Journal of Water Process Engineering 78 (2025) 108736

Contents lists available at ScienceDirect

Journal of Water Process Engineering

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

Hybrid machine learning model for disinfectant dosing in small-scale water treatment under data scarcity

Diego Takashi Sato [a,d], Orlando M. Oliveira Belo [c], Antonio P. Castro Junior [a,d,f], Viviane M. Gomes Pacheco [a,d,e], Cloves Gonçalves Rodrigues [e], Antonio Paulo Coimbra [b], Wesley Pacheco Calixto [a,b,d,*]

[a] Technology Research and Development Center, Federal Institute of Goias, Goiania, PC 75.250-000, Goias, Brazil
[b] Institute of Systems and Robotics, Coimbra University, PC 3030-290, Coimbra, Portugal
[c] Department of Informatics, School of Engineering, University of Minho, PC 4710-057, Braga, Portugal
[d] Electrical, Mechanical & Computer Engineering School, Federal University of Goias, Goiania, PC 74.605-010, Goias, Brazil
[e] Polytechnic and Arts School, Pontifical Catholic University of Goias, Goiania, PC 74.605-010, Goias, Brazil
[f] Directorate of Data Science, Court of Justice of the State of Goias, Goiania, PC 74.130-011, Goias, Brazil

ARTICLE INFO

Editor: Michael Short

Dataset link: https://doi.org/10.5281/zenodo.15770106

Keywords:
Artificial intelligence
Data scarcity
Disinfection by-products
Gradient boosting
Small-scale water treatment
Sustainable disinfection

ABSTRACT

Disinfection by-products, including trihalomethanes and haloacetic acids, pose persistent risks to human health and aquatic ecosystems, particularly in small-scale water treatment plants characterized by limited automation and incomplete monitoring records. This study proposes a hybrid model that integrates extreme gradient enhancement with seasonal trend decomposition, allowing incomplete time series to be partitioned into trend and seasonal components, thereby improving prediction stability and improving interpretability of variable influence. The main contribution is a method that explicitly addresses seasonal variability and data scarcity while preserving predictive accuracy under infrastructure constraints, achieving $R^2 \geq 0.90$ and RMSE values between 0.15 and 0.30. The model was validated in a real decentralized system, where it exhibited high performance even with data missing up to 30%, producing monthly reductions of approximately 450 g of trihalomethanes and 800 g of haloacetic acids, along with lower chlorine and fluoride consumption. By integrating technical, environmental, and economic dimensions, including measurable financial returns with a positive annual ROI and a short payback period, the approach provides a replicable solution for dosing control in data-limited contexts, aligned with the Sustainable Development Goal 6 of the United Nations and the advancement of responsible digital strategies in the water sector.

1. Introduction

Small and medium-sized water treatment plants (WTPs) face persistent challenges in maintaining safe drinking water standards. A major concern is the formation of disinfection by-products (DBPs), such as trihalomethanes (THMs) and haloacetic acids (HAAs), which are recognized as emerging contaminants due to their chronic toxicity and frequent occurrence in surface waters [1,2]. Exposure to these compounds has been associated with cancer and adverse reproductive outcomes, particularly in populations served by resource-limited systems [3]. Although the U.S. Environmental Protection Agency (EPA) and the World Health Organization (WHO) have established maximum contaminant levels of 80 μgL^{-1} for THM and 60 μgL^{-1} for HAA, exceedances are frequently reported in facilities lacking advanced monitoring and control, exacerbating both chemical and microbiological risks and underscoring the need for solutions that improve dosing precision while minimizing health and environmental impacts [4].

These challenges are even more pronounced in developing countries, which play a pivotal role in achieving the Sustainable Development Goal (SDG) 6, Clean Water and Sanitation [5–8]. Many WTPs in these regions still operate with outdated technologies, limited automation, and shortages of skilled personnel [9], leading to reduced efficiency and higher environmental and operational costs. Digital solutions such as artificial intelligence (AI) and the Internet of Things (IoT) have emerged as promising alternatives to improve process optimization, although their adoption remains restricted by institutional and regulatory barriers [10–12].

Developing solutions tailored to the operational realities of small-scale WTPs is therefore a strategic step toward enhancing water security

* Corresponding author at: Institute of Systems and Robotics, Coimbra University, PC 3030-290, Coimbra, Portugal.
E-mail addresses: diegotakashi@saneago.com.br (D.T. Sato), wesley.pacheco@isr.uc.pt (W.P. Calixto).

https://doi.org/10.1016/j.jwpe.2025.108736
Received 13 July 2025; Received in revised form 29 August 2025; Accepted 9 September 2025
Available online 19 September 2025

and sustainability [13]. In such facilities, limited automation often leads to disinfection failures, resulting in disease outbreaks, elevated DBP concentrations, and even the spread of antimicrobial resistance genes [6,7,14–17]. These vulnerabilities are further compounded by operational inefficiencies, and some plants report annual losses exceeding US$25,000.00 [18].

In response to these operational challenges, various artificial intelligence techniques — such as artificial neural networks (ANN), random forests (RF), and support vector machines (SVM) — have been applied to predict water quality variables and support WTP operations [19–22]. These approaches show promise in process optimization, reducing chemical waste and reducing energy consumption [10,20,23]. However, most studies remain limited to isolated predictions and do not explicitly address seasonality or real-time data gaps, which represent critical constraints in small-scale WTPs within developing regions [24].

Recent studies have extended the use of AI beyond conventional WTP operations, applying it to electrochemical processes and intelligent monitoring systems [25,26], while deep learning approaches have also emerged as promising tools for urban water management [27]. Despite these advances, adoption in decentralized treatment plants remains limited by irregular datasets, limited instrumentation, and the absence of integration into real-time operational workflows [19–22]. As a result, many efforts continue to rely on offline data or narrowly focus on single-parameter predictions.

Many existing approaches still rely on offline datasets or address only single parameters. For example, some studies predict chlorine concentration using neural networks, fuzzy systems, or predictive control [28,29], whereas others focus on the classification of the Water Quality Index (WQI) without disregarding dosing operations [30]. Fluoridation and multiparameter optimization are rarely incorporated [31,32]. Although recent advances have introduced real-time sensors and chlorine decay models [33,34], large-scale implementation remains constrained by operational and economic barriers. In developing regions, limited financial resources, scarce technical expertise, and irregular maintenance practices further impede the adoption of advanced instrumentation, ultimately lowering long-term sustainability [35].

Moreover, data quality in small-scale WTPs is often compromised by irregular sampling intervals, equipment downtime, and manual data entry, resulting in incomplete datasets [36]. Recent reviews emphasize the need for adaptable and explainable models that can maintain predictive accuracy under such imperfections [36,37], with similar challenges observed in building energy and HVAC systems operating under unbalanced and non-linear conditions [38–40]. Emerging solutions may involve IoT-based monitoring systems to enhance data continuity and support real-time optimization [10], provided that these technologies are tailored to the economic and operational constraints of small and medium-sized WTPs [37]. Together, these challenges highlight the urgent need for innovative frameworks that ensure reliable process control under seasonal variability and data scarcity, an issue that is directly addressed by the optimization model proposed in this study.

In this context, the present study hypothesizes that a hybrid model combining extreme gradient boosting (XGBoost) with seasonal trend decomposition using Loess (STL) can enhance the accuracy and sustainability of chlorine and fluoride dosing control in small-scale WTPs. The rationale for this integration is that hybrid machine learning models have consistently demonstrated superior predictive performance and improved interpretability of environmental parameters, outperforming isolated approaches when faced with incomplete datasets and complex temporal patterns. This potential has been illustrated in various applications, including the integration of XGBoost with STL to forecast rainfall [41], with LSTM to predict effluent quality [42], in hydrological models to forecast river flow [43], and in CNN-LSTM architectures to estimate pH and dissolved oxygen [44], collectively reinforcing their suitability for small-scale WTPs in developing regions.

Based on this rationale, the general objective of the present study is to develop a hybrid model that integrates XGBoost with seasonal trend decomposition using Loess (STL) to optimize chlorine and fluoride dosing in small-scale water treatment plants, ensuring predictive accuracy and operational sustainability under conditions of data scarcity and seasonal variability typical of developing regions. The specific objectives are: (i) to develop and validate the proposed hybrid model, ensuring resilience to incomplete datasets and seasonal fluctuations, (ii) to compare its performance with alternative artificial intelligence algorithms, including ANN-MLP and random forest, using accuracy metrics such as RMSE, R^2, and MAE, (iii) to reduce chlorine and fluoride dosing errors while maintaining residual concentrations within regulatory limits and minimizing the formation of disinfection by-products (THMs and HAAs), (iv) to quantify the environmental and economic benefits of the model, including chemical savings, waste reduction, financial returns, and payback period, and (v) to evaluate model interpretability through feature importance analysis, thereby enhancing transparency and supporting practical applicability in resource-limited contexts.

This study introduces a hybrid framework that integrates STL decomposition with XGBoost to ensure accurate dosing predictions under conditions of data scarcity and seasonal variability. Its novel contribution lies in demonstrating that STL-XGBoost can maintain high predictive precision ($R^2 \geq 0.90$) even with more than 30% missing data, while also providing tangible operational and environmental benefits. These include monthly reductions of approximately 450 g of trihalomethanes (THM) and 800 g of haloacetic acids (HAA), along with a decrease in chlorine and fluoride consumption and measurable financial returns. By jointly addressing technical, environmental, and economic dimensions, the proposed solution advances long-term sustainability and provides a replicable approach aligned with global sanitation goals. The following sections detail its development, validation, and broader implications.

To further substantiate these contributions, this study develops a predictive model that integrates sustainability metrics and maintains consistent performance under incomplete data conditions, achieving a coefficient of determination of $R^2 \geq 0.90$. The main text presents only the essential theoretical concepts required to understand the modeling approach, while supplementary details are provided in Section S2 of the Supporting Information. The manuscript is organized as follows: Section 2 addresses the challenges of disinfection and the role of AI, Section 3 describes the proposed model, Section 4 presents case studies and key findings, and Section 6 discusses implications for policy and sustainability.

2. Theoretical background

Optimization of chemical dosing in WTPs aligns with the principles of cleaner production by minimizing reagent waste, reducing energy use through automated decision making, and improving treated water quality. In this context, machine learning (ML) has been applied to improve operational efficiency and promote safer and more sustainable treatment practices.

2.1. Chlorination and fluoridation processes

Chlorination remains one of the most widely used disinfection methods due to its low cost and residual protection. However, chlorine-based reagents can react with natural organic matter, forming DBPs with recognized health risks [45]. Variables such as pH and organic load influence both disinfection efficacy and DBP formation [24].

Fluoridation, widely adopted to prevent dental caries, is regulated by WHO-endorsed guidelines for fluoride concentrations [46]. Overexposure may lead to fluorosis and other adverse effects [47].

Given the narrow safety margins and variable water quality, AI-based tools can support precise control of chlorine and fluoride levels. Predictive control and ML techniques, such as neural networks and decision trees, have shown potential to minimize overdosing risks and improve operational efficiency [48,49]. Additional technical details on disinfection mechanisms, fluoridation practices, and ML strategies are provided in Section S2 of the Supporting Information.

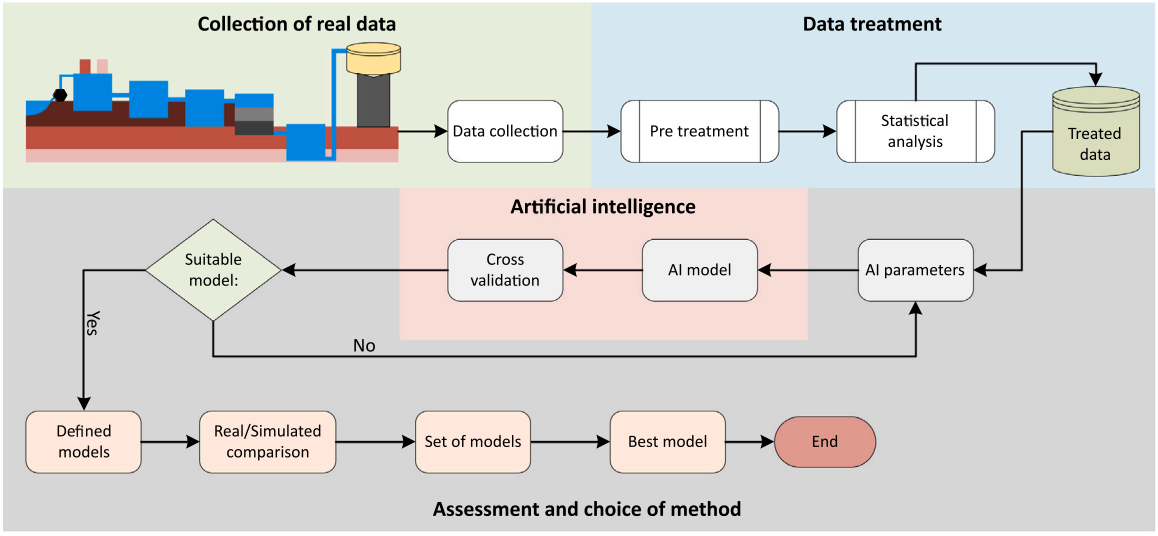

Fig. 1. Flowchart of the proposed methodology.

