Journal of Water Process Engineering 78 (2025) 108734

Contents lists available at ScienceDirect

Journal of Water Process Engineering

journal homepage: www.elsevier.com/locate/jwpe

Beyond point predictions: Quantifying uncertainty in *E. coli* ML-based monitoring

David Abert-Fernández [a], Ester Aguilera [a,b], Pere Emiliano [a,c], Fernando Valero [c], Hèctor Monclús [a,*]

[a] LEQUIA, Institute of the Environment, University of Girona, C/M. Aurèlia Capmany, 69, Girona, 17003, Spain
[b] Amphos 21 Consulting SL, Carrer de Veneçuela, 103, Sant Martí, 08019, Barcelona, Spain
[c] Ens d'Abastament d'Aigua Ter-Llobregat (ATL), Departament d'R+D+i i Control de Processos, Sant Martí de l'Erm, 2, 08970, Sant Joan Despí, Barcelona, Spain

ARTICLE INFO

Editor: Soroush Abolfathi

Keywords:
Machine learning
Drinking water
Uncertainty quantification
Conformal prediction intervals
Microbial risk forecasting

ABSTRACT

Machine learning regression models are increasingly used to improve management, decision-making, and monitoring of drinking water quality, leveraging growing data from real-time sensors and laboratory analyses. However, most models provide only point predictions, ignoring inherent uncertainty caused by unobserved factors that can produce varying outcomes under similar conditions. This study benchmarks state-of-the-art regression algorithms and uncertainty quantification methods for predicting *E. coli* concentrations in a drinking water catchment. Gradient-boosted decision trees (GBDT) proved effective for real-time tracking, with CatBoost achieving the lowest error (RMSLE = 0.877), improving on the naïve baseline (1.160) and outperforming Random Forest by 5 %. Uncertainty quantification techniques successfully generated valid prediction intervals to identify high-risk contamination events, with Conformalized Quantile Regression emerging as the most reliable method. By combining accurate GBDT predictions with well-calibrated uncertainty estimates, this approach enhances microbial water quality forecasting, offering improved risk assessment and supporting more robust decision-making in drinking water management.

1. Introduction

Ensuring a safe drinking water supply is a critical global public health priority. Water quality in drinking water treatment plants (DWTPs) is monitored by collecting data from both laboratory analyses and real-time sensor measurements. This data encompasses a wide range of chemical, physical, and microbiological parameters that reflect the current state of water quality. While initially used for routine monitoring, the growing volume of collected data offers far more potential when leveraged through machine learning (ML) techniques. These algorithms can unlock insights beyond simple monitoring, enabling tasks such as predicting optimal treatment dosages [1], detecting pipeline leaks [2], developing early warning systems [3], and enhancing process control and optimization [4–6]. By harnessing the power of data, ML can transform water management, driving safer and more efficient treatment practices, granting a powerful environmental risk assessment.

In environmental risk assessment, uncertainty arises at multiple stages, from data collection and sampling to the final stages of a model development and prediction [7]. This uncertainty can generally be categorized into two main types: aleatoric uncertainty, arising from inherent variability in natural systems, and epistemic uncertainty, which results from limited knowledge or data gaps. Each stage of environmental data acquisition and modeling introduces variability that contributes to aleatory uncertainty, arising from natural system fluctuations, and epistemic uncertainty, stemming from sampling, measurement, processing inconsistencies, or modeling assumptions, both of which affect the reliability of risk assessment. Quantifying this uncertainty is particularly relevant in contexts where there is health risk associated, e.g. in drinking water supply systems. Integrating this uncertainty into decision making process is essential for producing more robust and reliable outcomes. This can improve decision-making [8], detect critical events [9], produce prediction intervals (PIs) and predictive distributions of risk-related outcomes [10], optimize processes [11] and reduce false positives and false negatives [12], helping managers to avoid under and over estimations.

The traditional regression algorithms provide point predictions,

[*] Corresponding author.
E-mail address: hector.monclus@udg.edu (H. Monclús).

https://doi.org/10.1016/j.jwpe.2025.108734
Received 30 June 2025; Received in revised form 28 August 2025; Accepted 9 September 2025
Available online 17 September 2025

offering a single estimated value for a given feature. However, these models fail to account for the uncertainty inherent in environmental systems [13]. Inaccurate or overconfident ML outputs that do not consider uncertainty quantification can lead to missed hazards that may compromise public health [14]. While international standards emphasize the need to quantify the uncertainty in water quality laboratory measurements [15–17], the underlying principles are equally applicable to modeling approaches. When developing regression predictive models, the uncertainty can be considered by providing PI instead of single point predictions. These intervals provide a range within which the true value is likely to fall, with a specified level of confidence indicated by the user [18]. To contextualize the integration of uncertainty quantification within the broader modeling process, Fig. 1 presents a typical machine learning regression workflow, extended to include uncertainty-aware predictions. The process starts with data generation, acquisition, and pre-processing to prepare inputs for a regression algorithm, which yields point predictions evaluated for performance. Based on this evaluation, preprocessing and modeling can be readjusted as needed, and through calibration, point predictions can be converted into PI that capture uncertainty.

Despite their value, prediction intervals are not widely used in water management and modeling, but some studies have addressed this gap by employing various techniques. In wastewater systems, a Bayesian framework has been used to quantify uncertainty in water quantity and quality simulations [19] and signal decomposition combined with adaptive kernel density estimation has been applied to generate dynamic prediction intervals for effluent quality [20]. Gaussian Process Regression has been used to produce predictive distributions in a papermaking wastewater system [21] and Monte Carlo simulations have been used to evaluate *E. coli* variability and calculate related health risks in a karst aquifer [22]. Uncertainty has also been incorporated into drinking water pipe break modeling through Bayesian Belief Networks, enhancing prediction capability and supporting asset management decisions [23]. In recent years, conformal prediction (CP) has gained significant attention as a promising method for uncertainty quantification. It offers distinct advantages for generating prediction intervals because it is a distribution-free framework that provides coverage guarantees. Unlike many other methods, CP does not require any assumptions about the data distribution, making it highly flexible and broadly applicable. Despite its capabilities CP has seen limited application in the water field. For instance, in groundwater applications, MAPIE algorithms have been integrated with Gradient Boosting Decision Trees (GBDT) to estimate heavy metal concentrations [24]. In surface and recreational waters, conformalized quantile regression (CQR) has been used to generate prediction intervals for Enterococci concentrations [25]. Moreover, machine learning models have produced prediction intervals with CP for water quality parameters based on data from uncrewed surface vessels and hyperspectral UAV sensors [26]. In urban water systems, CP has been incorporated into a hybrid CNN-BiLSTM model to improve hourly demand forecasts [27]. Most of the UQ-prediction intervals algorithms are built upon an existing regression algorithm, like GBDTs, which are widely used for both regression and classification tasks. This makes well-established in data-driven models for water quality prediction [28] and can be used as a strong foundation for prediction interval algorithms. GBDTs are based on the concept of boosting, where weak learners, such as decision trees, are ensembled into a robust predictive model. In most recent benchmarking studies, GBDT algorithms have demonstrated a superior predictive performance over other models due their ability to handle complex relationships in data, handle missing values and provide robust predictions [29,30]. Given these strengths, GBDT models are especially promising for critical steps in drinking water treatment where predictions are vital. One such application is Quantitative Microbial Risk Assessment (QMRA).

QMRA is a methodology used to estimate the health risks posed by pathogenic microorganisms, from environmental concentrations to exposure. This approach is increasingly valuable in managing water reuse related health risks [31]. For instance, recent work has proved that coupling hydrological simulations with QMRA can be used to enable an effective risk management of microbial scenarios in stormwater reuse [32] and that can be used to quantify the risk associated to specific bacteria [33]. A key component to perform these tasks is the reliable estimation of microbial concentrations, which often serves as proxies for pathogen presence. Among these, *E. coli* is a widely used organism as fecal contamination indicator [34]. Its presence is strictly regulated in the European Union, with a parametric value set at 0 CFU/100 mL in drinking water under European Directives [35], and thresholds ranging from 250 to 900 CFU/100 mL for recreational waters [36]. The traditional microbial quantification methods for these indicators, which are performed by microbial cultures, can take up 24 h to yield final concentrations [37]. This delay hinders the ability to adapt the drinking water treatment to water quality fluctuations. To deal with this challenge, data driven models can be employed to estimate the microorganism's concentration in raw water. Predictive models are particularly valuable in this context, as they not only provide real-time monitoring to complement traditional laboratory cultures but also offer more robust inputs for guiding operational decisions. A huge number of different regression algorithms have been used to predict microorganisms in water sources, e.g. artificial neural networks [38], Gaussian Process [39], Zero-Inflated regression models [40], Random Forest, Bayesian Belief Networks [41], Tree-based pipeline optimization tools [42] and tree-based ensemble models [43]. However, this area is particularly challenging due to the uncertainty that arises in microbial quantification [44]. Various sources contribute to this variability, including human and equipment errors in weighting, pipetting, preparation, and the risk of contamination during culture medium sterilization [45]. By integrating uncertainty into these models, all these variabilities can be accounted

