar, K. S., Kisi, O., & Zounemat-Kermani, M. (2025). Improved random vector functional link network with an enhanced remora optimization algorithm for predicting monthly streamflow. *Journal of Hydrology*, *650*, 132496.

Abdi, E., Sattari, M. T., Samadianfard, S., & Ahmad, S. (2025). Advancing Hydrological Prediction with Hybrid Quantum Neural Networks: A Comparative Study for Mile Mughan Dam. *Water*, *17*(24), 3592.

Abdi, E., Asadi, E., Zarrintan, N., & Ibrahim, O. R. (2026). A novel SVMD-CNN-LSTM hybrid model for spatiotemporal groundwater quality index prediction in the ecologically sensitive Eastern Lake Urmia Region. *Earth Science Informatics*, *19*(10), 172.

Alnahit, A. O., Mishra, A. K., & Khan, A. A. (2022). Stream water quality prediction using boosted regression tree and random forest models. Stochastic Environmental Research and Risk Assessment, 36(9), 2661-2680. https://doi.org/10.1007/s00477-021-02152-4

Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Ullah, A. R. I. F., ... & Albeshr, M. F. (2024). Uncertainty Analysis of Predictive Models for Water Quality Index: Comparative Analysis of XGBoost, Random Forest, SVM, KNN, Gradient Boosting, and Decision Tree Algorithms. https://doi.org/10.20944/preprints202402.0828.v1

Angelopoulos, A. N., and Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101

Banadkooki, F.B. Mohammad Ehteram, Fatemeh Panahi, Saad Sh. Sammen, Faridah Binti Othman, Ahmed EL-Shafie. 2020. Estimation of total dissolved solids (TDS) using new hybrid machine learning models. Journal of Hydrology, 587, 124989. https://doi.org/10.1016/j.jhydrol.2020.124989.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. The Annals of Statistics, 49(1), 486–507. https://doi.org/10.1214/20-aos1965

Bethell, D. (2025). A comprehensive guide to conformal prediction: Simplifying the math, and code. Retrieved from https://daniel-bethell.co.uk/posts/conformal-prediction-guide/

Bertone, E., Rousso, B. Z., & Kufeji, D. (2023). A probabilistic decision support tool for prediction and management of rainfall-related poor water quality events for a drinking water treatment plant. Journal of Environmental Management, 332, 117209. https://doi.org/10.1016/j.jenvman.2022.117209

Beltran Velez, N., Grande, A. A., Nazaret, A., Kucukelbir, A., & Blei, D. (2024). Treeffuser: probabilistic prediction via conditional diffusions with gradient-boosted trees. Advances in Neural Information Processing Systems, 37, 118296-118325. https://doi.org/10.52202/079017-3756

Brum, M., Fan, F. M., Salla, M. R., & von Sperling, M. (2022). Analysis of a probabilistic approach for modelling and assessment of the water quality of rivers. Journal of Hydroinformatics, 24(4), 783-797. https://doi.org/10.2166/hydro.2022.157

Cai, W., Wei, R., Xu, L., & Ding, X. (2022). A method for modelling greenhouse temperature using gradient boost decision tree. Information Processing in Agriculture, 9(3), 343-354. https://doi.org/10.1016/j.inpa.2021.08.004

del Castillo, A. F., Garibay, M. V., Díaz-Vázquez, D., Yebra-Montes, C., Brown, L. E., Johnson, A., ... & Gradilla-Hernández, M. S. (2024). Improving river water quality prediction with hybrid machine learning and temporal analysis. Ecological Informatics, 82, 102655. https://doi.org/10.1016/j.ecoinf.2024.102655

Goggins, W. B., Woo, J., Ho, S., Chan, E. Y., & Chau, P. H. (2012). Weather, season, and daily stroke admissions in Hong Kong. International journal of biometeorology, 56, 865-872. https://doi.org/10.1007/s00484-011-0491-9

Gandhimathi, G., Chellaswamy, C., Geetha, T. S., & Arunmozhi, S. A. (2025). Integrating Regression and Boosting Techniques for Enhanced River Water Quality Monitoring in the Cauvery Basin: A Seasonal and Sustainable Approach. Water Environment Research, 97(7), e70128. https://doi.org/10.1002/wer.70128