2.2. Machine learning techniques for water treatment plants optimization

ML techniques contribute to Cleaner Production by reducing chemical waste and improving predictive accuracy in incomplete and non-linear datasets [50]. In WTPs, models such as ANN, RF, and XGBoost have demonstrated their effectiveness in predicting optimal doses and supporting process control [51]. These models can capture non-linear dynamics and adjust parameters based on historical data [52].

Time-series gaps are common in operational datasets and can affect the reliability of the model. STL decomposition and ensemble strategies mitigate this problem by segmenting trend, seasonality, and noise components prior to recombination [53]. Integrating ML with decomposition techniques improves model performance under incomplete data, reduces energy costs [54], and supports environmental and economic gains, including reduced greenhouse gas emissions and operational savings [10]. Further details on time-series imputation, STL decomposition, and ML applications are provided in the Supporting Information.

3. Methodology

To optimize disinfectant dosing under data scarcity, this study adopts a hypothetical-deductive framework [55], integrating machine learning and time series decomposition to model and simulate the behavior of water treatment plants (WTP). The approach enables predictive analysis of operational variables and supports system optimization without physical modifications to the infrastructure. The process begins with the intake of surface water and concludes with the distribution of potable water. Fig. 1 illustrates the methodological workflow.

3.1. Real system characteristics for application

To be eligible for this study, a WTP must meet two key criteria: (i) surface water intake and (ii) a conventional treatment system that includes flocculation, sedimentation, filtration, chlorination and fluoridation. In addition, the plant must routinely measure and record relevant parameters, providing feedback to adjust chemical dosages, either manually or automatically. These characteristics constitute the necessary requirements guiding the development and efficient operation of the WTP, as illustrated in Fig. 2.

The focus on surface water sources is justified by their predominance in conventional water treatment plants with chlorination and fluoridation. Groundwater sources were excluded because they exhibit fundamentally different physicochemical and microbiological characteristics, requiring a distinct modeling framework. In general, groundwater has a lower content of turbidity and organic matter, with protein-like compounds of smaller molecular weights (< 3000 [Da]), whereas surface waters are enriched with humic and fulvic substances of higher molecular weight [56]. These compositional differences result in substantially higher DBP formation potentials in surface waters, with trihalomethane levels 50%–80% higher than in groundwater [57,58]. Hydrophobic and aromatic compounds in surface water further promote the formation of trihalomethanes and haloacetic acids [56]. Consequently, the predictive architecture developed in this study is not directly transferable to groundwater systems. Table 1 summarizes the main water quality differences relevant to the operation of the WTP.

Table 1
Comparison of environmental indicators between surface and groundwater sources in WTPs.

Indicator	Surface Water	Groundwater
Turbidity [NTU]	High (5–40)	Low (<5)
Total Dissolved Solids [mgL^{-1}]	200–800	80–300
Biochemical Oxygen Demand [mgL^{-1}]	6.2	4.5
Total Suspended Solids [mgL^{-1}]	20–100	<10
Total Coliforms [MPN/100 mL]	> 1000	<100
Formed THM [μgL^{-1}]	80–120	30–50
Formed HAA [μgL^{-1}]	40–90	10–30
Cytotoxicity Index	High	Low

3.2. Data preprocessing and application of predictive algorithms

After selecting the WTP, the data were obtained in dataframe format, including raw water flow rate V_{ab} [m^3/min], raw water turbidity T_{ab} [NTU], chlorine dosage D_{Cl} [mgL^{-1}], fluoride dosage D_F [mgL^{-1}], chlorine concentration in treated water M_{Cl} [mgL^{-1}], fluoride concentration in treated water M_F [mgL^{-1}], treated water temperature M_T [°C], treated water pH M_{pH}, and treated water flow rate V_{at} [m^3/min]. To ensure consistency across these variables, data are resampled at a uniform interval (e.g., one minute), aligned with the native frequency of the most responsive sensors (chlorine and fluoride dosing pumps). This temporal granularity provides sufficient precision for real-time monitoring while balancing computational efficiency with predictive accuracy, preventing unnecessary data inflation that would occur with higher frequency resampling.

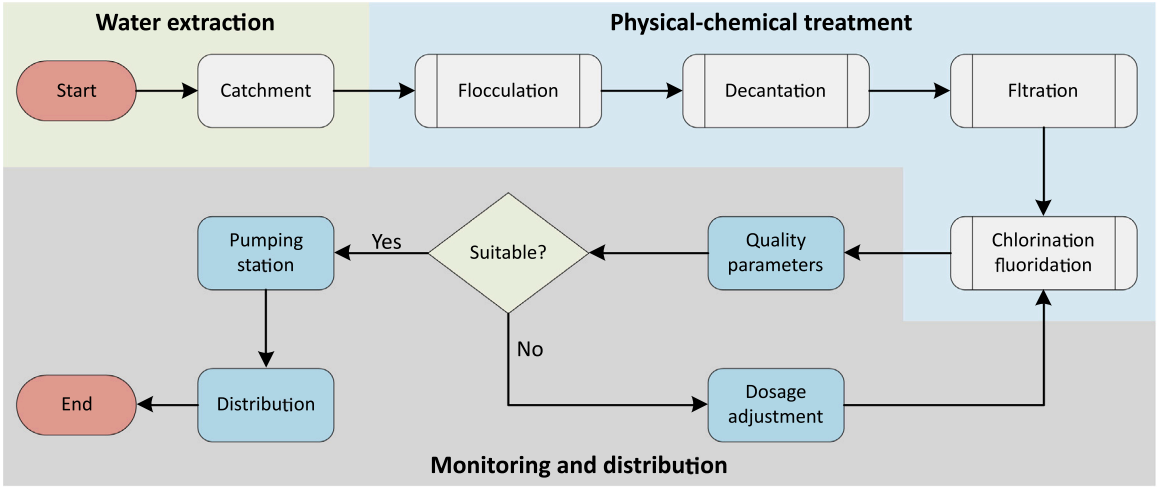

Fig. 2. Hypothetical flow of requirements from the WTP for inclusion in the study.

After temporal standardization of the series, the pre-processing focused on filling the gaps and ensuring the consistency of the measurement. In cases of missing data, polynomial linear regression functions (PSF) and STL decomposition were applied, with STL emphasized in this study due to its lower computational cost [59] and its alignment with the digital sustainability goals [60]. Theoretical foundations are provided in Section S2 of the Supporting Information. Taking into account the interdependence among the monitored variables, which directly influence both dosage decisions and final water quality (e.g. turbidity and pH affect disinfectant demand and stability), a descriptive statistical analysis was performed prior to model application. This analysis included mean (μ), standard deviation (σ), maximum (V_{max}), minimum (V_{min}), seasonal patterns, and correlations (ρ) between parameters.

3.3. Computational modeling and application of machine learning

Given the time-series nature of the data, after pre-processing and statistical analysis, the series were segmented into uninterrupted contiguous intervals of at least one month to ensure the identification of seasonal patterns in the system. The remaining data were used for training, focusing on segmentation and data independence. Validation data represented at least 20% of the dataset. If a single month did not meet this threshold, additional months were incrementally included until the criterion was satisfied. The models were constructed using Random Forest (RF), Extreme Gradient Boosting (XGBoost), and multilayer perceptron algorithms from the artificial neural network (ANN-MLP) with varying numbers of layers. The objective was to optimize the doses of chlorine and fluoride in the system. Fig. 3 illustrates the flow of processed data within the proposed methodology, in which the variable ε is the error that defines one of the stopping criteria. The justification for the STL-XGBoost integration is provided in Section S1 of the Supporting Information.

The RF algorithm constructs multiple decision trees from bootstrap samples of the training dataset, and the final prediction is obtained by averaging (in regression) or majority vote (in classification). Its simplicity reduces the number of tuning iterations, saving processing time and reducing energy demand. The ensemble structure helps mitigate overfitting, but increases memory requirements during training and inference. Mathematically, the RF model is expressed in (1), where T_b represents the decision tree bth and $\mathbf{x}$ is the input attribute vector:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}), \tag{1}$$

The XGBoost algorithm sequentially builds decision trees, with each iteration correcting the residuals of the previous. The predictions of all trees are combined and the model is updated to minimize a global loss function with regularization to prevent overfitting. This ensures efficient models with improved generalization. Its additive model is defined in (2), where $\mathbf{x}_i$ represents the attributes of observation i, f_k is the kth regression tree, and $\mathcal{F}$ is the space of all possible trees. The objective function in iteration t is given by (3), which combines the loss function $l(\cdot)$ with the regularization term $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda|w|^2$, where T is the number of leaves, w the leaf weight vector, γ the complexity cost and λ the regularization coefficient:

$$\hat{y}i = \sum k = 1^K f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \tag{2}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l!\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t), \tag{3}$$

The ANN-MLP model processes input data through interconnected layers of neurons. Each neuron performs a weighted linear combination followed by a non-linear activation function, extracting complex features in hidden layers and generating the final prediction at the output layer. Backpropagation adjusts the weights and biases iteratively to minimize the loss function, enabling non-linear approximation of variable relationships. Although effective in capturing complex patterns, ANN-MLP requires intensive computational resources and hyperparameter tuning, potentially increasing the carbon footprint of large-scale or cloud-based systems [61,62]. Its functional representation is given in (4), where $a_j^{(l)}$ is the activation of the neuron j in the layer l, $w_{ij}^{(l)}$ the weight between neurons i and j, $b_j^{(l)}$ the bias, and $\phi(\cdot)$ the activation function:

$$a_j^{(l)} = \phi!\left(\sum_{i=1}^{n} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right), \tag{4}$$

For RF and ANN-MLP, a performance evaluation was performed on the test or validation datasets, evaluating generalizability. In XGBoost, gradient boosting ensured efficient learning between iterations. Interpretability was incorporated into the modeling process through feature importance analysis: mean decrease in impurity (MDI) for RF and gain metric for XGBoost, reflecting the contribution of each variable to model predictions. These procedures were integrated into the computational modeling framework to support subsequent interpretability assessments.

3.4. Evaluation metrics, cross-validation and convergence criteria

For RF, XGBoost and ANN-MLP, the performance of the model was evaluated using the mean square error (RMSE), the mean absolute error (MAE), the coefficient of determination (R^2) and the Pearson

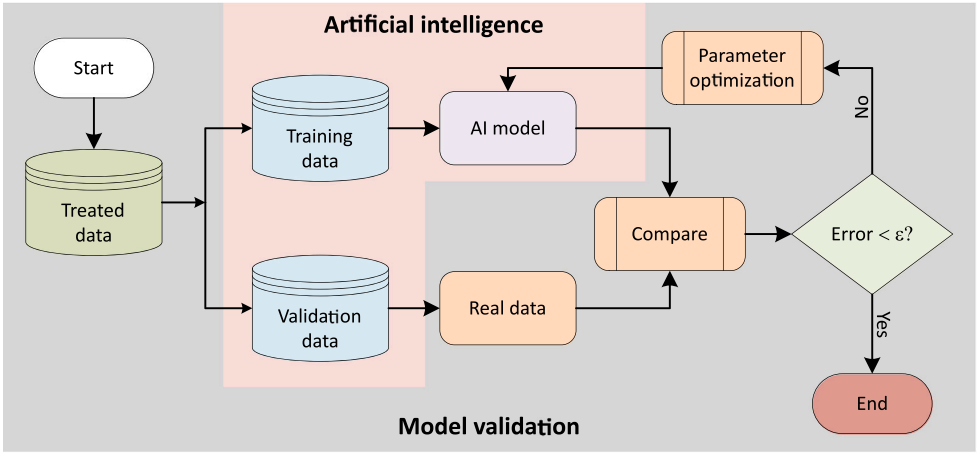

Fig. 3. Division of data in treated database for validation of the proposed model.

correlation coefficient (r). These indicators quantify prediction accuracy, explained variance, and linear association between observed and predicted values, and their selection must account for the specific characteristics of the regression problem addressed.

When selecting the evaluation method, it is necessary to consider the problem's nature, particularly in historical time-series regression. RMSE is widely used for regression problems because of its stronger penalization of larger errors, whereas MAE provides a more robust measure when outliers are present. Cross-validation was also applied to assess performance. For RF, XGBoost, and ANN-MLP, k-fold cross-validation was used, dividing the dataset into k partitions, training the model on $k-1$ partitions, and testing on the remaining one. This process was repeated k times so that each partition served once as a test set.

Convergence criteria were defined to determine when training should stop, to avoid overfitting, and to ensure computational efficiency. In RF, these criteria included the maximum number of trees (n_a) and the minimum leaf growth (c_{mf}). In XGBoost, the criteria involved n_a, the minimum sample size per leaf (t_{maf}), and the minimum gain (γ), which specifies the minimum reduction in the loss function required for a new split. In ANN-MLP, convergence was defined by the maximum number of epochs (ϵ), the minimum improvement in validation (Δ_{min}) and patience (α), indicating the number of consecutive epochs without improvement allowed before stopping training. Adjusting these criteria to the characteristics of the dataset improved the efficiency and stability of the trained models.