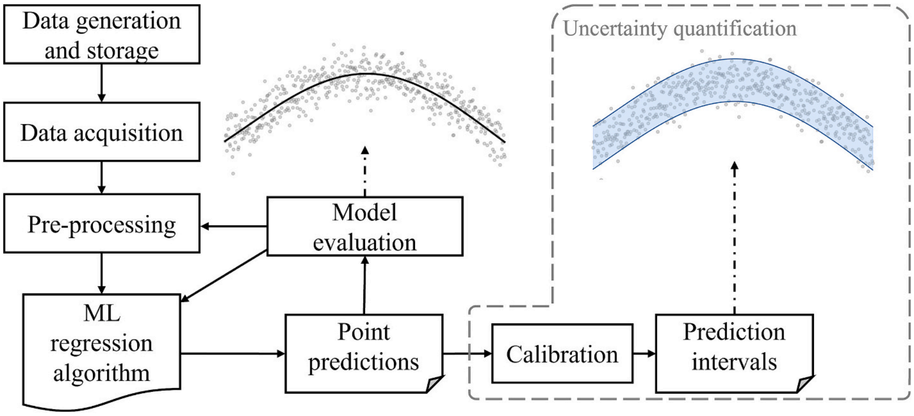

Fig. 1. General Machine Learning implementation pipeline.

for making predictions more reliable, supporting better risk management and improving the utility of QMRA in operational contexts [46]. In this way, the algorithms provide prediction intervals instead of single point estimates, resulting in a range of possible outcomes which allow for more conservative decisions under uncertain scenarios. For instance, this approach enables more informed decisions regarding chlorination strategies, allowing for adjustments in disinfectant dosing to ensure microbial safety while minimizing the formation of disinfection byproducts [47]. This approach becomes even more critical in an increasingly anthropized environment, where pressures on water bodies and evolving contamination patterns intensify the variability and unpredictability of raw water quality [48]. In such contexts, predictive models that incorporate UQ are essential for maintaining resilient water treatment operations. By explicitly accounting for uncertainty, these models can also serve as early warning systems, detecting peak contamination events that may not be reflected in historical data, as demonstrated in other fields [49,50]. These insights can guide the implementation of targeted control measures, in line with World Health Organization guidelines for drinking water [51]. UQ holds significant potential for improving real-time monitoring of microbial contamination in drinking water, yet it has not been applied in this context. Although some predictive models have shown promise in addressing various case studies, none have evaluated GBDT algorithms for FIB in drinking water catchments. Developing a tool capable of detecting microbiological risk concentration and producing uncertainty-aware prediction intervals in real-time would overcome delays on cultures results and would suppose an enhancement in microbiological risk assessment of a DWTP.

The aim of this paper is to develop and benchmark approaches for prediction intervals to improve the reliability of machine learning predictions by quantifying uncertainty, thereby addressing the lack of uncertainty quantification in drinking water quality modeling. To achieve this, two key objectives were defined: (i) to benchmark various gradient-boosted tree regression models and random forest against baseline models in order to identify the most suitable point prediction model for a specific case study, and (ii) to compare different model-agnostic and distribution-free algorithms, including CP for the first time in drinking water quality modeling for generating prediction intervals based on the predictions of the best-performing model identified in the benchmark. As a case study, this research examines the modeling of *E. coli* in a drinking water catchment, representing the first evaluation of uncertainty quantification algorithms to produce prediction intervals in this context, evaluating these intervals with a suitable metric, and suggesting different methods for incorporating uncertainty in regression problems.

2. Materials and methods

2.1. Study case

This study was conducted on the Llobregat DWTP catchment, located in Catalonia, Northeastern Spain (Fig. 2). The DWTP is managed by Ens d'Abastament d'Aigua Ter-Llobregat (ATL), the main drinking water supply company in the Barcelona Metropolitan Area, serving approximately 5.5 million inhabitants. The Llobregat River, which provides the surface water source for the DWTP, is subject to high anthropogenic pressure, receiving effluents from both urban and industrial wastewater treatment plants [52]. As a result, significant fluctuations of FIB occur in key water quality parameters, including organic matter concentration, salinity, temperature, and microbial contamination.

2.2. Data processing

The target variable in this study is *E. coli*, a key indicator of recent fecal contamination that is strongly correlated with public health risks [53]. To develop the predictive model, a comprehensive dataset was compiled, covering physicochemical, microbial, meteorological, and hydrological parameters collected daily from October 2000 to December 2023. Routine inlet water quality samples are collected at the DWTP intake by the Catalan Water Agency (ACA), the public agency responsible for water resource management in Catalonia. *E. coli* concentrations are measured every weekday at 7:00 a.m., while physicochemical parameters are continuously monitored through online sensors. A statistical summary of the water quality parameters and river flow conditions at the DWTP intake is provided in Table 1.

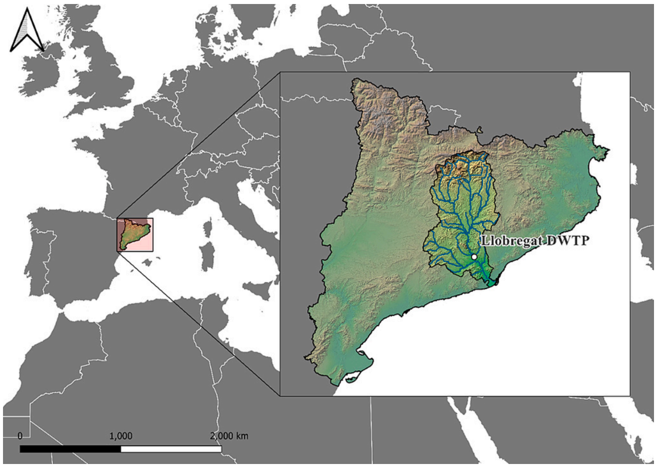

Fig. 2. Location of Llobregat DWTP operated by ATL and detailed region of Llobregat basin.

To enhance the model's predictive capabilities, multiple meteorological and hydrological monitoring points were selected to capture changes in river flow that could be influenced by wastewater discharge of the key drivers of fluctuations FIB concentrations. Precipitation data from three different locations within the Llobregat River basin were obtained from Meteorological Service of Catalonia, while real-time flow data from 12 monitoring points along the river and its tributaries were provided by ACA.

To account for the effects of previous rainfall events on *E. coli* concentrations, lagged precipitation features were incorporated into the dataset. Specifically, precipitation data from 1, 2, and 3 previous daily time steps were used for each observation. The number of lagged variables was determined based on domain knowledge, river basin hydrodynamics, and previous studies on microbial contamination transport in

Table 1
Statistical summary of quality parameters at Llobregat river, from 2000 to 2024.

	Units	Mean	Standard deviation	Min	Max
E. coli	MPN/ 100 mL	5116.88	12,916.47	73.00	244,200.00
UV absorbance 254 nm	m^{-1}	7.73	2.36	2.08	29.37
Ammonia	mg N-NH_4^+/L	0.24	0.23	0.00	3.68
Chlorides	mg Cl^-/L	251.74	89.52	29.72	1304.56
Conductivity	μS/cm	1277.43	283.16	531.00	4830.00
Nitrates	mg N-NO_3^-/L	8.98	3.70	1.77	34.69
Nitrites	mg N-NO_2^-/L	0.25	0.19	0.00	1.15
Dissolved oxygen	mg O_2/L	6.06	1.90	2.11	10.97
Total organic carbon (TOC)	mg C/L	4.46	1.36	0.91	10.89
Temperature	°C	15.76	6.46	2.40	28.90
Turbidity	NTU	69.87	224.72	0.26	8560.00
pH	pH units	8.11	0.17	7.32	9.03

surface water systems [54].

E. coli data shows a highly right-skewed distribution, indicating the presence of extreme values and deviation from a Gaussian distribution. As shown in Table 1, *E. coli* concentrations reach nearly 200,000 MPN/100 mL, suggesting occasional contamination peaks. This pattern indicates the presence of extreme and out-of-distribution values, which must be considered when selecting appropriate predictive models and uncertainty quantification methods.

For model training and evaluation, the dataset was split into two periods: 2000 to 2020 as the training set, and 2020 to 2023 as the testing set, corresponding approximately to an 85 % training and 15 % testing split. This temporal split was chosen due to the large size of the dataset, allowing for robust model development on extensive historical data while reserving recent data for unbiased evaluation. For uncertainty quantification, 20 % of the training data was set aside as a calibration set. All model training, calibration, and prediction interval construction were performed without using the testing set, which was strictly held out and used solely for final evaluation of model performance and prediction interval validity.

The main steps of the performed computational workflow are presented in Algorithm 1.