Ghaemi, E., Tabesh, M., & Nazif, S. (2022). Improving the ARIMA model prediction for water quality parameters of urban water distribution networks (case study: CANARY dataset). International Journal of Environmental Research, 16(6), 98. https://doi.org/10.1007/s41742-022-00482-x

Hajihosseinlou, M., Maghsoudi, A., & Ghezelbash, R. (2023). A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. Natural Resources Research, 32(6), 2417-2438. https://doi.org/10.1007/s11053-023-10249-6

Islam, M. J., Salekin, S. U., Anzum, A., Zaman, N., Khan, A. A. A., Sarkar, D., ... & Hossain, M. T. (2024). Machine Learning-Driven Water Quality Index Prediction: Enhancing Accuracy with Gradient Boosting and Explainable AI for Sustainable Water Monitoring. Applied Agriculture Sciences, 2(1), 1-14. https://doi.org/10.25163/agriculture.2110031

Ibrahim, O. R., Valizadeh, G., Meahrayen, M. A., Sarwari, S., & Abdi, E. (2026). A Combination Model Based on Sequential General Variational Mode Decomposition Method for Improving Lake Water Level Prediction. *Environmental Processes*, *13*(2), 36.

Jain, S., Bawa, A., Mendoza, K., Srinivasan, R., Parmar, R., Smith, D., ... & Johnston, J. M. (2025). Enhancing prediction and inference of daily in-stream nutrient and sediment concentrations using an extreme gradient boosting based water quality estimation tool-XGBest. Science of the Total Environment, 963, 178517. https://doi.org/10.1016/j.scitotenv.2025.178517

Lei, A., Lam, K. S. P., & Hu, Z. (2011). Comparison of two sampling methods when studying periphyton colonization in Lam Tsuen River, Hong Kong, China. Chinese Journal of Oceanology and Limnology, 29(1), 141-149. https://doi.org/10.1007/s00343-011-9961-5

Liu, Y., Zheng, N., Yang, S., Liu, F., Liu, M., & Chen, Y. (2025). Identification of water pollution sources in the Daluxi River using kernel principal component analysis and gradient boosting decision tree. Environmental Earth Sciences, 84(9), 236.

Ly, Q. V., Nguyen, X. C., Lê, N. C., Truong, T. D., Hoang, T. H. T., Park, T. J., ... & Hur, J. (2021). Application of Machine Learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the Han River, South Korea. *Science of The Total Environment*, *797*, 149040.

Miller, T., Durlik, I., Adrianna, K., Kisiel, A., Cembrowska-Lech, D., Spychalski, I., & Tuński, T. (2023). Predictive modeling of urban lake water quality using machine learning: A 20-year study. *Applied Sciences*, *13*(20), 11217.

Mendil, M., Mossina, L., and Vigouroux, D. (2023). PUNCC: A python library for predictive uncertainty calibration and conformalization. In Conformal and Probabilistic Prediction with Applications (pp. 582–601). PMLR.

Merabet, K., Di Nunno, F., Granata, F., Kim, S., Adnan, R. M., Heddam, S., ... & Zounemat-Kermani, M. (2025). Predicting water quality variables using gradient boosting machine: global versus local explainability using SHapley Additive Explanations (SHAP). Earth Science Informatics, 18(3), 1-34. https://doi.org/10.1007/s12145-025-01796-y

Nayan, A. A., Kibria, M. G., Rahman, M. O., & Saha, J. (2020, November). River water quality analysis and prediction using GBM. In 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT) (pp. 219-224). IEEE. https://doi.org/10.1109/icaict51780.2020.9333492

Nallakaruppan, M. K., Gangadevi, E., Shri, M. L., Balusamy, B., Bhattacharya, S., & Selvarajan, S. (2024). Reliable water quality prediction and parametric analysis using explainable AI models. Scientific Reports, 14(1), 7520. https://doi.org/10.1038/s41598-024-56775-y

Nazaret, A. O. (2025). Probabilistic Machine Learning for Predictions and Causal Discovery in Health Informatics (Doctoral dissertation, Columbia University).

Najwa Mohd Rizal, N., Hayder, G., Mnzool, M., Elnaim, B. M., Mohammed, A. O. Y., & Khayyat, M. M. (2022). Comparison between regression models, support vector machine (SVM), and artificial neural network (ANN) in river water quality prediction. Processes, 10(8), 1652. https://doi.org/10.3390/pr10081652

Olbert, A. I., Diganta, M. T. M., Bamal, A., Burke, W., Sajib, A. M., Abioui, M., ... & Uddin, M. G. (2025). Developing river water quality prediction model incorporating reliable indexing approach. Journal of Environmental Sciences. https://doi.org/10.2139/ssrn.5214350

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In Machine Learning: ECML 2002 (pp. 345–356). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-36755-1_29

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. Advances in Neural Information Processing Systems, 32.