In machine learning, hyperparameters are defined prior to training and are not automatically adjusted by algorithms. These hyperparameters regulate aspects such as model complexity, learning rate, and regularization. Their proper definition directly influences model performance. Four optimization strategies were considered: (i) Bayesian optimization, which applies statistical techniques to model the relationship between hyperparameters and the evaluation metric, (ii) genetic algorithms, which evolve populations of hyperparameter combinations, (iii) grid search, which performs an exhaustive search over predefined hyperparameter combinations, and (iv) random search (RS), which tests random combinations to identify promising configurations.

Building on these strategies, the model calibration was carried out in two stages: (i) an initial broad exploration of the solution space using Random Search, and (ii) local refinement with grid search around the most promising configurations. For ANN-MLP, different hidden layer architectures were tested by varying the number of neurons per layer, with regularization techniques applied to prevent overfitting. Training used early stopping combined with cross-validation to balance computational cost and model stability. For RF, the effect of the number of candidate variables in each split was assessed. In XGBoost, combinations of hyperparameters related to tree depth, learning rate, and column and row sampling were explored, with regularization parameters defined to control model complexity. These procedures were guided by empirical criteria and references from the literature, ensuring the development of stable and generalizable models compatible with the operational constraints of small-scale WTPs.

3.5. Analytical modeling of DBP formation, economic performance, and solid waste generation

This methodological component establishes the analytical framework to quantify the formation of disinfection by-products (DBP), assess the economic performance of the optimized dosing system and estimate the generation of solid waste in the water treatment plant. The estimation of trihalomethanes (THMs) and haloacetic acids (HAAs) was based on the proportional model proposed by Miklos et al. [63], which describes DBP production as a direct proportional relationship between the chlorine dose applied and the resulting variation in the concentrations of the by-products. This formulation assumes pseudofirst-order kinetics and constant operational conditions such that reductions or increases in chlorine dosage almost linearly translate into changes in the formation of THM and HAA. This simplification enables a practical estimation of the potential for DBP formation in water treatment plants under realistic operational scenarios. The reduction in THM (ΔTHM) is expressed as:

$$\Delta\text{THM} = \left(\frac{\mu_{D_{Cl}} - D_{Cl}^{O}}{\mu_{D_{Cl}}} \right) \times \text{THM}_B \tag{5}$$

The reduction in HAA (ΔHAA) is given by:

$$\Delta\text{HAA} = \Delta D_{Cl} \times \alpha \times Q_M \tag{6}$$

where $\mu_{D_{Cl}}$ denotes the historical average chlorine dose, D_{Cl}^{O} the optimized dose, THM_B the baseline THM concentration, α the proportionality coefficient for HAA formation and Q_M the average monthly treated water volume [m^3/month]. These baseline values were derived from sources in the literature and operational records of the water treatment plant prior to the implementation of the model. The proportionality coefficient α represents the ratio between the applied chlorine dose and the estimated HAA formation, expressed as mg HAA/mg Cl_2. In this study, α was set at 0.09, a value supported by previous investigations that reported ranges between 0.017 and 0.106 mg HAA/mg Cl_2 on different aqueous matrices [64,65]. The adoption of this value reflects the typical conditions of tropical surface waters and was validated through consistency with the operational records of the treatment plant, thus ensuring agreement between the predicted reduction in chlorine dose and the historical levels of HAA observed in the system.

The subsequent economic analysis was structured into three categories: (i) direct savings from reduced sodium hypochlorite and sodium

fluorosilicate, (ii) operational savings from lower labor costs due to automated dosing, and (iii) indirect savings resulting from avoidance of regulatory fines for noncompliance. The financial parameters considered included the unit cost of chemicals [US$/kg], the average monthly salary of plant operators [US$/month], and the reference value of the regulatory fines [US$]. These reference values were obtained from the operational records of the local water utility, market quotations for chemical products, and values established in applicable regulatory resolutions. Such assumptions served as the basis for calculating the return on investment (ROI) and the payback period. Implementation costs comprised expenditures on sensors, equipment, and software. The ROI was defined as:

$$\text{ROI} = \left(\frac{A_E - I_C}{I_C} \right) \times 100 \tag{7}$$

where A_E is the estimated total annual savings and I_C the annual implementation cost. The payback period (months) was calculated as:

$$\text{Payback} = \frac{I_C}{A_E/12} \tag{8}$$

The generation of solid waste (sludge) was estimated from the mass of total suspended solids (TSS) removed during treatment, including organic matter, clay particles and aluminum hydroxide formed during coagulation with aluminum sulfate. TSS values were inferred from the turbidity readings and converted to the mass load [mg/L] using the Bratby method [66]. The seasonal classification was based on the turbidity thresholds: dry season (<10 NTU) and rainy season (> 80 NTU). The sludge-to-coagulant ratio was calculated using coefficients from the literature [66,67], assuming typical sludge production per kilogram of aluminum sulfate applied.

3.6. Statistical evaluation of model performance

The statistical evaluation of the ANN-MLP, RF, and XGBoost models was carried out in three stages: (i) definition of evaluation metrics (RMSE, R^2, and MAE), (ii) application of the Mann–Whitney U test to compare prediction errors with the ideal value of zero, selected for its robustness in non-normal distributions, and (iii) visual inspection of residuals. Random subsamples of size N were used in each iteration to ensure representativeness and reduce bias. The significance level was established at $\alpha = 0.05$, and p-values below this threshold were considered statistically significant.

Complementary visualization techniques included boxplots to represent error dispersion, residual plots to detect systematic deviations, and scatter plots comparing predicted and observed values against the $p = r$ line. The final selection of the model considered consistency across all evaluation metrics, statistical significance of error comparisons, absence of residual trends, and the ability to jointly predict chlorine and fluoride doses.

4. Results

This section presents the results of the proposed methodology, including the analysis of input variables, the predictive performance of the ANN-MLP, RF and XGBoost models, and a comparison between standard and optimized configurations. The precision and stability of the model were assessed using RMSE, R^2 and MAE, while the error distribution was examined to identify potential biases. The results are organized into three stages: (i) analysis and preprocessing of input data, (ii) validation of predictive models, and (iii) evaluation of technical, environmental and economic impacts. All analyses were performed in R 4.3.1 and Python 3.9, using the `caret`, `xgboost`, `randomforest`, `ggplot2`, `reshape2`, `scales`, and `metrics` packages in R, and `pandas`, `numpy`, and `matplotlib` in Python. All datasets and codes used in this study are openly available in the repository described in Sato & Calixto [68].

4.1. Data source, preprocessing and variable selection

Data were collected from the Brazabrantes Water Treatment Plant (WTP), located in the Metropolitan Region of Goiânia, Goiás, Brazil, and operated by Saneamento de Goiás S.A. (Saneago), through a request under the Access to Information Law [69]. Brazabrantes lies within the Paraná Basin and is part of the Meia Ponte River Basin Committee, covering approximately 1% of its area. The municipality has a urban water coverage of 93.55% and a metering rate of 97.67%, both above the national averages. The Brazabrantes Water Treatment Plant abstracts water from the Ribeirao Cachoeira, a tributary of the Meia Ponte River. Although the municipal concession allows for a maximum withdrawal of 20 Ls^{-1}, the operational flow is restricted to 14 Ls^{-1} due to infrastructure constraints. The intake point is located at the coordinates $16°26'57.3''S$, $49°22'15.9''W$, approximately 450 m downstream of the WTP, which is located at a higher elevation near the urban perimeter (see Fig. 4).

The treatment process at the Brazabrantes Water Treatment Plant includes surface water abstraction, turbidity monitoring, proportional coagulant dosing, flocculation–sedimentation in a clarification tank, filtration, and final disinfection with chlorination and fluoridation in contact chambers. Operational control is ensured by monitoring the temperature, pH, and residual disinfectants, with treated water subsequently pumped to an elevated reservoir for distribution. Sedimentation sludge undergoes thickening, mechanical dewatering, and stabilization before reuse as a soil substrate in reforestation programs with native, non-fruit-bearing species, in accordance with public health regulations. An overview of these stages is provided in Fig. 5, while further operational details are described in the Supporting Information.

Chlorination and fluoridation values at this WTP comply with international guidelines for potable water [70,71]. For systems without natural fluoride sources, artificial fluoridation is maintained between 0.5 and $1.0, \text{mg/L}$ (maximum $1.5, \text{mg/L}$), while disinfection with chlorine derivatives (e.g., sodium hypochlorite) requires residual concentrations of 0.2–$1.0, \text{mg/L}$ to prevent microbial regrowth in the distribution network. Ferric chloride is applied for flocculation-decantation and sodium fluorosilicate is applied for fluoridation. The choice of Brazabrantes WTP reflects its systematic monitoring of parameters, the availability of historical operational data, and its representativeness of approximately 80% of the plants operated by *Saneago S.A.*.

In Fig. 5, orange rectangles with a vertical separator indicate where the variables are measured and recorded in the system. Fig. 6 presents some equipment, environments and process indicators observed in the WTP of Brazabrantes. The operation begins with the intake of surface water, where V_{ab} is measured and turbidity T_{ab} is corrected with coagulant (Fig. 6(a–b)). The water then flows through the flocculation-decantation tank (Fig. 6(c)), followed by filtration units with activated carbon to remove excess particles and correct taste and odor (Fig. 6(d–e)). It then enters the contact tank, where chlorine and fluoride are dosed and measurements such as M_T and M_{pH} are taken (Fig. 6(f)). If the parameters are within acceptable ranges, the treated water is pumped to an elevated reservoir and distributed to the population. Monitoring data from these stages are stored locally and later transmitted to a remote repository.

All recorded information from the WTP in Brazabrantes, covering the period from October 28, 2022, to July 23, 2024. The dataset includes: (i) raw water turbidity T_{ab}, (ii) levels of the supported reservoir N_{RA} and elevated reservoir N_{RE}, (iii) electrical information on current, voltage, power and frequency of the coagulant dosing pump, (iv) coagulant dosage D_S, (v) reservoir level in the subdistrict of Deuslandia, Goias, (vi) electrical information on current, voltage, power and frequency of the fluoride dosing pump, (vii) fluoride dosage D_F, (viii) raw water flow rate V_{ab}, (ix) cumulative historical volume of raw water, (x) treated water flow rate V_{at}, (xi) historical volume of treated water, (xii) electrical information on current, voltage, power and frequency

Fig. 4. Georeferenced map of the water intake at the Brazabrantes WTP, Goiás, Brazil.

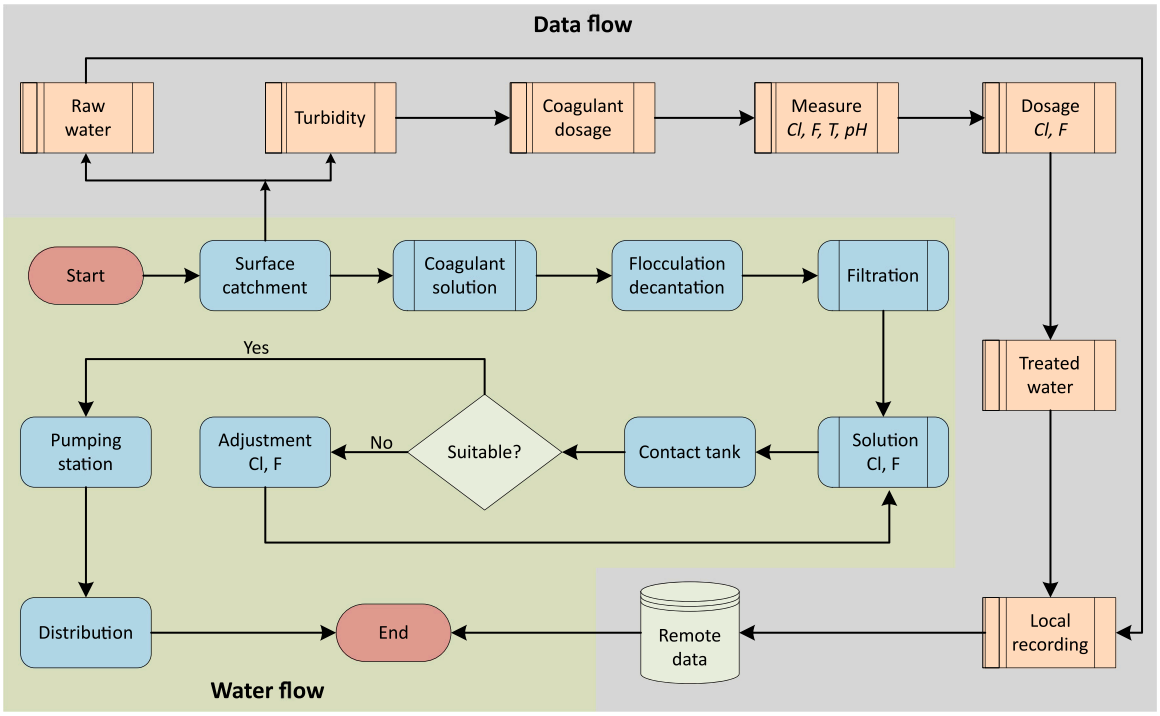

Fig. 5. Brazabrantes water treatment system, Goias/Brazil.

of the chlorine dosing pump, (xiii) chlorine dosage D_{Cl} and (xiv) measurements of pH M_{pH}, fluoride M_F, chlorine M_{Cl} and treated water temperature M_T. Based on these records, the variables most correlated with chlorine and fluoride doses were selected for further analysis, as summarized in Table 2.