Algorithm 1. Point prediction model benchmarking.

```
1 Data Acquisition: collect meteorological, hydrological and water quality real data.
2 Data Splitting: Partition data into training (80%) and validation (20%).
3 Model benchmarking: for each regression model:
      Train on the training set
      Make point predictions on the validation set
      Compute performance metric score on validation set
6 PI evaluation: For each PI algorithm:
      Train PI algorithm with the best-performing point prediction model
      Generate prediction intervals for the validation set
      Compute PI performance metric
```

All the machine learning tasks were implemented in *Python 3.12.*

2.3. Point prediction

A benchmarking study was conducted to evaluate the predictive performance of different machine learning models for *E. coli* concentration forecasting. The study compared a naïve baseline method, which assumes that the most recent observation persists as the next predicted value, against a Random Forest (RF) model, and three GBDT algorithms, which were XGBoost (XGB), CatBoost (CB) and LightGBM (LGBM).

The naïve method was included as a baseline reference, while RF and GBDT models were selected for their ability to capture complex, nonlinear relationships in environmental data. To ensure that the most reliable point-prediction model was used for subsequent prediction interval estimation, the model that achieved the best balance between performance and the over- and under-prediction ratio in the benchmarking study was selected as the base model. To ensure optimal model performance, hyperparameter tuning was conducted using the Tree-structured Parzen Estimator algorithm, implemented via the Optuna library [55]. Each model underwent 60 min of optimization, allowing for an efficient search across the hyperparameter space. The following sections provide a brief description of each model, with further details available in Supplementary Text S1 and S2.

2.3.1. Naïve method

To establish a baseline for comparison, we employed the Naïve persistence model, which assumes that the most recent observation remains unchanged in the subsequent time step. This simple yet effective approach is commonly used in time series forecasting as a reference to assessing the improvement provided by more sophisticated models. Mathematically, it is expressed as:

$$Y_t = Y_{t-1} \tag{1}$$

where Y_t represents the predicted value at time t, and $Y_{t\text{-}1}$ is the observed value at the previous time step.

2.3.2. Random Forest

RF is an ensemble ML technique used for both classification and regression tasks [56]. It models complex relationships between predictor variables by combining multiple decision trees and aggregating their output, using majority voting for classification and averaging for regression. Each tree is built using a random subset of the training data (bootstrap sampling), and at each node, a random subset of features is considered for splitting. This randomness helps decorrelate trees, reducing overfitting and improving generalization to unseen data. RF is particularly effective for high-dimensional datasets and can handle missing values while maintaining robustness against noise.

2.3.3. Gradient boosting decision trees

GBDT are a family of ensemble learning algorithms used for regression and classification tasks. Unlike RF, which builds trees independently, GBDT trains trees sequentially, where each new tree corrects the errors made by previous ones. This process iteratively minimizes a loss function, improving the model's overall accuracy over multiple boosting iterations. Three GBDT models were evaluated in this study: CB, XGB and LGBM, which were implemented with Python 3.12 using the libraries scikit-learn 1.6.1 [57], XGBoost 2.1.4 [58], LightGBM 4.6.0 [59], and CatBoost 1.2.8, respectively [60].

2.3.4. Point prediction metrics

To assess the predictive performance of the models, the Root Mean Square Logarithmic Error (RMSLE) was selected as the loss function. RMSLE is defined as:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(x_i+1) - \log(y_i+1)\right)^2} \tag{2}$$

where x_i represents the predictive value, y_i is the observed value and n is the total number of observations.

The choice of RMSLE for FIB quantification is driven by several key factors related to the nature of the data. First, FIB concentrations often follow highly skewed, non-linear, making traditional error metrics less effective. Additionally, measurement errors in culture-based microbial quantification tend to grow exponentially at higher concentrations. By applying a logarithmic transformation, RMSLE stabilizes variance across different concentration levels, mitigating this effect. Finally, unlike mean absolute error or root mean squared error, which focus on absolute differences, RMSLE emphasizes relative deviations, making it particularly suitable for contamination levels that span multiple orders of magnitude [61].

Furthermore, since the logarithmic scale is asymmetric, the RMSLE metric inherently penalizes underpredictions more than overpredictions. This property is critical in drinking water safety, as underestimating FIB concentration poses a direct risk to public health, potentially leading to inadequate treatment response. By contrast, a slight overprediction results in more conservative safety measures, which are preferable in this context.

2.4. Prediction interval

PIs provide a range within which a future observation is expected to fall, given a specified confidence. Unlike confidence intervals, which quantify the uncertainty of estimated parameters (e.g., the mean of a population), PIs capture uncertainty at the individual predictions level, incorporating both model uncertainty and intrinsic data variability. This makes PIs particularly valuable in environmental and microbial risk assessment, where uncertainty can arise from multiple sources, including measurement errors and dynamic systems fluctuations.

In this study, the methods for uncertainty quantification and PI generation were selected based on two key criteria: i) model agnosticism: the methods should be applicable to wide range of predictive models without requiring specific structural assumptions. ii) Distribution-free properties: the techniques should not rely on predefined statistical distributions, making them more flexible for real-world microbial water quality data. Based on these criteria, the following PI estimation techniques were benchmarked: a naïve data split approach, resampling methods (CV+, Jackknife+, and a split method [62]), quantile regression (QR) and CQR. For this case study, the algorithms were designed to target a 90 % prediction interval coverage, meaning that under typical conditions, 90 % of true outcomes are expected to fall within the predicted intervals, thereby providing a reliable measure of uncertainty while avoiding excessively wide intervals. The only assumption for the implemented models is the exchangeability of the data, meaning that the joint probability distribution is unchanged if the order of the observations is permuted. In the case of a highly anthropized, high-flow river, *E. coli* levels show little to no temporal autocorrelation because conditions vary substantially over short time scales due to irregular human inputs and rapid water turnover, making the samples effectively independent in time and thus approximately exchangeable. The resampling methods were implemented through MAPIE v1.0.0 library [63], QR was implemented with the native algorithm of each GBDT library, and the CQR approach was implemented using adaptations from the solution of Probabilistic Forecasting I: Temperature competition-winning methodology.

2.4.1. Metric for prediction interval

To evaluate the accuracy and quality of the PI, the Winkler Interval Score (WIS) was used. This metric evaluates both width of the interval and whether the observed value falls within it, providing a balance measures of interval reliability [64]. The WIS penalizes excessively wide intervals, as well as cases where the observed value falls outside the predicted bounds. Formally the WIS is defined as (Eq. (3)).

$$W(y;L,U) = \begin{cases} U-L & \text{if } L \le y \le U, \\ (U-L) + \dfrac{2}{\alpha}(L-y) & \text{if } y < L, \\ (U-L) + \dfrac{2}{\alpha}(y-U) & \text{if } y > U. \end{cases} \tag{3}$$

where y is the observed value, L and U represents the lower and upper bounds of the prediction interval, respectively and α is the penalty factor for predictions outside the interval.

The penalty factor (α) determines the level of penalization applied when the observed value falls outside the prediction interval. A lower α value results in stricter penalties for non-coverage, while a higher α value tolerates more outliers in exchange for narrower intervals. In this study, $\alpha = 0.1$ was chosen, placing greater emphasis on penalizing non-coverage rather than increasing interval width. This choice reflects the importance of ensuring robust uncertainty quantification, particularly in a public health context, where underestimation of microbial risk can have significant consequences. The evaluation of the model with WIS was computed over the validation set of data, which was not used by the algorithm during the training.

3. Results and discussion

3.1. Point-prediction modeling benchmarking

The predictive performance of GBDT algorithms, RF, and a naïve baseline was benchmarked using RMSLE for forecasting *E. coli* concentrations.

3.1.1. Model performance

Table 2 presents the RMSLE scores, number of underpredictions and overpredictions, and the underprediction-to-overprediction ratio (U/O ratio) for each model.

All ML models outperformed the naïve baseline, which recorded an RMSLE of 1.160, confirming the added predictive value of ML-based approaches. Among the benchmarked models, LGBM achieved the lowest RMSLE (0.869), followed closely by CB (0.877), XGB (0.878), and RF (0.917). The small differences in RMSLE indicate that all four models provide comparable accuracy, yet their individual strategies for handling errors differ significantly. Analyzing more in detail the results, while LGBM exhibited the best RMSLE score, it also recorded the highest number of underpredictions, resulting in the largest U/O ratio (1.075).

Table 2
Descriptive analysis of performance and bias of regression models.