Raheja, H., Goel, A., & Pal, M. (2024). A novel approach for prediction of groundwater quality using gradient boosting-based algorithms. ISH Journal of Hydraulic Engineering, 30(3), 281-292. https://doi.org/10.1080/09715010.2024.2316615

Satish, N., Anmala, J., Rajitha, K., & Varma, M. R. (2024). A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. Ecological Informatics, 80, 102500. https://doi.org/10.1016/j.ecoinf.2024.102500

Sharafi, M., Abdi, E., Mohebbiyan, M., & Samadianfard, S. (2024). Prediction of daily evapotranspiration using the strategy of combining tree models with empirical Hargreaves method. *Water and Soil Science*, *34*(2), 107-119.

Shah, M. I., Alaloul, W. S., Alqahtani, A., Aldrees, A., Musarat, M. A., & Javed, M. F. (2021). Predictive modeling approach for surface water quality: development and comparison of machine learning models. *Sustainability*, *13*(14), 7515.

Sattari, M. T., Feizi, H., Colak, M. S., Ozturk, A., Ozturk, F., & Apaydin, H. (2021). Surface water quality classification using data mining approaches: Irrigation along the Aladag River. Irrigation and Drainage, 70(5), 1227-1246. https://doi.org/10.1002/ird.2594

Sattari, M. T., Mirabbasi, R., Jarhan, S., Shaker Sureh, F., & Ahmad, S. (2020). Trend and abrupt change analysis in water quality of Urmia Lake in comparison with changes in lake water level. Environmental monitoring and assessment, 192(10), 623. https://doi.org/10.1007/s10661-020-08577-8

Sierra, I. S. (2024). Predictive Understanding of Lake Water Temperature and Dissolved Oxygen Profiles Across the Red River Basin Through Interpretable Machine Learning (Master's thesis, The University of Texas at El Paso).

Singh, P., Hasija, T., & Ramkumar, K. R. (2024, June). Leveraging ML with XGBoost, CatBoost and LGBoost Classifiers to Optimize Water Quality Assessment and Prediction. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (pp. 1-6). IEEE. https://doi.org/10.1109/iciteics61368.2024.10625322

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction, 129, 103827. https://doi.org/10.1016/j.autcon.2021.103827

Song, C., & Yao, L. (2022). A hybrid model for water quality parameter prediction based on CEEMDAN-IALO-LSTM ensemble learning. *Environmental Earth Sciences*, *81*(9), 262.

Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications, 83(12), 35307-35334. https://doi.org/10.1007/s11042-023-16737-4

Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic learning in a random world. Springer.

Yao, K., Chen, Y., Li, Y., Zhang, X., Zhu, B., Gao, Z., ... & Hu, Y. (2024). Water quality prediction of small-micro water body based on the intelligent-algorithm-optimized support vector machine regression method and unmanned aerial vehicles multispectral data. Sustainability, 16(2), 559.

Yu, R., & Zhang, C. (2021). Early warning of water quality degradation: A copula-based Bayesian network model for highly efficient water quality risk assessment. Journal of environmental management, 292, 112749. https://doi.org/10.1016/j.jenvman.2021.112749

Zheng, Y., Li, W., Fang, C., Feng, B., Zhong, Q., & Zhang, D. (2023). Investigating the impact of weather conditions on urban heat island development in the subtropical city of Hong Kong. Atmosphere, 14(2), 257.

Zhao, S., Liu, R., Liu, Y., Zeng, T., Chen, C., & Xu, L. (2025). A water quality prediction model based on modal decomposition and hybrid deep learning models. Water, 17(2), 184.

Zhou, S., Song, C., Zhang, J., Chang, W., Hou, W., & Yang, L. (2022). A hybrid prediction framework for water quality with integrated W-ARIMA-GRU and LightGBM methods. Water, 14(9), 1322. https://doi.org/10.3390/w14091322.



Copyright © 2026 The Author(s). This is an open-access article published by Khazar Journal of Water and Environment under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.