In Table 2, several correlations stand out as particularly relevant for model performance. For example, V_{ab} and V_{at} exhibit a strong positive correlation ($r \approx 0.92$), reflecting the operational consistency between the flows of raw and treated water, which directly influences chemical dosing strategies. Similarly, the chlorine dosage (D_{Cl}) shows a moderate to strong correlation with both flow rates ($r \approx 0.82$–0.89), indicating that variations in hydraulic load substantially affect the amount of disinfectant applied. In contrast, the negative correlation between M_F and D_F ($r \approx -0.18$) suggests compensatory adjustments during operation, where higher measured fluoride levels tend to be followed by lower doses to maintain target concentrations. These patterns clarify how plant dynamics shape the relationships between input and output variables and support the inclusion of these parameters in predictive models.

To ensure data reliability, outliers were removed based on operational constraints and technician reports. The adopted criteria were: $2 \leq M_{pH} < 8$, $0 \leq D_{Cl} < 5$ mg/L, $0 \leq D_F < 3$ mg/L, $10\ °C \leq M_T < 50\ °C$, $0 \leq V_{ab}, V_{at} < 1\ m^3/min$, $0 \leq M_{Cl} < 5$ mg/L, and $0 \leq M_F < 10$ mg/L. After data cleaning, descriptive statistics of the filtered data set, including mean (μ), standard deviation (σ), maximum (V_{max}), and minimum (V_{min})—are presented in Table 3. Parameters

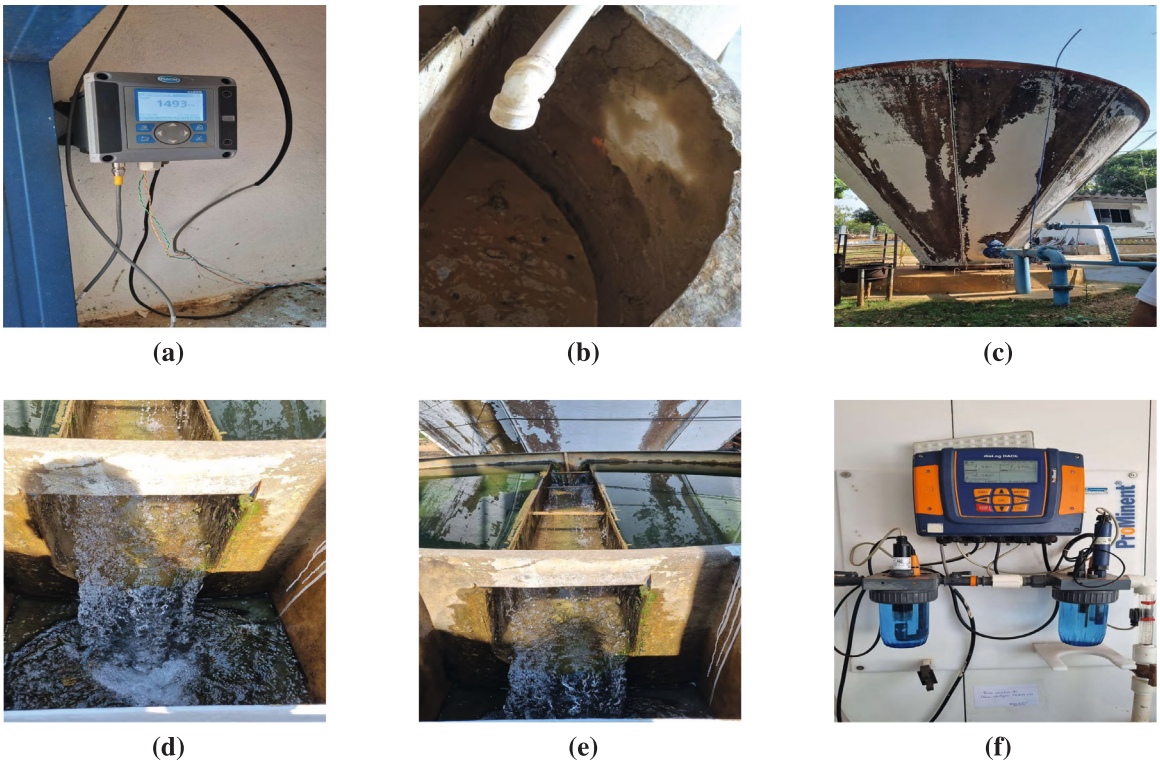

Fig. 6. Equipment, environments and process: (a) ferric chloride solubilization tank in batch, (b) turbidity meter in NTU, (c) floc-decanter, (d) floating/coarse solids retention screen at the outlet of the filter-decanter, before entering the upflow filter, (e) surface water collection chute in the filter-decanter and (f) automated chlorine, fluoride, M_T and M_{pH} meter.

Table 2
Correlation between some observed variables and the dosage of chlorine and fluoride.

	M_{Cl}	M_F	M_{pH}	M_T	V_{at}	V_{ab}	D_{Cl}	D_F
M_{Cl}	–							
M_F	-1.00×10^{-1}	–						
M_{pH}	5.00×10^{-2}	2.20×10^{-1}	–					
M_T	-2.00×10^{-2}	2.10×10^{-1}	-1.60×10^{-1}	–				
V_{at}	4.80×10^{-1}	-2.40×10^{-1}	3.00×10^{-2}	0	–			
V_{ab}	4.60×10^{-1}	-2.40×10^{-1}	3.00×10^{-2}	0	9.20×10^{-1}	–		
D_{Cl}	4.30×10^{-1}	-1.80×10^{-1}	4.00×10^{-2}	-2.00×10^{-2}	8.90×10^{-1}	8.20×10^{-1}	–	
D_F	4.30×10^{-1}	-1.80×10^{-1}	4.00×10^{-2}	-2.00×10^{-2}	8.90×10^{-1}	8.20×10^{-1}	9.80×10^{-1}	–

Table 3
Descriptive statistics of the filtered dataset.

Parameter	μ	σ	V_{max}	V_{min}
D_{Cl}	1.45	1.61	4.82	0.00
M_{Cl}	6.20×10^{-1}	4.93×10^{-1}	2.83×10^{1}	1.00×10^{-2}
M_{pH}	6.62	4.04×10^{-1}	7.93	2.67
M_T	2.58×10^{1}	2.64	4.51×10^{1}	1.72×10^{1}
M_F	9.07×10^{-1}	3.81×10^{-1}	9.14	0.00
D_F	8.60×10^{-1}	9.50×10^{-1}	2.82	0.00
V_{ab}	4.00×10^{-1}	3.92×10^{-1}	1.00	0.00
V_{at}	3.65×10^{-1}	3.84×10^{-1}	1.00	0.00
T_{ab}	1.31×10^{2}	1.55×10^{2}	3.00×10^{3}	2.00

such as D_{Cl} (μ = 1.45 mg/L) and M_{pH} (μ = 6.62) fall within typical operational ranges.

To assess the reliability of the gap filling methods, 20% of the known values were randomly removed. The Polynomial Spline Regression Functions (PSF) and STL methods were then applied to the remaining 80%, and their performance was evaluated using RMSE, R^2 and MAE. The results summarized in Table 4 consistently demonstrate the superiority of STL over PSF, with RMSE reductions of up to 18% for key variables such as D_{Cl} (from 1.04×10^{-1} to 9.11×10^{-2}). This improvement, combined with stable performance across all parameters evaluated, supports the adoption of STL as the reference approach to handling missing data in this study, thus validating a central methodological choice for the construction of subsequent predictive models.

These results confirm the reliability of STL for handling missing data in real-time monitoring. The variables listed in Table 2 were used for model construction. The original dataset comprised 599,386 historical records from the analyzed period, totaling approximately 81 MB of structured data. These records, exported from a relational database in `*.csv` format, included variables used to calculate the Water Quality

Table 4
Evaluation metrics for PSF and STL methods used in gap filling.

Parameter	RMSE PSF	RMSE STL	R^2 PSF	R^2 STL	MAE PSF	MAE STL
D_{Cl}	1.04×10^{-1}	9.11×10^{-2}	9.96×10^{-1}	9.97×10^{-1}	8.08×10^{-3}	5.57×10^{-3}
M_{Cl}	2.27×10^{-2}	1.84×10^{-2}	9.98×10^{-1}	9.99×10^{-1}	1.34×10^{-3}	1.16×10^{-3}
M_{pH}	1.97×10^{-2}	1.47×10^{-2}	9.98×10^{-1}	9.99×10^{-1}	1.47×10^{-3}	1.09×10^{-3}
M_T	7.34×10^{-2}	6.54×10^{-2}	9.99×10^{-1}	9.99×10^{-1}	4.75×10^{-3}	3.39×10^{-3}
M_F	2.34×10^{-2}	2.02×10^{-2}	9.96×10^{-1}	9.97×10^{-1}	2.71×10^{-3}	2.12×10^{-3}
D_F	6.11×10^{-2}	5.35×10^{-2}	9.96×10^{-1}	9.97×10^{-1}	4.78×10^{-3}	3.29×10^{-3}
V_{ab}	2.74×10^{-2}	2.27×10^{-2}	9.95×10^{-1}	9.97×10^{-1}	4.90×10^{-3}	3.94×10^{-3}
V_{at}	2.69×10^{-2}	2.29×10^{-2}	9.95×10^{-1}	9.96×10^{-1}	5.31×10^{-3}	3.68×10^{-3}
T_{ab}	6.86×10^{-5}	5.07×10^{-5}	9.99×10^{-1}	9.99×10^{-1}	6.26×10^{-8}	5.32×10^{-8}

Table 5
Performance of standard × optimized models based on RMSE, R^2, and MAE.

Model	Standard						Optimized					
	Chlorination			Fluoridation			Chlorination			Fluoridation		
	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE
ANN-MLP	0.531	0.898	0.205	0.318	0.896	0.139	0.139	0.994	0.041	0.129	0.983	0.035
	0.526	0.897	0.816	0.294	0.909	0.128	0.302	0.960	0.102	0.089	0.992	0.018
							0.301	0.961	0.104	0.085	0.993	0.019
RF	0.085	0.997	0.012	0.051	0.997	0.007	0.074	0.998	0.008	0.054	0.997	0.006
	0.088	0.996	0.011	0.052	0.996	0.006	0.074	0.998	0.008	0.054	0.997	0.005
XGBoost	0.278	0.970	0.112	0.172	0.966	0.073	0.089	0.997	0.016	0.083	0.998	0.014
	0.281	0.969	0.115	0.172	0.967	0.072	0.056	0.997	0.009	0.054	0.997	0.009

Values are rounded to three decimal places. Boldface indicates the best result in each row.

Index (WQI). Data were collected every 15 min until May 11, 2023, after which the sampling rate increased to one minute. To ensure uniform input of the model and enable real-time analysis, all data were resampled to one-minute intervals. Following the application of gap-filling techniques, the dataset expanded to 909,250 records, covering the period from October 31, 2022, to July 23, 2024. STL consistently outperformed PSF.

4.2. Predictive model implementation and validation

The consolidated dataset was used to evaluate model performance based on RMSE (emphasizing larger deviations), R^2 (proportion of variance explained) and MAE (average error magnitude). A statistical analysis of the WTP samples was performed to compare the parameters with internationally recommended limits [72]. Of the 909,250 records, 908,054 met the chlorination criterion (≈ 99%). Regarding fluoridation, 866,621 records were within the recommended range (≈ 95%). These results indicate that the operation of the analyzed WTP complies with international guidelines for hypochlorite and artificial fluoridation disinfection, thus supporting reliable modeling of operational behavior.

For model development and internal validation, data were partitioned so that the period from January 1 to March 30, 2023 (89 days), comprising 131,041 records (≈ 25% of the dataset), was reserved for testing and validation. The implementations were carried out in R using the `caret` package. A 10-fold cross-validation procedure was adopted, with the random seed fixed to ensure reproducibility. Three methods were applied: ANN-MLP, RF, and XGBoost. Two simulation scenarios were generated: one with default hyperparameters and another with optimized hyperparameters.

After optimization, the tree-based models (RF and XGBoost) consistently outperformed ANN-MLP. For chlorine dosage prediction, RF achieved the best performance (RMSE = 0.074, R^2 = 0.998) with candidate variables N_{rc} = 4 per split. For the prediction of fluoride dosage, XGBoost produced the most accurate results (RMSE = 0.054, R^2 = 0.997) using a configuration with learning rate η = 0.2 and maximum tree depth D_M = 6. Performance comparisons between default and optimized configurations, evaluated by RMSE, R^2, and MAE, are summarized in Table 5, which presents the complete analysis and results for all algorithms.