Model	RMSLE	Underpredictions	Overpredictions	U/O ratio
Naïve	1.160	375	396	0.947
CB	0.877	342	455	0.752
XGB	0.878	363	434	0.837
LGBM	0.869	413	384	1.075
RF	0.917	361	436	0.828

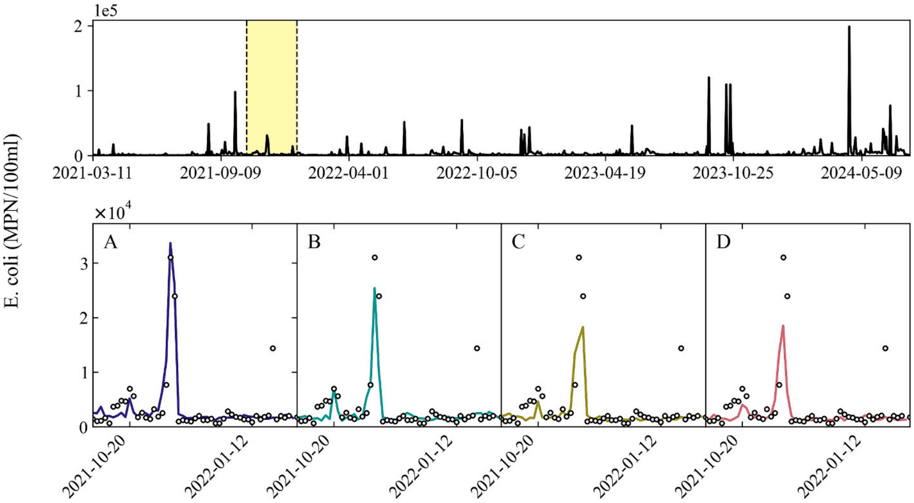

Fig. 3. Global time series of observed *E. coli* concentrations (top panel) with a yellow-shaded region indicating the selected interval. The lower panels provide a zoomed-in view of this period, comparing observed values (markers) with model predictions: (A) CB, (B) XGB, (C) LGBM, and (D) RF.

This suggests that LGBM priories accuracy in low-concentration scenarios, potentially at the cost of underestimating peak contamination levels. Conversely, CB achieved the lowest U/O ratio (0.751) by minimizing underpredictions, making it more conservative in its risk estimation.

3.1.2. Predicted vs observed E. coli values

Fig. 3 presents a time series plot comparing observed *E. coli* concentrations with model predictions over a specific period that includes both baseline values and a peak contamination event.

CB enhanced predictive performance during peak events by minimizing underpredictions, making it particularly valuable for early warning systems and risk mitigation strategies. Its superior peak detection outperformed the other models. The model's tendency to favor overpredictions in uncertain or extreme cases reflects a prioritization of safety, helping to ensure that high-risk events are less likely to be missed. These results are closely followed by XGB, which showed a lower capacity to accurately quantify peak events but demonstrated overall good performance for lower concentration values. A similar pattern was observed for RF, which exhibited a prediction trend comparable to XGB but with higher associated errors. LGBM, on the other hand, provided more stable predictions within commonly observed concentration ranges, making it a reliable option under regular operational conditions. These results highlight the trade-offs between models: CB tends to prioritize safety by reducing underpredictions, while LGBM excels in low-to-moderate range predictions but underestimates peak

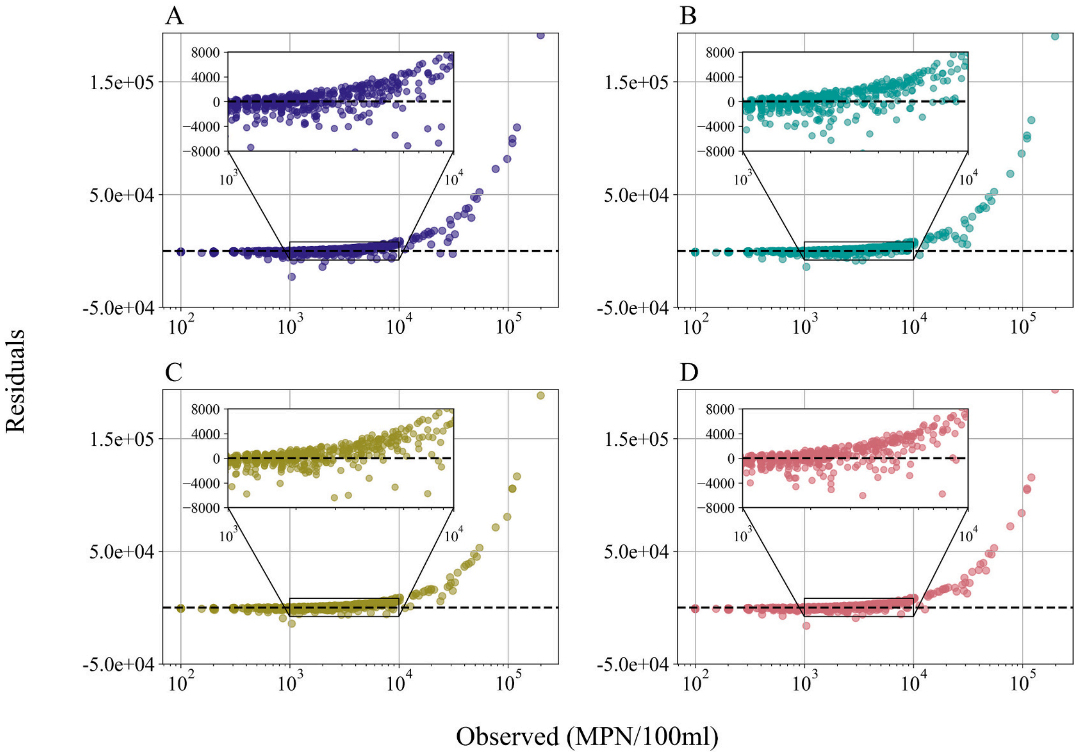

Fig. 4. Residuals of the models along different observed ranges. (A) CB model, (B) XGB model, (C) LGBM model, (D) RF model. The insets in each plot show a zoom-in view of the residuals in the observed range from 10^3 to 10^4.

contamination levels. This behavior aligns with the findings in Table 2, where LGBM had the lowest RMSLE but also the highest U/O ratio.

A detailed comparison of predicted versus observed *E. coli* concentrations is shown in Supplementary Fig. S1, illustrating the distribution of prediction errors alongside model-fitted and perfect agreement reference lines.

3.1.3. Residuals analysis

Fig. 4 displays the residuals (difference between observed and predicted values) across different *E. coli* concentration ranges for the four evaluated models. The residual analysis helps identify systematic biases and variations in model performance across contamination levels.

Across all models, prediction errors increase significantly at higher *E. coli* concentrations, showing a systematic bias in high concentrations and reflecting the challenges of accurately predicting extreme contamination events. This behavior is likely due to the scarcity of high-concentration samples in the training set, limiting the models' ability to generalize in these ranges.

As shown in Fig. 4, LGBM exhibits a greater dispersion of residuals above zero compared to the other models, indicating a higher degree of underestimation. This aligns with LGBM's higher U/O ratio observed in Table 2. In contrast, CB exhibits more positive residuals in *high E. coli* ranges, indicating fewer underpredictions. Its residuals are also closer to zero than those of other models, suggesting it better captures contamination levels, which is a crucial factor for early warning systems. The zoom-in boxes highlight residual behavior in low-moderate concentration ranges, where most models exhibit lower residual variance, indicating higher predictive accuracy in moderate contamination scenarios. RF and XGB have also shown residuals with a tendency to underestimate high *E. coli* concentrations. However, these models have adopted an intermediate approach between CB and LGBM, exhibiting a higher number of underpredictions than CB and less than LGBM, but with residuals closer to zero.

From a public health and decision-making perspective, underprediction is more critical than overprediction, as it may lead to undetected contamination risks. In this context, CB appears to be the most reliable model for risk assessment, as it reduces the likelihood of underestimating peak contamination events. LGBM, while achieving the lowest RMSLE, tends to underpredict high-risk cases, which may limit its applicability in early warning systems. XGB and RF offer a balanced alternative, providing consistent performance across different concentration ranges, but without a strong prioritization of safety-driven predictions. These residual patterns reinforce the importance of quantifying uncertainty using prediction intervals, as discussed in the following Section 3.2, to provide decision-makers with more reliable risk assessments.

3.2. Prediction interval benchmarking

To generate the PI, multiple distribution-free and model-agnostic methods were benchmarked. CB was selected as the base model, given its slightly superior performance in point-prediction benchmarking (Section 3.1), particularly in minimizing underpredictions, which is crucial in risk-sensitive applications. Model performance was assessed using the mean WIS, a metric that balances interval width and coverage quality. Additional key indicators included the percentage of samples below and above the interval bounds, as well as overall coverage.

Table 3
Descriptive analysis of the performance in model benchmarks.

Model	WIS ($\times 10^4$)	% samples below	Coverage %	% samples above	Mean interval width ($\times 10^4$)
Naive	2.34	12.0	83.7	4.3	1.4
CV	2.69	5.1	93.1	1.7	2.39
CV+ minmax	4.53	1.7	97.7	0.5	4.31
CV+	2.76	3.9	94.9	1.2	2.36
Jackknife-minmax-AB	2.47	5.3	92.2	2.5	1.5
Jackknife+ AB	2.59	10.5	85.8	3.6	1.03
CQR	2.08	4.5	92.4	3.1	1.57
QR	2.19	13.0	83.7	3.3	1.17

3.2.1. Model performance of prediction interval

Table 3 compares the benchmarked prediction interval methods using WIS, coverage, and mean interval width. CQR consistently outperformed other approaches, achieving the lowest WIS (2.08 × 10^4), solid coverage (92.4 %), and a balanced proportion of samples outside the interval bounds—demonstrating its reliability and efficiency in capturing uncertainty without compromising actionability.