In addition to the superior performance of the ensemble models, the analysis in Table 5 highlights important patterns regarding optimization for each target variable. Although RF achieved operational excellence with N_{rc} = 4 for chlorination, it exhibited a relevant trade-off for fluoridation. While N_{rc} = 2 maximized R^2 = 0.997, N_{rc} = 4 was more effective in minimizing MAE = 0.005. This subtle yet critical difference underscores that the selection of the optimal hyperparameter depends on the performance metric prioritized for the intended application.

XGBoost, in turn, demonstrated greater resilience. Its optimized configuration not only achieved the best overall performance for fluoridation but also delivered highly competitive results for chlorination, with RMSE = 0.056 and R^2 = 0.997, values comparable to those of RF. This outcome suggests a slightly superior generalizability when dealing with variations in the quality of raw water. Although ANN-MLP remained third in the overall comparison, the magnitude of its post-optimization improvement was notable, with reductions in RMSE of up to 73% for chlorination. However, this performance gain came at the cost of substantially higher computational complexity and relative instability across runs compared with the more stable tree-based models. Table 6 details the hyperparameter configurations underlying these findings.

The optimization process, which combined random sampling with grid search, yielded the optimal architecture for each model, as presented in Table 6. For ANN-MLP, superior performance was achieved with more complex architectures (e.g. [11, 4, 8] for chlorination) coupled with minimal weight decay values ($W_D = 10^{-5}$ or 10^{-6}), confirming the need for sufficient regularized modeling capacity to prevent overfitting. Convergence was efficiently ensured using `early stopping` (`patience` = 20, `max_epochs` = 500), with these parameters validated under the constraints of the target hardware (Raspberry Pi 4B).

For the ensemble models, optimization identified more compact and efficient configurations. The optimal number of candidate variables per split (N_{rc}) for RF was 4 for chlorination and 5 for fluoridation, reflecting the distinct nature of the two prediction tasks. In XGBoost, a lower learning rate (η = 0.2) combined with a moderate maximum depth ($6 \leq D_M \leq 8$) was required to achieve the best performance, allowing more careful and generalizable learning. In general, the performance gains documented in Table 5 are directly attributable to the

Table 6
Standard and optimized hyperparameters used in internal validation.

Model	Hyperparameters standard				Hyperparameters optimized			
	S_z/S_1	S_2	S_3	W_D	S_1	S_2	S_3	W_D
ANN-MLP	3.00	–	–	0.000	11.00	4.00	8.00	0.000010
	5.00	–	–	0.000	10.00	11.00	2.00	0.000001
	–	–	–	0.000	10.00	12.00	2.00	0.000001
RF	2.00	–	–	–	4.00	–	–	–
	4.00	–	–	–	5.00	–	–	–
XGBoost	0.400	3.00	0.800	0.750	0.200	6.00	0.800	0.750
	0.400	3.00	0.800	1.000	0.200	8.00	0.800	0.750

Numeric values correspond to the tested configurations, where "–" indicates a non-applicable parameter.

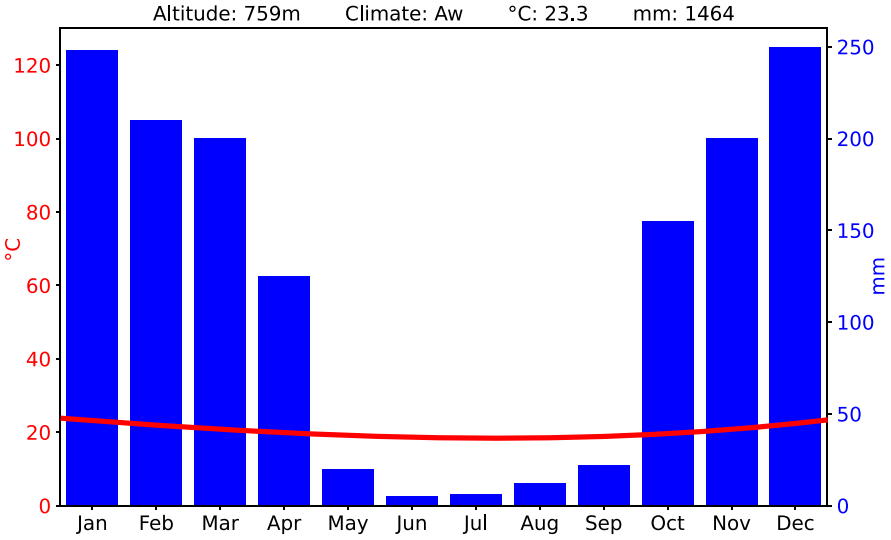

Fig. 7. Climatic context of the study region based on monthly precipitation and temperature, highlighting rainy and dry seasons.

hyperparameter configurations detailed in Table 6. Optimization was not merely incremental, but essential to fully exploit the potential of each algorithm, resulting in robust and resilient models capable of delivering reliable predictions for operational control of chlorine and fluoride dosing, with tangible impacts on reducing variability and reagent consumption.

4.3. External validation under seasonal conditions

The historical analysis of precipitation and temperature provides the climatic context for external validation. Fig. 7 shows that the region has a limited annual variation in mean temperature but two well-defined seasons: (i) a rainy period from October to March and (ii) a dry period from April to September, with minimum precipitation in July. Within this framework, the effectiveness of the optimized models was assessed through comparative performance analysis, where the predictions of ANN-MLP, RF and XGBoost were tested against two independent datasets, A22 and A24. Dataset A22, collected between October and December 2022 during the rainy season, reflects higher turbidity fluctuations and higher variability in system performance.

Dataset A24 corresponds to April 2024, at the beginning of the dry season, when the system operation was more stable following routine equipment adjustments. The choice of these contrasting seasonal contexts enables the evaluation of the generalization of the model under distinct hydrological and operational conditions. Rainfall patterns directly influence treatment performance: (i) during the wet season, increased surface runoff increases organic matter loads and turbidity, requiring dosage adjustments, whereas (ii) in the dry season, restrictions on water abstraction introduce different operational challenges. The analysis includes the calculation of the following performance metrics: (i) RMSE, (ii) MAE, and (iii) R^2. In addition, the Mann–Whitney U statistical test is applied using subsamples of size $N = 1000$. Table 7 reports the values of RMSE, R^2, MAE, and values-p for each model and the parameter evaluated, highlighting the individual performance of each approach.

As shown in Table 7, the optimized RF and XGBoost models achieved the best performance for chlorine and fluoride predictions, respectively. Both models outperformed their standard counterparts in all evaluation metrics, with significant reductions in RMSE and MAE and increases in R^2. For fluoride, the optimized XGBoost reached RMSE = 0.15 and $R^2 = 0.95$, indicating a high degree of precision and explanatory power. All p-values were below 0.05, confirming the statistical significance of the observed improvements. Although the optimized ANN-MLP showed progress, its results remained inferior to those of the other optimized models. To complement these quantitative results, an analysis of the error distribution was performed to assess consistency and detect potential outliers.

To evaluate the error distribution and identify potential outliers, Fig. 8 was created. The central line in each boxplot represents the median of the errors, while the box edges correspond to the first and third quartiles, defining the interquartile range. The boxplots illustrate the error dispersion for each model, complementing the quantitative analysis of the performance metrics. The optimized models are observed to exhibit lower dispersion, suggesting more accurate predictions. For chlorine, optimized models display smaller boxes and shorter whiskers, indicating a more concentrated error distribution. The optimized RF model stands out because it has a median close to zero and a reduced number of outliers, implying greater consistency in predictions. For fluoride, the optimized XGBoost shows a lower dispersion and values close to zero, further highlighting its precision and stability.

Visual analyses of residuals were performed to identify potential patterns or systematic deviations from actual values. Fig. 9 displays, on

Table 7
Evaluation metrics results for predictive models of chlorine and fluoride.

Parameter	Model	RMSE	R^2	MAE	p-valor
chlorine	ANN-MLP	9.30×10^{-1}	4.40×10^{-1}	6.80×10^{-1}	1.11×10^{-8}
	RF	9.90×10^{-1}	3.80×10^{-1}	6.80×10^{-1}	5.33×10^{-29}
	XGBoost	9.60×10^{-1}	4.10×10^{-1}	6.50×10^{-1}	3.40×10^{-28}
	ANN-MLP Optimized	6.60×10^{-1}	7.20×10^{-1}	5.60×10^{-1}	9.99×10^{-5}
	RF Optimized	4.00×10^{-1}	9.00×10^{-1}	3.60×10^{-1}	1.07×10^{-10}
	XGBoost Optimized	7.90×10^{-1}	6.00×10^{-1}	7.00×10^{-1}	1.11×10^{-8}
fluoride	ANN-MLP	5.10×10^{-1}	4.50×10^{-1}	4.40×10^{-1}	3.17×10^{-225}
	RF	3.00×10^{-1}	8.10×10^{-1}	2.10×10^{-1}	3.38×10^{-232}
	XGBoost	2.50×10^{-1}	8.70×10^{-1}	1.60×10^{-1}	2.37×10^{-163}
	ANN-MLP Optimized	1.80×10^{-1}	9.30×10^{-1}	1.30×10^{-1}	7.17×10^{-71}
	RF Optimized	1.90×10^{-1}	9.20×10^{-1}	1.30×10^{-1}	3.70×10^{-207}
	XGBoost Optimized	1.50×10^{-1}	9.50×10^{-1}	9.00×10^{-2}	7.32×10^{-18}

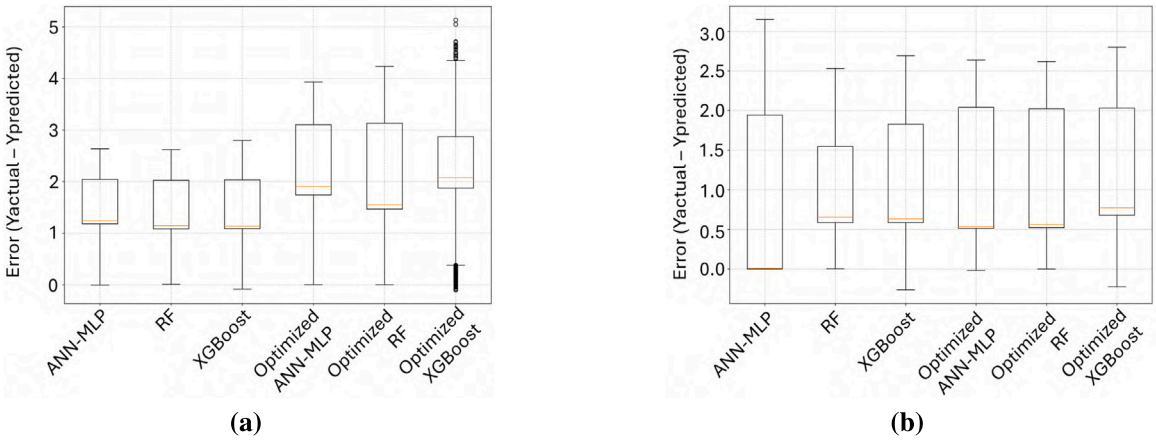

Fig. 8. Error distribution of predictive models for: (a) chlorine and (b) fluoride.

the x-axis, the sequence of predictions in the observations, while the y-axis represents the differences between the actual and predicted values generated by the models. The blue points indicate individual errors for each observation, and the horizontal red line denotes the zero-error reference, where the predictions perfectly align with the actual values.

The optimized XGBoost is observed to exhibit a lower dispersion around the zero error line, indicating a higher prediction accuracy. The smaller dispersion and the reduced occurrence of points far from the red line indicate fewer extreme errors. The optimized ANN-MLP shows reduced error variability compared to its non-optimized version, however, its dispersion remains higher than that of the XGBoost, suggesting lower consistency in predictions.

This visual analysis of residuals indicates that the optimized models exhibit a higher concentration of errors close to zero, reinforcing their accuracy. The absence of systematic patterns in the residuals suggests that the errors are randomly distributed, indicating a lack of bias and proper model behavior. Furthermore, the reduced occurrence of outliers in the XGBoost highlights greater stability and improved generalization capacity, which supports the absence of overfitting in the predictions. Fig. 10 illustrates the alignment between predicted and actual values, with the red line representing the ideal relationship $p = r$.

For the ANN-MLP, a considerable dispersion of points is observed around the ideal line, with tendencies to underestimate or overestimate values across different intervals, suggesting a lower prediction accuracy. In the optimized ANN-MLP, despite improvements, non-linear patterns and a higher concentration of points distant from the red line remain noticeable, indicating limitations in accurately capturing the actual values. The RF model demonstrates a distribution closer to the ideal line with reduced point dispersion. However, the presence of some distant points indicates difficulties in predicting certain values with high precision. In the optimized RF model, a more uniform and concentrated distribution is observed around the ideal line, with significant reductions in deviations and closer alignment of the points with the red line. These results confirm that optimization improved the generalization and accuracy of the model.

The residual distribution confirms that the optimized models are neither over- or under-fitted. Although all models improved after optimization, the RF model yielded the best results for chlorine, and XGBoost outperformed others for fluoride. However, for chlorine, the optimized RF model achieved the best performance, with RMSE = 0.40, R^2 = 0.91, MAE = 0.36 and p-value = 1.07×10^{-10}. For fluoride, optimized XGBoost delivered the best results, with RMSE = 0.15, R^2 = 0.95, MAE = 0.09 and p-value = 7.32×10^{-18}.

Among all the configurations evaluated, the optimized XGBoost achieved the best overall performance in predicting both chlorine and fluoride. The criteria supporting this choice include: (i) the lowest RMSE values for both parameters, (ii) a high R^2, indicating strong explanatory power for data variance, (iii) residuals with no discernible trends, reinforcing the absence of overfitting and (iv) consistent predictions across different periods and sub-samples. These findings were evaluated through external validation.

Residual dispersion observed in Fig. 9 (residual plots) and Fig. 10 (scatter plots of observed × predicted values) can be attributed to mechanistic factors associated with plant operation, water chemistry, and measurement limitations. In addition, the error distribution in Fig. 8 confirms the magnitude and variability of these deviations. Abrupt seasonal increases in turbidity, particularly during rainfall events, increase coagulant demand and indirectly affect chlorine and fluoride dosages, generating discrepancies not fully captured by the models. Variations in pH and temperature alter chlorine speciation and reaction kinetics, leading to localized prediction errors. Dosing pump inertia introduces delays in achieving target concentrations, while sensor noise, downtime, and manual interventions further contribute to measurement inconsistencies. Finally, changes in reservoir levels modify hydraulic

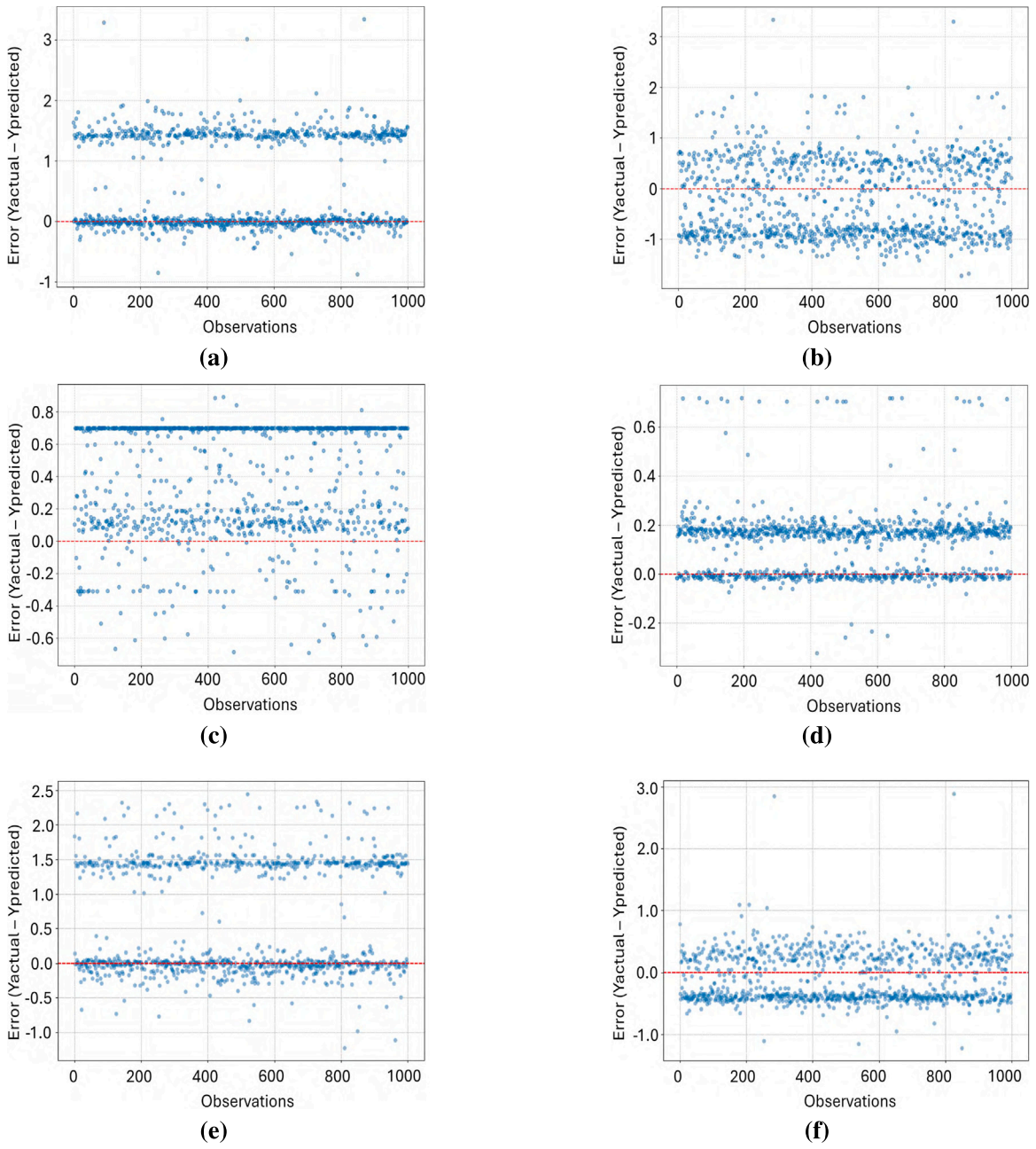

Fig. 9. Residuals of predictive models across observations: (a) XGBoost without optimization for chlorine, (b) optimized XGBoost for chlorine, (c) ANN-MLP without optimization for fluoride, (d) optimized ANN-MLP for fluoride, (e) RF without optimization for chlorine and (f) optimized RF for chlorine.

regimes, influencing the dilution and the disinfectant contact time. Together, these mechanistic factors explain the residuals observed, while the overall precision of the optimized models remained high.

For the RF model, Figs. 9(e) and 9(f) show residuals with lower dispersion compared to ANN-MLP and with consistency levels similar to XGBoost. After optimization, the residuals become more concentrated around the zero reference line, with a marked reduction in extreme values. These findings complement the quantitative metrics (Table 7), confirming the predictive capacity of RF compared to ANN-MLP and its performance comparable to XGBoost in chlorine dosage prediction.

4.3.1. External validation and comparative performance

To evaluate the performance of the model under different operating conditions, two external datasets were used: A22, collected from October 31, 2022, to December 31, 2022 (62 days, 129,601 records), and A24, collected from April 1, 2024, to April 12, 2024 (12 days, 25,084 records). Initial validations were conducted using models without hyperparameter optimization. As presented in Table 8, RF yielded the best chlorine predictions in both samples. For fluoride, the best results were obtained with ANN-MLP in A22 and XGBoost in A24. Fig. 11 illustrates the behavior of chlorine and fluoride doses over a 24-hour period in A22, comparing the measured values of the system with those predicted by the models.

The same external datasets (A22 and A24) were also used to validate the models after hyperparameter optimization. Table 9 shows that RF achieved the lowest RMSE and MAE values for chlorine in both samples, while ANN-MLP obtained the highest R^2. For fluoride, ANN-MLP outperformed the other models in A22, while XGBoost showed superior accuracy in all metrics in A24. Fig. 12 illustrates the behavior of chlorine and fluoride doses over a 24-hour period in the A24 sample, comparing the actual values of the system with the predictions generated by the proposed models.

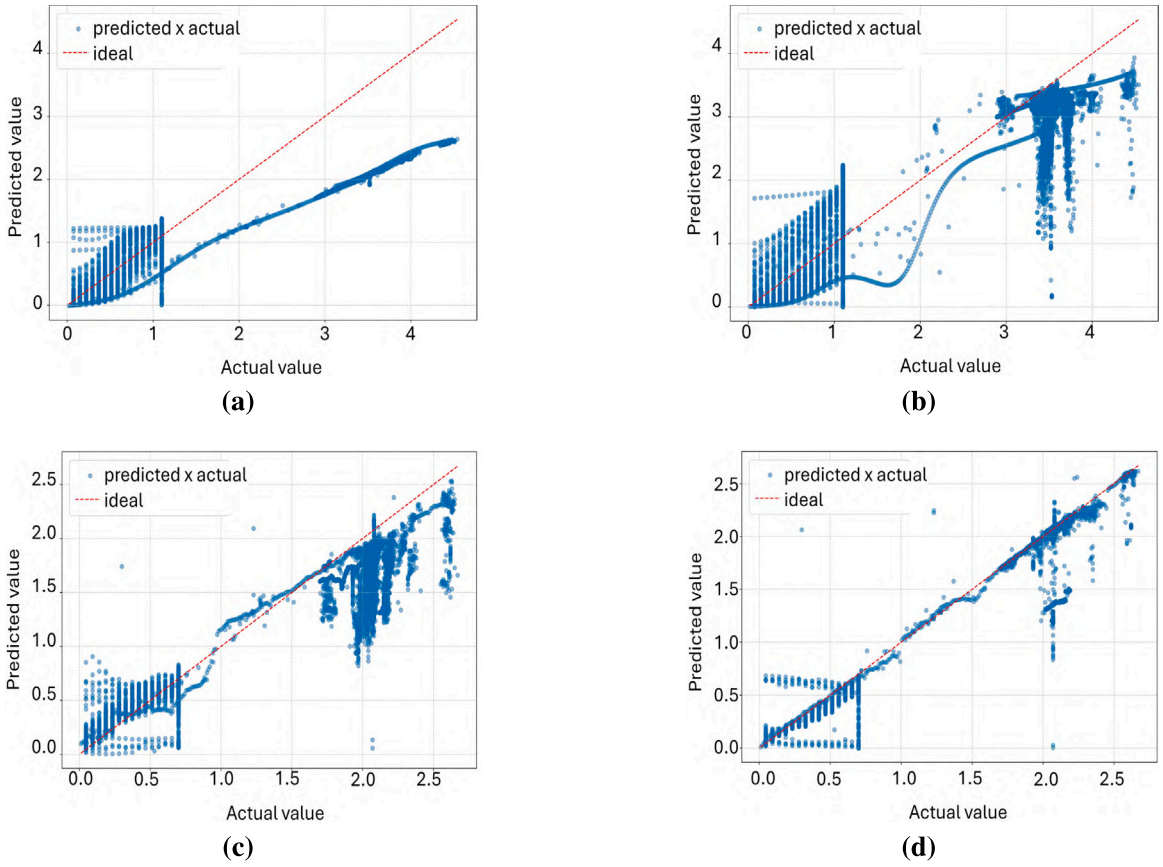

Fig. 10. Scatter plot of actual × predicted values: (a) ANN-MLP without optimization for chlorine, (b) optimized ANN-MLP for chlorine, (c) RF without optimization for fluoride and (d) optimized RF for fluoride.

Table 8
External validation comparing model outputs with actual values collected in the system.

Sample	Models	Chlorination			Fluoridation		
		RMSE	R^2	MAE	RMSE	R^2	MAE
2022	ANN-MLP	4.21×10^{-1}	9.64×10^{-1}	$\mathbf{2.31 \times 10^{-1}}$	$\mathbf{2.07 \times 10^{-1}}$	$\mathbf{9.60 \times 10^{-1}}$	$\mathbf{1.04 \times 10^{-1}}$
2022	RF	$\mathbf{3.10 \times 10^{-1}}$	$\mathbf{9.87 \times 10^{-1}}$	2.47×10^{-1}	3.37×10^{-1}	9.54×10^{-1}	2.40×10^{-1}
2022	XGBoost	8.11×10^{-1}	8.14×10^{-1}	5.90×10^{-1}	2.51×10^{-1}	9.54×10^{-1}	1.61×10^{-1}
2024	ANN-MLP	3.25×10^{-1}	$\mathbf{8.54 \times 10^{-1}}$	1.66×10^{-1}	2.72×10^{-1}	6.50×10^{-1}	1.35×10^{-1}
2024	RF	$\mathbf{2.68 \times 10^{-1}}$	8.47×10^{-1}	$\mathbf{1.42 \times 10^{-1}}$	1.75×10^{-1}	8.65×10^{-1}	9.88×10^{-2}
2024	XGBoost	3.05×10^{-1}	8.12×10^{-1}	1.62×10^{-1}	$\mathbf{1.60 \times 10^{-1}}$	$\mathbf{8.71 \times 10^{-1}}$	$\mathbf{8.61 \times 10^{-2}}$

Table 9
External validation comparing optimized model outputs with actual values collected in the system.