While QR showed the second-best WIS (2.19 × 10^4), its coverage fell below the 90 % target, with 13 % of samples below the lower bound, confirming a tendency to produce intervals that are too narrow and prone to underestimating risk, especially during contamination peaks.

Jackknife-based methods generally offered a better trade-off than CV-based methods. The Jackknife-minmax-AB variant ensured high coverage through conservative interval construction, resulting in a lower WIS than standard Jackknife+, though at the cost of wider intervals. Conversely, Jackknife+AB yielded narrower intervals but missed more extreme values, increasing WIS penalties.

CV-based methods all exceeded the 90 % coverage threshold but produced broader intervals overall. Notably, CV-minmax delivered the highest coverage (97.7 %) with the widest intervals (and the highest WIS: 4.53 × 10^4), while CV and CV+ struck a better balance between width and reliability.

In drinking water monitoring, prediction intervals must be both dependable and practical. Methods like QR, which sacrifice coverage for narrowness, may be unsuitable for risk-sensitive environments. CQR stands out as the most balanced solution: it ensures robust coverage, maintains actionable interval widths, and is distribution-free, making it adaptable to diverse datasets and ideal for integration into early warning systems and decision support tools.

3.2.2. Predicted intervals

Fig. 5 provides a visual representation of the PIs over a selected period of N days, highlighting how different methods respond to fluctuations in *E. coli* concentrations. The black line represents the observed values, while the blue shaded areas denote the predicted intervals for each method.

During peak concentration events, all models tend to produce wider intervals, reflecting the higher uncertainty associated with these extreme values. This aligns with the findings in Fig. 5, where models exhibited larger residuals in high-concentration ranges. Jackknife-minmax-AB, CQR and QR tend to generate narrower intervals compared to other methods, particularly in peak scenarios, although QR does not reach the target coverage, as seen in Table 3. This suggests both CQR and Jackknife-minmax-AB methods efficiently balance coverage and interval width, reducing excessive overestimation of uncertainty.

In low-concentration periods, Jackknife-minmax-AB, CQR, and QR consistently yield tighter prediction intervals, indicating higher confidence in predictions when contamination levels are stable. This property is advantageous in operational settings, where precise risk estimation is needed without unnecessary uncertainty inflation.

Overall, this Fig. 5 reinforces the quantitative findings from Table 3, demonstrating that CQR and Jackknife-based methods provide the most adaptive prediction intervals, adjusting their width dynamically based on contamination levels. These characteristics make them particularly well-suited for early warning systems and decision support frameworks, ensuring that uncertainty is accounted for without compromising actionability.

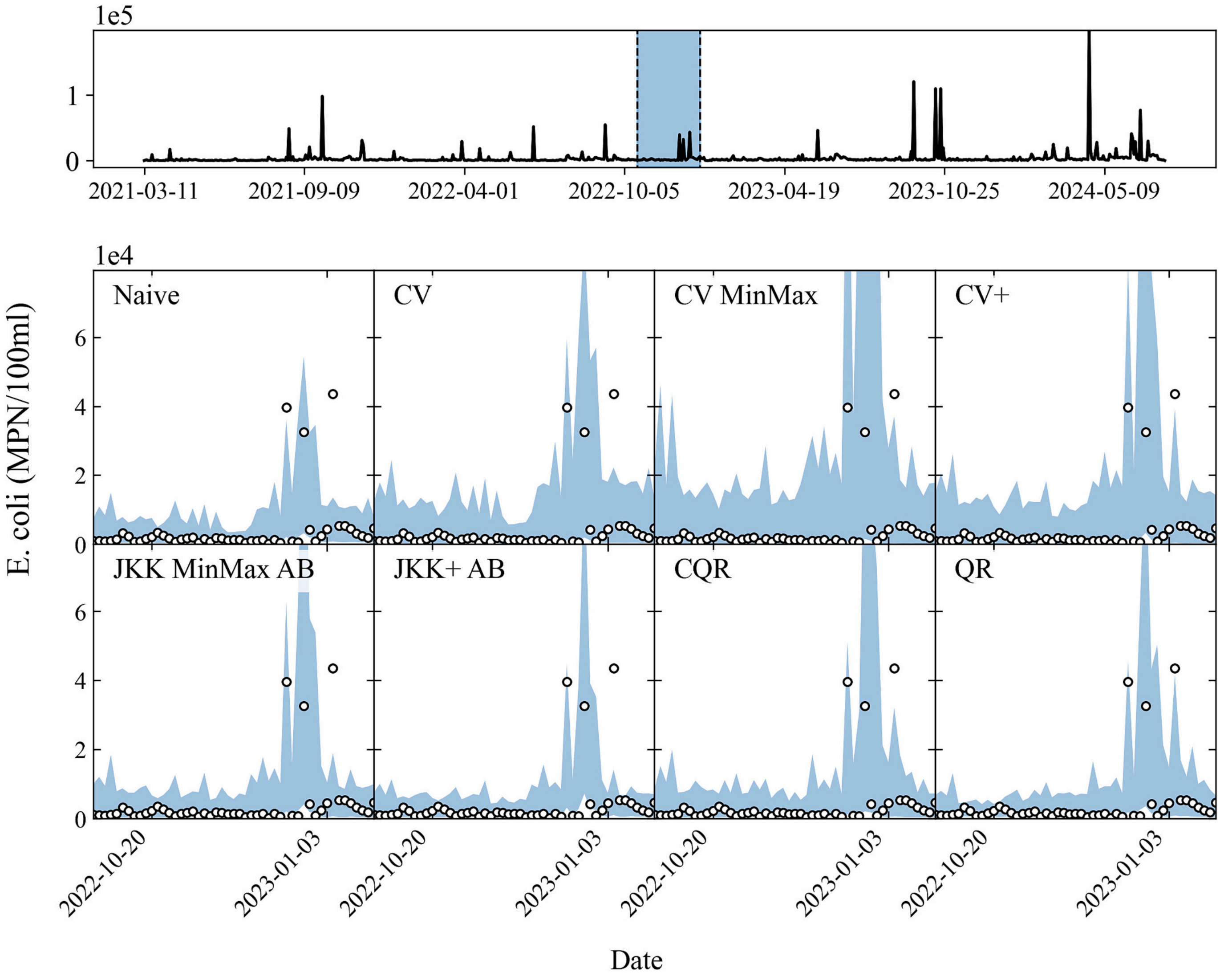

Fig. 5. Time series of observed *E. coli* concentrations (top panel), with the blue-shaded region indicating the period selected for detailed analysis. The lower panels provide an expanded view of this interval, showing 90 % prediction intervals generated by different models. Observed values are overlaid as markers, and shaded areas represent model uncertainty.

3.2.3. Coverage analysis

Fig. 6 shows a detailed view of the coverage and interval width of the prediction interval benchmarked algorithms across E. coli concentration ranges. Three distinct concentration ranges can be distinguished based on coverage: low values ($<10^3$ MPN/100 mL), medium values, which are the most common and range approximately between 10^3 and $10^{3.5}$/ 10^4 MPN/100 mL, and high values ($>10^4$ MPN/100 mL). The global trade-offs between coverage and interval follows the same pattern for all the algorithms. Although the overall coverage reaches 90 % as seen in Table 3, the coverage in low values does not reach the target for almost any of the models. In medium ranges, which are the most common ones, the predictions cover most of the observed values. In high *E. coli* values, the coverage slightly decreases in all algorithms while the interval width increases in these ranges.

QR, Jackknife+AB, and Naïve methods showed the lowest coverage at low ($<10^3$) and medium ($<10^4$) E. coli concentrations. Although they produced the narrowest intervals in these ranges, their failure to capture the true values indicates that they do not adequately represent the variability in the data, resulting in invalid prediction intervals.

At the other extreme, CV-minmax generated the widest intervals across the full range of concentrations, achieving consistently high coverage. This conservatism ensures that true values are rarely missed but reduces the informativeness of predictions. Other CV-based approaches, such as CV and CV+, are somewhat more adaptive: their intervals are narrower and more informative, yet still expand at higher E. coli levels, reflecting the inherent uncertainty in those regions.

CQR and Jackknife-minmax-AB provide a more balanced solution, producing informative intervals with relatively high coverage across all concentrations. CQR performs especially well at mid-range levels, capturing central variability while keeping intervals reasonably narrow, whereas Jackknife-minmax-AB slightly outperforms at the highest concentrations.

Across all methods, interval width generally increases with *E. coli* concentration, reflecting greater uncertainty at extreme values. This pattern suggests that wider intervals can serve as an early indicator of potentially elevated *E. coli* levels, offering valuable information for risk assessment. Narrow intervals at lower concentrations indicate higher confidence but may underrepresent rare high-concentration events if coverage is insufficient.