Sample	Models	Chlorination			Fluoridation		
		RMSE	R^2	MAE	RMSE	R^2	MAE
2022	ANN-MLP	1.90×10^{-1}	9.89×10^{-1}	6.37×10^{-2}	$\mathbf{6.11 \times 10^{-2}}$	$\mathbf{9.96 \times 10^{-1}}$	$\mathbf{1.39 \times 10^{-2}}$
2022	RF	$\mathbf{1.49 \times 10^{-1}}$	$\mathbf{9.93 \times 10^{-1}}$	$\mathbf{3.93 \times 10^{-2}}$	1.28×10^{-1}	9.85×10^{-1}	3.17×10^{-2}
2022	XGBoost	1.87×10^{-1}	9.89×10^{-1}	5.76×10^{-2}	1.07×10^{-1}	9.89×10^{-1}	3.85×10^{-2}
2024	ANN-MLP	4.17×10^{-1}	9.44×10^{-1}	9.64×10^{-2}	2.47×10^{-1}	9.44×10^{-1}	$\mathbf{4.40 \times 10^{-2}}$
2024	RF	$\mathbf{4.03 \times 10^{-1}}$	$\mathbf{9.48 \times 10^{-1}}$	$\mathbf{7.79 \times 10^{-2}}$	2.45×10^{-1}	9.45×10^{-1}	5.39×10^{-2}
2024	XGBoost	4.11×10^{-1}	9.46×10^{-1}	9.43×10^{-2}	$\mathbf{2.38 \times 10^{-1}}$	$\mathbf{9.48 \times 10^{-1}}$	6.05×10^{-2}

A comparison between Table 8 (without optimization) and Table 9 (with optimization) indicates consistent improvements in R^2 and MAE between models. Although RMSE slightly increased for chlorine in some cases, predictions remained aligned with real system behavior, as illustrated in Figs. 11(b) and 12(d). For fluoride, this trend was observed in most models, except ANN-MLP. Although such patterns may indicate potential overfitting, external validation confirms that all models retained generalization capacity under operational conditions.

The differences in RMSE, R^2, and MAE between the models remained within the 10^{-2} order of magnitude, and all predicted doses

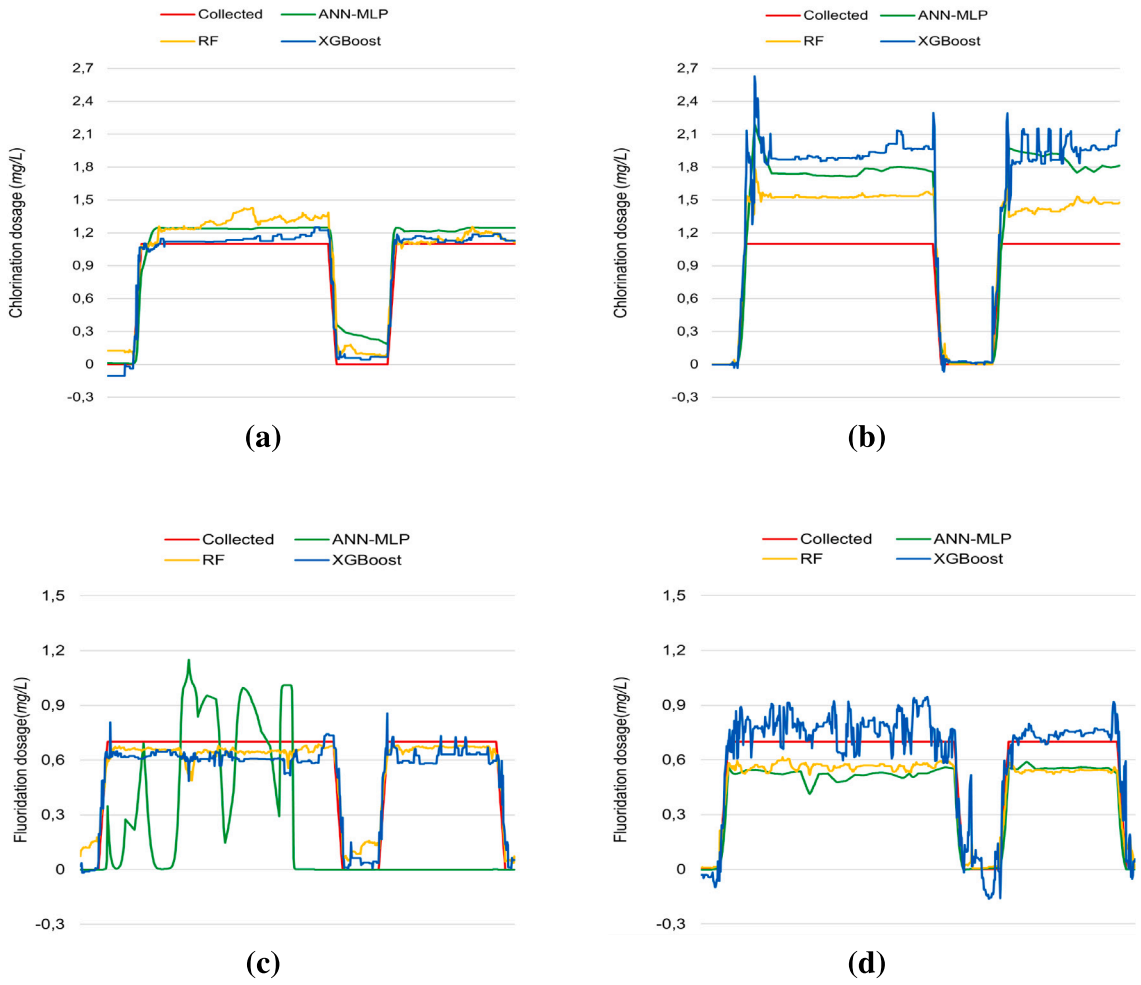

Fig. 11. Real system values and model predictions with sample from 2022: (a) chlorine without hyperparameters optimization, (b) chlorine with hyperparameters optimization, (c) fluorine without hyperparameters optimization and (d) fluorine with hyperparameters optimization.

complied with the internationally recommended limits for chlorine and fluoride. In terms of computational demands, RF requires more memory and processing power during training due to its ensemble structure, despite involving fewer hyperparameters. In contrast, ANN-MLP and XGBoost offer faster inference but require extensive tuning of parameters such as the number of neurons per layer, learning rate, activation function (ANN-MLP) and tree-related hyperparameters such as depth, learning rate, and number of estimators (XGBoost). Improper choices can increase training time and raise the risk of overfitting, especially in large datasets or when technical expertise is limited. Although both methods include regularization mechanisms, their parameter complexity demands careful configuration to ensure reliable generalization. Beyond performance metrics and computational considerations, the interpretability of each model was examined to assess its practical applicability in decision support.

The Tables 8 and 9 indicate that even without hyperparameter optimization, the models achieved good agreement with field observations, with RF standing out for chlorination and ANN-MLP or XGBoost for fluoridation, depending on the period analyzed. After optimization, MAE consistently decreased and R^2 increased, confirming enhanced predictive stability under different operational conditions, although some cases exhibited marginal increases in RMSE. In real WTP applications, these results demonstrate that machine learning models can reliably support chlorine and fluoride dosing decisions while ensuring compliance with international limits, balancing accuracy with computational feasibility, and providing flexibility across diverse monitoring contexts and data availability.

4.3.2. Feature importance and interpretability of predictive models

To enhance the interpretability of the model, a feature importance analysis was performed for both RF and XGBoost, following the principles of explainable AI (XAI). In both models, the most influential variables for predicting chlorine and fluoride concentrations were residual chlorine, temperature, and pH. These features consistently ranked highest in XGBoost, while RF yielded similar importance patterns with minor variations, likely due to its ensemble-based architecture.

These findings provide valuable guidance for operational decisions in water treatment, particularly in identifying the environmental parameters that most influence dosing behavior. The ability to highlight key variables increases the transparency of the model, a prerequisite for the practical implementation of AI systems in the sanitation sector [36, 37]. This interpretability supports more strategic sensor placement and monitoring practices, contributing to the development of resilient and explainable solutions tailored to resource-limited settings. Furthermore, the ability to trace the influence of the variables enables integrated evaluations of technical performance, environmental conditions, and economic results.

4.4. Technical, environmental and economic impact assessment

The accuracy of the chemical dosing was substantially improved after optimization of the model. For chlorine, the RF model maintained residual concentrations within the regulatory range of 0.2–1.0 mgL^{-1}, achieving RMSE = 0.40 and R^2 = 0.91. For fluoride, XGBoost maintained values within 0.5–1.0 mgL^{-1}, with RMSE = 0.15 and R^2 = 0.95.

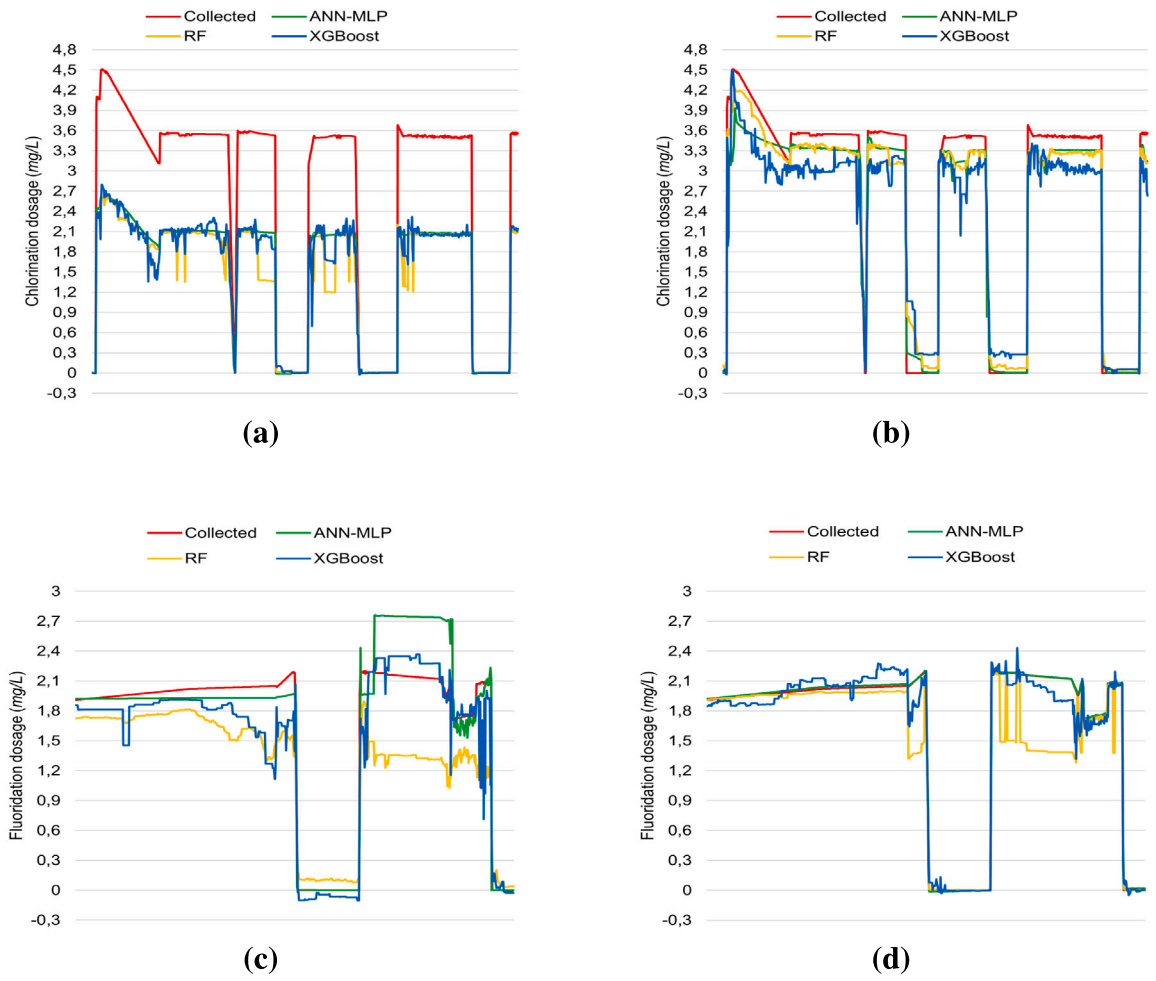

Fig. 12. Real system values and model predictions with sample from 2024: (a) chlorine without hyperparameters optimization, (b) chlorine with hyperparameters optimization, (c) fluorine without hyperparameters optimization and (d) fluorine with hyperparameters optimization.

Compared to conventional strategies, such as manual adjustment or linear regression models, which in similar contexts generally produce 10%–15%, the proposed models reduced manual dosing errors by up to 30%.

The reduction in DBP formation was quantified using operational data (Table S3 and Table S4), combined with proportional models relating chlorine dosage to the generation of by-products. From Table S3, baseline values were extracted: average chlorine dosage D_{Cl} = 1.45 mgL^{-1}, mean $pH \approx 6.6$ and water temperature around 25.8 °C. The estimated reduction in THM was derived using a simplified proportional model, following Miklos et al. [63]. Using $\mu_{D_{Cl}} = 1.45$ mgL^{-1}, an optimized dose $D_{Cl}^{O} = 1.20$ mgL^{-1}, and a regulatory limit THM_B = 80 μgL^{-1}, a reduction in chlorine 17% led to a decrease of 13.6 μgL^{-1} in THM formation. This value was extrapolated to a monthly reduction of 493.5 g of THM using the plant flow rate of 14 Ls^{-1}.