Overall, CQR and Jackknife-minmax-AB emerge as the most suitable methods for this study. By balancing coverage and interval width, they provide reliable and informative uncertainty estimates across the full range of E. coli concentrations, making them the most practical choices for monitoring and risk assessment.

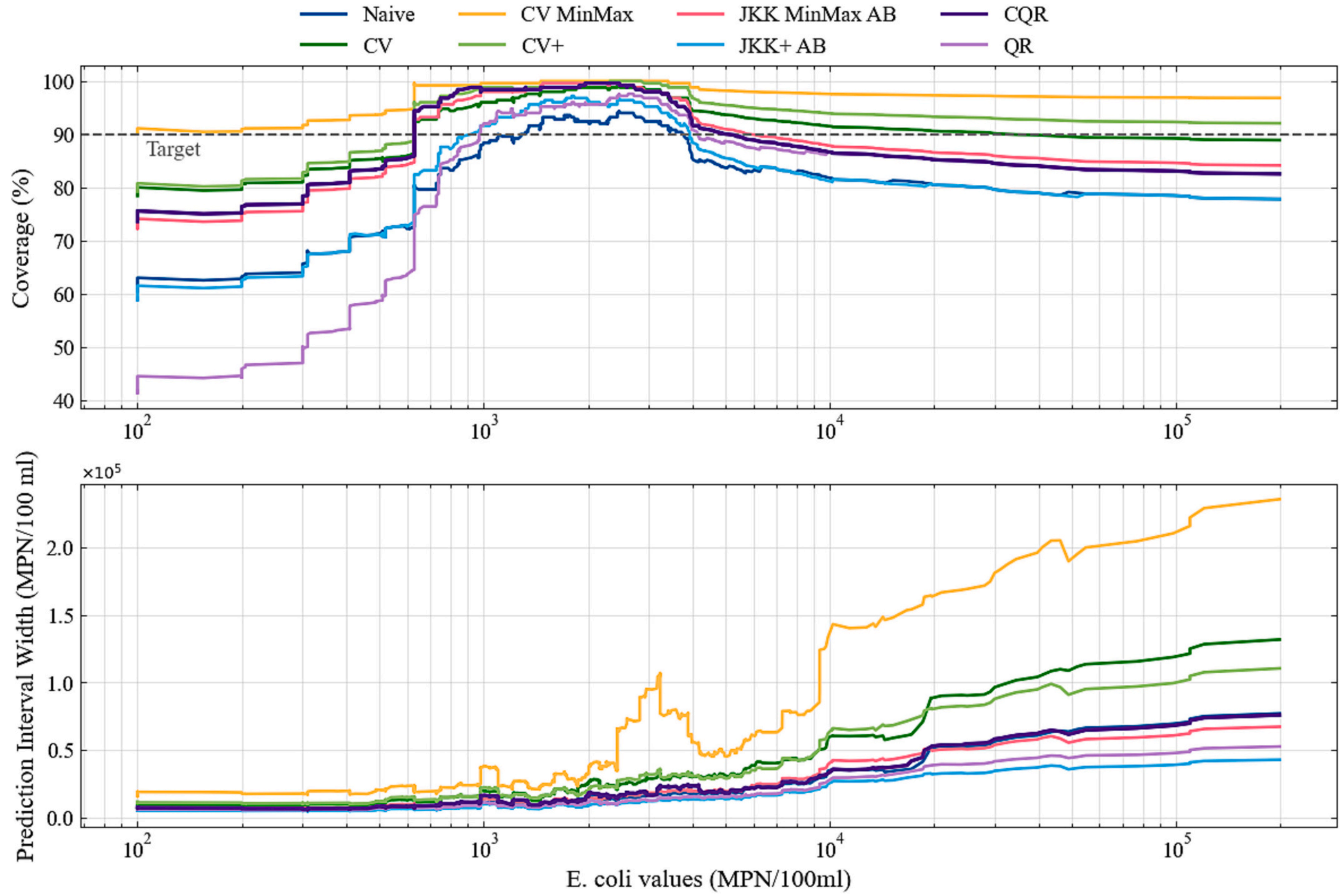

Fig. 6. Prediction interval widths and coverage across observed *E. coli* concentrations. Both interval and coverage were smoothed to highlight trends across the observed *E. coli* range.

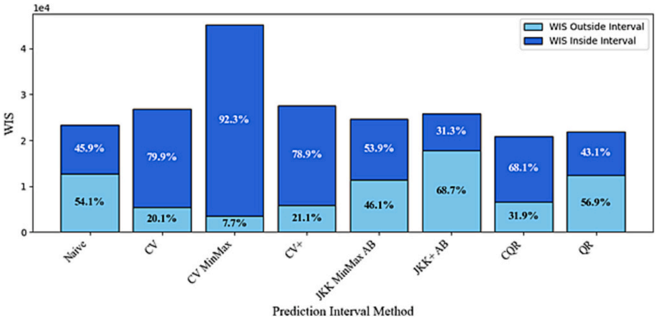

Fig. 7. Decomposition of the WIS for each prediction interval method, distinguishing between WIS contributions from predictions inside and outside the interval.

3.2.4. *Prediction interval reliability*

Fig. 7 presents a decomposition of the WIS into contributions from predictions that fall inside and outside the interval bounds. This breakdown provides a clearer picture of each method's overall performance while also illustrating how they manage miscoverage, an essential aspect of reliability in uncertainty quantification.

While methods like CQR and Jackknife+AB achieved similar coverage and average interval widths, their WIS values are distributed differently. This difference stems from how they respond to miscoverage: Jackknife+AB tends to compensate for missed observations, especially at the extremes, by producing excessively wide intervals, which sharply increases its WIS penalty. In contrast, CQR maintains more consistent and moderate interval widths even when failing to fully capture extreme values, resulting in a lower and more stable WIS.

A similar pattern of trade-offs emerges when comparing the CV-based and Jackknife-minmax-AB methods. CV-minmax stands out for its near-perfect coverage but does so at the cost of very broad intervals and the highest WIS, highlighting the inefficiency of overly conservative approaches. On the other hand, CV, CV+, and Jackknife-minmax-AB strike a more balanced compromise. Their WIS contributions are more evenly distributed between coverage failures and interval width, indicating a reasonable calibration of uncertainty. Notably, CV+ improves slightly over CV in terms of coverage, with only a minor increase in WIS, offering a practical middle ground between robustness and usability. By contrast, QR and the Naïve model perform poorly, with high WIS values driven largely by frequent and substantial miscoverage. Their intervals are too narrow to capture the true variability of *E. coli* concentrations, particularly in high-risk situations, resulting in significant penalties that reflect their unreliability in predictive risk assessment.

The CQR method remains the most balanced approach, minimizing WIS while maintaining good coverage and reasonable interval width. This aligns with broader findings in microbial water quality forecasting where this method produces risk-based prediction intervals [25]. Jackknife+-AB, despite its strong performance in coverage, shows a slightly excessive widening of intervals when failing to cover observations, resulting in higher WIS penalties and making it less optimal. CV+ and Jackknife-minmax-AB provide reasonable alternatives for balancing coverage and width, particularly in moderate-risk scenarios where overconfidence needs to be avoided. QR and the used Naïve method are the least reliable in terms of uncertainty quantification, as they lead to more frequent and severe out-of-interval penalties, making them unsuitable for early warning systems and risk mitigation strategies. These results underscore the importance, as noted in recent water quality modeling literature [20] of selecting interval prediction methods that achieve target coverage without excessive interval expansion, thus ensuring actionable and interpretable uncertainty quantification in microbial risk assessment [31]. These results reinforce the importance of selecting an interval prediction method that not only meets target coverage but also controls interval expansion efficiently, ensuring actionable and interpretable uncertainty estimates in microbial risk assessment.

4. Conclusions

The benchmarked point-prediction models showed similar performance in RMSLE scores, with LGBM, CB, XGB, and RF achieving comparable accuracy. However, differences emerged in how each model balanced underprediction and overprediction. LGBM prioritized performance in low concentration ranges, leading to a higher underprediction-to-overprediction ratio, whereas CB minimized underpredictions, making it a more conservative and risk-averse model. This trade-off is crucial when applying machine learning models to microbial risk assessment, where underestimating contamination events could have serious consequences. Although these trade-offs can be calibrated during training, peak contamination events remain particularly challenging to predict with point predictions. Uncertainty quantification can help address this limitation, not by precisely quantifying the magnitude of such events, but by identifying anomalous scenarios where the model's confidence decreases, thereby flagging situations that warrant closer investigation.

Among the uncertainty quantification methods, Jackknife-AB+ and CQR demonstrated the best performance, generating prediction intervals with strong guarantees. These methods achieved high coverage and controlled interval widths, ensuring that predictions were both informative and actionable. Notably, CQR adapted well to peak contamination events, making it particularly well-suited for early warning systems. Both CQR and Jackknife-AB+ produced narrow intervals during low-concentration scenarios, which is particularly useful for routine monitoring and regulatory compliance, as high precision in the absence of elevated risk enables more efficient resource allocation. Relative to the observed scale, their mean interval widths remain within an order of magnitude of the variability in *E. coli* concentrations, providing meaningful and actionable bounds for operational use.

A key insight from this study is that relying solely on point predictions limits the effectiveness of microbial risk assessment, particularly in scenarios with significant uncertainty. Incorporating uncertainty quantification through prediction intervals in predictions provides decision-makers with a clearer understanding of potential risks and confidence levels, enabling more informed and effective responses to contamination events. The uncertainty quantified in this study represents all sources of variability, including both environmental fluctuations and model limitations. Measurement noise, while present, is negligible compared to the observed changes in E. coli concentrations and does not significantly influence predictions. Part of the model error arises inherently from the uncertain and dynamic nature of the environment and cannot be separated from the model's outputs. Consequently, the prediction intervals capture the combined effects of environmental variability and model imperfection, providing a realistic estimate of total uncertainty and supporting informed interpretation of microbial risk under highly variable conditions.

In this case study, centered on predicting *E. coli* concentrations in the highly anthropized Llobregat River catchment, the pronounced skewness of the data highlights the significant impact of human activities on water quality. Since the catchment is heavily influenced by diverse and often untracked anthropogenic factors, there is inherently high uncertainty associated with the predictions. This strong asymmetry creates one of the most challenging scenarios for predictive modeling, as it requires capturing both frequent low-concentration values and rare but critical peak contamination events. The approach proved successful in this difficult context, with prediction intervals effectively representing the uneven distribution and achieving the targeted coverage levels. These results demonstrate the robustness of the methodology in providing reliable microbial risk estimates under complex and highly variable conditions. This supports its applicability not only to similarly impacted water sources but also to other ecosystems influenced by hydrometeorological and anthropogenic factors.

Future work should focus on integrating uncertainty quantification methodologies into QMRA decision-making frameworks, enabling guidance that recommends conservative actions when uncertainty is high and more flexible or optimized measures when uncertainty is low. In drinking water treatment, the current operational tendency is to over-disinfect to avoid microbial health risks, which can lead to the increased formation of disinfection by-product. By incorporating well-calibrated prediction intervals, water managers can better assess the likelihood and magnitude of contamination events, providing a basis for balancing microbial safety with chemical by-product formation. Adopting these UQ approaches allows managers to navigate uncertainty and incomplete data more effectively, ultimately improving water safety assessments, optimizing treatment plant control processes, and enhancing public health protection.

CRediT authorship contribution statement

David Abert-Fernández: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ester Aguilera:** Writing – original draft, Software. **Pere Emiliano:** Resources. **Fernando Valero:** Validation, Resources. **Hèctor Monclús:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

David Abert-Fernández thanks UdG for a predoctoral grant under the program IFUdG2023/2 and Hèctor Monclús gratefully acknowledges the Ramon y Cajal Research Fellowship (RYC2019- 026434-I). This study was supported ShERLOcK (PID2020-112615RA-I00) and WaterCLUE (CNS2023-143664) projects, financed by the Ministerio de Ciencia e Innovación (MICIN) and Agencia Estatal de Investigación (AEI) (Spain). LEQUIA has been recognized as a consolidated research group by the Catalan government (2021-SGR-1352).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jwpe.2025.108734.

Data availability

Data will be made available on request.

References

[1] S. Lin, J. Kim, C. Hua, M.-H. Park, S. Kang, Coagulant dosage determination using deep learning-based graph attention multivariate time series forecasting model, Water Res. 232 (2023) 119665, https://doi.org/10.1016/j.watres.2023.119665.

[2] S. Nahas, D. Ayala-Cabrera, Application of ground penetrating radar and machine learning tools in the identification of cracks in buried water pipes and creation of 3-

D models of the water leakage, 2023, https://doi.org/10.5194/egusphere-egu23-16353.
[3] Y. Xie, Y. Chen, Q. Wei, H. Yin, A hybrid deep learning approach to improve real-time effluent quality prediction in wastewater treatment plant, Water Res. 250 (2024) 121092, https://doi.org/10.1016/j.watres.2023.121092.
[4] Y. Ding, Q. Sun, Y. Lin, Q. Ping, N. Peng, L. Wang, et al., Application of artificial intelligence in (waste)water disinfection: emphasizing the regulation of disinfection by-products formation and residues prediction, Water Res. 253 (2024) 121267, https://doi.org/10.1016/j.watres.2024.121267.
[5] S. Moradi, A. Omar, Z. Zhou, A. Agostino, Z. Gandomkar, H. Bustamante, et al., Forecasting and optimizing dual media filter performance via machine learning, Water Res. 235 (2023) 119874, https://doi.org/10.1016/j.watres.2023.119874.
[6] D. Yun, D. Kang, J. Jang, A.T. Angeles, J. Pyo, J. Jeon, et al., A novel method for micropollutant quantification using deep learning and multi-objective optimization, Water Res. 212 (2022) 118080, https://doi.org/10.1016/j.watres.2022.118080.
[7] D.J.C. Skinner, S.A. Rocks, S.J.T. Pollard, A review of uncertainty in environmental risk: characterising potential natures, locations and levels, J. Risk Res. 17 (2014) 195–219, https://doi.org/10.1080/13669877.2013.794150.
[8] J.C. Cresswell, Y. Sui, B. Kumar, N. Vouitsis, Conformal Prediction Sets Improve Human Decision Making, 2024, https://doi.org/10.48550/ARXIV.2401.13744.
[9] R. Laxhammar, Conformal Anomaly Detection: Detecting Abnormal Trajectories in Surveillance Applications, University of Skövde, Skövde, 2014.
[10] V. Vovk, J. Shen, V. Manokhin, M. Xie, Nonparametric predictive distributions based on conformal prediction, Mach. Learn. 108 (2019) 445–474, https://doi.org/10.1007/s10994-018-5755-8.
[11] D. Rehman, F. Sheriff, J.H. Lienhard, Quantifying uncertainty in nanofiltration transport models for enhanced metals recovery, Water Res. 243 (2023) 120325, https://doi.org/10.1016/j.watres.2023.120325.
[12] A. Marandon, Conformal link prediction for false discovery rate control, 2024, https://doi.org/10.48550/arXiv.2306.14693.
[13] H. Tyralis, G. Papacharalampous, A review of predictive uncertainty estimation with machine learning, Artif. Intell. Rev. 57 (2024) 94, https://doi.org/10.1007/s10462-023-10698-8.
[14] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013−2023), Comput. Biol. Med. 165 (2023) 107441, https://doi.org/10.1016/j.compbiomed.2023.107441.
[15] Water Quality – General Requirements and Guidance for Microbiological Examinations, International Organization for Standardization, Geneva, Switzerland, 2018.
[16] Water Quality – Sampling – Part 1: Guidance on the Design of Sampling Programmes, International Organization for Standardization, Geneva, Switzerland, 2023.
[17] General Requirements for the Competence of Testing and Calibration Laboratories, International Organization for Standardization, Geneva, Switzerland, 2017.
[18] Chatfield C. Calculating, Interval Forecasts, J. Bus. Econ. Stat. 11 (1993) 121–135, https://doi.org/10.1080/07350015.1993.10509938.
[19] G. Freni, G. Mannina, Bayesian approach for uncertainty quantification in water quality modelling: the influence of prior distribution, J. Hydrol. 392 (2010) 31–39, https://doi.org/10.1016/j.jhydrol.2010.07.043.
[20] T. Liu, W. Liu, Z. Liu, H. Zhang, W. Liu, Ensemble water quality forecasting based on decomposition, sub-model selection, and adaptive interval, Environ. Res. 237 (2023) 116938, https://doi.org/10.1016/j.envres.2023.116938.
[21] X. Wan, X. Li, X. Wang, X. Yi, Y. Zhao, X. He, et al., Water quality prediction model using Gaussian process regression based on deep learning for carbon neutrality in papermaking wastewater treatment system, Environ. Res. 211 (2022) 112942, https://doi.org/10.1016/j.envres.2022.112942.
[22] S.K. Sarker, R.T. Dapkus, D.M. Byrne, A.E. Fryar, J.M. Hutchison, Quantifying temporal dynamics of *E. coli* concentration and quantitative microbial risk assessment of pathogen in a Karst Basin, Water 17 (2025) 745, https://doi.org/10.3390/w17050745.
[23] R.A. Francis, S.D. Guikema, L. Henneman, Bayesian belief networks for predicting drinking water distribution system pipe breaks, Reliab. Eng. Syst. Saf. 130 (2014) 1–11, https://doi.org/10.1016/j.ress.2014.04.024.
[24] J. Zhang, Y. Xuan, J. Lei, L. Bai, G. Zhou, Y. Mao, et al., Heavy metals prediction system in groundwater using online sensor and machine learning for water management: the case of typical industrial park, Environ. Pollut. 374 (2025) 126270, https://doi.org/10.1016/j.envpol.2025.126270.
[25] A. Cheena, K. Dost, N. Straathof, J. Wicker, T. Sarris, Don't Swim in Data: Real-time Microbial Forecasting for New Zealand Recreational Waters, 2025, https://doi.org/10.2139/ssrn.5230457.
[26] J. Waczak, A. Aker, L.O.H. Wijeratne, S. Talebi, A. Fernando, P.M.H. Dewage, et al., Characterizing water composition with an autonomous robotic team employing comprehensive in situ sensing, hyperspectral imaging, machine learning, and conformal prediction, Remote Sens. 16 (2024) 996, https://doi.org/10.3390/rs16060996.
[27] O. Iwakin, F. Moazeni, Improving urban water demand forecast using conformal prediction-based hybrid machine learning models, J Water Process Eng 58 (2024) 104721, https://doi.org/10.1016/j.jwpe.2023.104721.
[28] X. Yan, T. Zhang, W. Du, Q. Meng, X. Xu, X. Zhao, A comprehensive review of machine learning for water quality prediction over the past five years, J. Mar. Sci. Eng. 12 (2024) 159, https://doi.org/10.3390/jmse12010159.
[29] D. McElfresh, S. Khandagale, J. Valverde, V.P. C, B. Feuer, C. Hegde, et al., When Do Neural Nets Outperform Boosted Trees on Tabular Data?, 2024, https://doi.org/10.48550/arXiv.2305.02997.
[30] A. Shmuel, O. Glickman, T. Lazebnik, A Comprehensive Benchmark of Machine and Deep Learning Across Diverse Tabular Datasets, 2024, https://doi.org/10.48550/arXiv.2408.14817.
[31] E. Clements, C. Van Der Nagel, K. Crank, D. Hannoun, D. Gerrity, Review of quantitative microbial risk assessments for potable water reuse, Environ. Sci.: Water Res. Technol. 11 (2025) 542–559, https://doi.org/10.1039/D4EW00661E.
[32] B. Szelag, L. De Simoni, A. Kiczko, M. Sgroi, A.L. Eusebi, F. Fatone, Towards stormwater reuse risk management plans: methodology and catchment scale evaluation of QMRA, Sci. Total Environ. 964 (2025) 178552, https://doi.org/10.1016/j.scitotenv.2025.178552.
[33] E. Amoueyan, S. Ahmad, J.N.S. Eisenberg, B. Pecson, D. Gerrity, Quantifying pathogen risks associated with potable reuse: a risk assessment case study for Cryptosporidium, Water Res. 119 (2017) 252–266, https://doi.org/10.1016/j.watres.2017.04.048.
[34] S.T. Odonkor, J.K. Ampofo, Escherichia coli as an indicator of bacteriological quality of water: an overview, Microbiol. Res. 4 (2013) 2, https://doi.org/10.4081/mr.2013.e2.
[35] Directive (EU) 2020/2184 of the European Parliament and of the Council of 16 December 2020 on the quality of water intended for human consumption (recast), 2020.
[36] Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EECL, 2006.
[37] W.C. Lipps, E.B. Braun-Howland, T.E. Baxter, American Public Health Association, American Water Works Association, Water Environment Federation (Eds.), Standard Methods for the Examination of Water and Wastewater, 24th edition, American Public Health Association, Washington, 2023.
[38] Z. Zhang, Z. Deng, K.A. Rusch, Modeling fecal coliform Bacteria levels at Gulf Coast beaches, Water Qual Expo Health 7 (2015) 255–263, https://doi.org/10.1007/s12403-014-0145-3.
[39] H. Mohammed, I.A. Hameed, R. Seidu, Comparison of Adaptive Neuro-Fuzzy Inference System (ANFIS) and Gaussian Process for Machine Learning (GPML) Algorithms for the Prediction of Norovirus Concentration in Drinking Water Supply, in: A. Hameurlain, J. Küng, R. Wagner, S. Sakr, I. Razzak, A. Riyad (Eds.), Trans. Large-Scale Data- Knowl.-Centered Syst. XXXV 10680, Springer Berlin Heidelberg, Berlin, Heidelberg, 2017, pp. 74–95, https://doi.org/10.1007/978-3-662-56121-8_4.
[40] H. Mohammed, I.A. Hameed, R. Seidu, Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway, Sci. Total Environ. 628–629 (2018) 1178–1190, https://doi.org/10.1016/j.scitotenv.2018.02.140.
[41] A. Panidhapu, Z. Li, A. Aliashrafi, N.M. Peleato, Integration of weather conditions for predicting microbial water quality using Bayesian belief networks, Water Res. 170 (2020) 115349, https://doi.org/10.1016/j.watres.2019.115349.
[42] E. Sokolova, O. Ivarsson, A. Lillieström, N.K. Speicher, H. Rydberg, M. Bondelind, Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data, Sci. Total Environ. 802 (2022) 149798, https://doi.org/10.1016/j.scitotenv.2021.149798.
[43] L. Li, J. Qiao, G. Yu, L. Wang, H.-Y. Li, C. Liao, et al., Interpretable tree-based ensemble model for predicting beach water quality, Water Res. 211 (2022) 118078, https://doi.org/10.1016/j.watres.2022.118078.
[44] L.I. Forster, Measurement uncertainty in microbiology, J. AOAC Int. 86 (2003) 1089–1094, https://doi.org/10.1093/jaoac/86.5.1089.
[45] S. Niemela, Uncertainty of measurement of microbiological counts, Téc Lab (2017) 744–748.
[46] W.P. Oosterhuis, H. Bayat, D. Armbruster, A. Coskun, K.P. Freeman, A. Kallner, et al., The use of error and uncertainty methods in the medical laboratory, Clin. Chem. Lab. Med. 56 (2018) 209–219, https://doi.org/10.1515/cclm-2017-0341.
[47] L. Godo-Pla, J.J. Rodríguez, J. Suquet, P. Emiliano, F. Valero, M. Poch, et al., Control of primary disinfection in a drinking water treatment plant based on a fuzzy inference system, Process. Saf. Environ. Prot. 145 (2021) 63–70, https://doi.org/10.1016/j.psep.2020.07.037.
[48] S. Gavrilaş, F.-L. Burescu, B.-D. Chereji, F.-D. Munteanu, The impact of anthropogenic activities on the catchment's water quality parameters, Water 17 (2025) 1791, https://doi.org/10.3390/w17121791.
[49] C. Xu, T.K. Nguyen, E. Dixon, C. Rodriguez, P. Miller, R. Lee, et al., Can We Detect Failures Without Failure Data? Uncertainty-aware Runtime Failure Detection for Imitation Learning Policies, 2025, https://doi.org/10.48550/ARXIV.2503.08558.
[50] S. Schmidt, P.S. Heyns, Localised gear anomaly detection without historical data for reference density estimation, Mech. Syst. Signal Process. 121 (2019) 615–635, https://doi.org/10.1016/j.ymssp.2018.11.051.
[51] World Health Organization, Guidelines for Drinking-water Quality: Fourth Edition Incorporating First Addendum, 4th ed., World Health Organization, Geneva, 2017 (1st add.).
[52] L. Godo-Pla, P. Emiliano, F. Valero, M. Poch, G. Sin, H. Monclús, Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural network: uncertainty and sensitivity analysis, Process. Saf. Environ. Prot. 125 (2019) 317–327, https://doi.org/10.1016/j.psep.2019.03.017.
[53] R.G. Price, D. Wildeboer, *E. coli* as an indicator of contamination and health risk in environmental waters, *Escherichia coli* -Recent Advances on Physiology, Pathogenesis and Biotechnological Applications (2017), https://doi.org/10.5772/67330.
[54] A. Tornevi, O. Bergstedt, B. Forsberg, Precipitation effects on microbial pollution in a river: lag structures and seasonal effect modification, PloS One 9 (5) (2014) e98546, https://doi.org/10.1371/journal.pone.0098546.

[55] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, 2019, https://doi.org/10.48550/ARXIV.1907.10902.

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

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[58] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, ACM, San Francisco California USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[59] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Proces. Syst. 30 (2017) 3146–3154.

[60] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, 2017, https://doi.org/10.48550/ARXIV.1706.09516.

[61] A. Aldrees, M.F. Javed, A.T. Bakheit Taha, A. Mustafa Mohamed, M. Jasiński, M. Gono, Evolutionary and ensemble machine learning predictive models for evaluation of water quality, J. Hydrol. Reg. Stud. 46 (2023) 101331, https://doi.org/10.1016/j.ejrh.2023.101331.

[62] R.F. Barber, E.J. Candès, A. Ramdas, R.J. Tibshirani, Predictive inference with the jackknife+, Ann. Stat. 49 (2021) 486–507, https://doi.org/10.1214/20-AOS1965.

[63] V. Taquet, V. Blot, T. Morzadec, L. Lacombe, N. Brunel, MAPIE: an open-source library for distribution-free uncertainty quantification, 2022, https://doi.org/10.48550/ARXIV.2207.12274.

[64] R.L. Winkler, A decision-theoretic approach to interval estimation, J. Am. Stat. Assoc. 67 (1972) 187–191, https://doi.org/10.1080/01621459.1972.10481224.