For HAA, the reduction was estimated using the same dataset and a proportionality factor $\alpha = 0.09$mg of HAA per mg of chlorine applied. With $\Delta D_{Cl} = 0.25$ mgL^{-1}, baseline HAA = 90 μgL^{-1} (from Table 10), and a treated monthly volume $Q_M = 36,288$ m^3, the estimated monthly reduction was 816.5 g. These values reflect pre-optimization conditions and quantify the environmental benefits of the optimized system.

Accurate dosing improves microbiological safety by reducing underdosing and limits the generation of DBP by avoiding overdosing. These results reinforce regulatory compliance and operational reliability in small-scale WTPs, according to WHO and the Brazilian Ministry of Health guidelines.

4.4.1. Economic assessment

Economic gains were estimated from three main components: (i) reduction in chemical inputs, (ii) savings in labor costs, and (iii) avoided regulatory penalties. A reduction of 11.5% in sodium hypochlorite use yielded 349.20 kg/year in savings, valued at US$136.19 (US$0.39/kg). Sodium fluorosilicate reduction added US$5.60 (1.13 kg/year at US$4.96/kg), totaling US$141.79 in direct chemical savings.

Automation allowed discontinuation of the 12 × 36 shift scheme. Based on operator salaries (US$2,249.84/ month), annual labor savings were estimated at US$26,998.03, corresponding to one full-time equivalent. No formal dismissals occurred, with savings derived from reduced overtime and standby hours. In addition, penalties were considered. According to Garibay-Rodriguez et al. [73], high-level regulatory sanctions range from US$7,887.33 to US$15,774.67. A conservative estimate of US$7,887.33 was adopted. Table 11 presents these savings.

Assuming implementation costs of US$6,700.00, the ROI was estimated at approximately 420%, with a payback period of 2.3 months. This value should be interpreted as indicative, as variations in chemical prices, labor costs, and regulatory penalties can change the estimate by about ±5–10%.

4.4.2. Sludge generation and environmental impact

The generation of solid waste results from (i) the removal of suspended solids and (ii) the formation of $Al(OH)_3$ flocs from the hydrolysis of aluminum sulfate. In dry periods, the turbidity reaches $10NTU$,

Table 10
Performance metrics of the AI optimization model and parameters used for by-product estimation.

Metric	Value	Regulatory Limit	Improvement
THM reduction[a]	493.5 g/month	80 μgL^{-1}	17%
HAA reduction[a]	816.5 g/month	60 μgL^{-1}	25%
Chlorine RMSE	0.40 mgL^{-1}	–	30% error reduction
Fluoride RMSE	0.15 mgL^{-1}	–	35% error reduction
Mean chlorine dosage[a]	1.45 mgL^{-1}	–	Table S3
Optimized chlorine dosage[a]	1.20 mgL^{-1}	–	Model output
Chlorine dosage reduction[a] ΔD_{Cl}	0.25 mgL^{-1}	–	Direct calculation
Baseline THM concentration	80 μgL^{-1}	80 μgL^{-1}	EPA/WHO guideline
Baseline HAA concentration	90 μgL^{-1}	60 μgL^{-1}	Table 1
Monthly treated volume[a]	36,288 m^3	–	$14 L/s \times 30$ days

[a] Estimated local guideline value, data from Saneago S.A. operational records.

Table 11
Estimated components of annual savings.

Component	Amount [US$]
Reduction in chemical inputs (chlorine and fluoride)	142.17
Reduction in labor costs	26,998.03
Avoided penalties (Regulatory Resolution)	7,887.33
Total estimated savings	**35,027.53**

Table 12
Performance of optimized AI models for chlorine and fluoride (external validation).

Model	Chlorine			Fluoride		
	RMSE	MAE	R^2	RMSE	MAE	R^2
ANN-MLP	0.66	0.56	0.72	0.18	0.13	0.93
RF	**0.40**	**0.36**	**0.90**	0.19	0.13	0.92
XGBoost	0.79	0.70	0.60	**0.15**	**0.09**	**0.95**

corresponding to 15 mgL^{-1} of suspended solids or 215 kg/month. During the rainy season, the turbidity increases to $115 NTU$, increasing the sludge to 2000 kg/month.

According to John [66] and Crittenden et al. [67], the coagulant ratios of sludge typically range from 0.25 to 0.6, kg/kg. In our study, the dry season ratio of 0.24, kg/kg indicated high efficiency, while the rainy season ratio of 1.1, kg/kg suggested underdosing or excess sediment. These deviations reflect the influence of operational factors such as fluctuations in pH and water temperature, as well as variations in coagulant purity and reactivity, which affect the effective conversion of aluminum salts into hydroxides and, consequently, the amount of solids generated. Therefore, real-time turbidity-based dosing is recommended to maintain treatment performance, with broader sustainability implications.

4.4.3. Model performance summary and limitations

Independent test sets confirmed the accuracy of the model. As shown in Table 12, optimized RF yielded RMSE = 0.40 and $R^2 = 0.90$ for chlorine, XGBoost achieved RMSE = 0.15 and $R^2 = 0.95$ for fluoride. The integration of STL improved both interpretability and adaptability.

Minor performance drops occurred in cases of long data gaps, sensor anomalies, or operational plateaus (e.g., stable pH or flow). These were more evident in the early phases with limited historical data. Table 12 and Section 5 detail mitigation strategies to maintain the reliability of the model under such constraints. Historically, small-scale WTPs have relied on manual rules or static models, unable to capture nonlinearities or seasonal changes [12,14]. More recent approaches, like ARIMA and spline interpolation, offer better smoothing but falter under long gaps or non-stationarity.

To address these issues, the proposed framework integrates STL to: (i) decompose series into trend, seasonality, and residuals, (ii) handle missing data and irregular sampling and (iii) improve feature extraction. Preliminary tests showed that XGBoost alone performed poorly during abrupt regime shifts, reinforcing the use of STL+XGBoost for more interpretable and resilient performance in data-scarce environments.

5. Discussion

This study developed and validated a hybrid STL-XGBoost model to optimize chlorine and fluoride doses in small-scale water treatment plants. The results support the central hypothesis that the integration of seasonal trend decomposition (STL) with XGBoost improves predictive accuracy and promotes operational sustainability under data scarcity and seasonal variability conditions. Furthermore, they demonstrate that all the objectives defined in the introduction were achieved. The discussion was structured to explicitly address each objective, establishing a direct link between the results obtained and the intended goals of the study.

5.1. Model development, validation, and comparative performance

The strategic alignment between model selection and STL-based preprocessing was confirmed by the results. Seasonal trend decomposition reformulated the complex time-series forecasting problem into a supervised learning task with tabular data. This transformation reduced reliance on models designed for long-range temporal dependencies, such as LSTM, and allowed the use of well-established algorithms for static tabular structures, including XGBoost [74] and Random Forest. These algorithms achieved an adequate balance between predictive accuracy, computational efficiency, and operational applicability within the context analyzed. In particular, XGBoost excelled in optimizing dosing in small-scale water treatment plants due to its inherent ability to handle missing values and capture complex, nonlinear relationships, such as seasonal flow variations during rainy periods. This reduced the need for extensive preprocessing and ensured predictive resilience under non-stationary conditions.

The outperformance of XGBoost, which was decisive in achieving the study objectives, is grounded in specific algorithmic advantages that address real operational constraints. Its built-in regularization mechanisms, including L1/L2 penalties and pruning, mitigate overfitting even under noisy conditions, a relevant feature given the frequent sensor anomalies observed in the field. Its scalability, supported by parallel processing that enables rapid updates even on low-cost hardware such as the Raspberry Pi, makes it suitable for remote treatment plants with limited infrastructure. This combination of efficiency, resilience, and adaptability sets XGBoost apart from ANN-MLP, which is sensitive to data imbalance, and from Random Forest, which entails higher memory requirements. These characteristics directly contributed to the resilience of the model when handling incomplete data and to its better performance in comparative analysis.

The comparative analysis conducted in this study is grounded in the well-documented performance of artificial intelligence methods compared to traditional control strategies. Conventional approaches, typically manual or based on linear regression, are limited in both

Table 13
Comparative summary of machine learning metrics applied to water resources and water quality: previous studies ×this study.

Parameter/Metric	Previous studies	This study
Coverage of metrics (RMSE, MAE, R^2)	Partial, typically restricted to one or two metrics [75–78]	Comprehensive: RMSE, MAE, R^2, and standard deviation
External validation	Rare or absent [79–81]	Included (datasets A22 and A24)
Variable interpretation (XAI)	Limited exploration [26,75,80]	Highlighted (variable importance, transparency)
Handling of missing data	Limited, generally through deletion or simple interpolation [26,52,75,79]	Advanced (STL decomposition, reliable imputation)
Techno-environmental-economic integration	Largely absent [50,52,82]	Developed (impact on DBPs, ROI, sludge management)
Resilience under different seasonal conditions	Rarely addressed [81,83]	Assessed (wet season vs. dry season)
Applicability to small-scale DWTPs	Commonly overlooked [26,76,78,79]	Demonstrated as feasible and scalable

accuracy and resilience, often leading to excessive chemical consumption and increased operational costs [20,23,28]. In contrast, artificial intelligence techniques such as ANN-MLP, Random Forest, and XGBoost have demonstrated higher predictive accuracy along with measurable operational, economic, and environmental benefits [19,21]. The results obtained in this study not only corroborate, but also extend previous findings by addressing a relevant methodological gap. Consistent with Solaimany-Aminabad et al. [19], the suitability of tree-based algorithms for handling heterogeneous inputs and pronounced seasonality was confirmed. However, by integrating trend and seasonal decomposition (STL), specifically tailored to the challenges of small-scale treatment plants, the resilience of the model was improved against incomplete datasets and seasonal fluctuations, thus directly fulfilling the study objectives.

For comprehensive validation against the stated objectives, the results were contextualized in light of the existing literature and supported by empirical evidence. As summarized in Table 13, this study advances previous work by performing a full evaluation (RMSE, MAE, R^2 and standard deviation), incorporating external validation, applying interpretability techniques and integrating technical, environmental and economic dimensions, an approach rarely achieved in previous investigations. This framework underscores both the methodological innovation and the practical relevance of the proposed model for small-scale water treatment plants.

The empirical validation consistently reinforces the identified advantages. The scatter plot of observed versus predicted values, Fig. 10, highlights the systematic trends of each model, demonstrating that the optimized STL–XGBoost algorithm consistently reduces deviations throughout the operational range. This reliable accuracy is critical for the calibration and adjustment of the reagent dosing. The temporal analysis, Fig. 11, confirms the model's ability to accurately reproduce complex concentration dynamics, even under operational disturbances, thus providing direct support for real-time decision-making. Finally, the residual variability analysis, Fig. 9, shows a narrower error distribution in the optimized model, translating into greater operational reliability by reducing the likelihood of critical dose failures. Taken together, this body of evidence demonstrates the enhanced resilience of the model under data limitations and its performance advantage over alternative approaches.

5.2. Operational, environmental, and economic benefits

This subsection addresses the objectives of reducing dosing errors while ensuring regulatory compliance, as well as quantifying environmental and economic benefits. The challenge of chlorine and fluoride dosing in resource-limited settings, often linked to microbiological risks or the excessive formation of harmful by-products (THM and HAA) [12,14,84], is directly addressed by the proposed STL-XGBoost model. The results demonstrate that accurate AI-supported predictions can significantly reduce dosing errors, maintain compliance, and minimize the formation of by-products, thus delivering measurable operational, environmental, and economic gains.

The optimized models exhibited lower dispersion around the mean, reflected in reduced standard deviations, therefore increasing the reliability by lowering the probability of erroneous decisions. Beyond the methodological advance, the study also demonstrated tangible operational benefits, including a reduction in fluoride consumption of approximately $15\% \pm 3\%$ ($\approx$ 50 kg/year) and sodium hypochlorite consumption of $18\% \pm 2\%$ ($\approx$ 120 kg/year), resulting in annual savings of about US$ 240. For applications in computationally constrained contexts, strategies such as edge computing and Secure Multiparty Computation (MPC) may enable local processing and decentralized model training, reducing reliance on centralized infrastructure and associated costs.

The results represent a substantial advance over the state of the art and demonstrate alignment with the Sustainable Development Goals (SDGs). Regarding SDG 6 (Clean Water and Sanitation), the optimized XGBoost model achieved 99% compliance with WHO guidelines, supported by consistently low RMSE values. In relation to SDG 12 (Responsible Consumption and Production), accurate dosing predictions led to a 17% reduction in THM formation and a 25% reduction in HAA formation. Compared to manual procedures, dosing errors were reduced by approximately 45%, strengthening the potential of artificial intelligence approaches for water treatment management. This study highlights the broad applicability and tangible benefits of incorporating the proposed AI framework into disinfectant control in small- and medium-scale water treatment plants. By integrating environmental protection, economic efficiency and reproducibility, the approach supports the transition toward cleaner, safer, and more resilient sanitation systems, strengthening progress toward SDG 6.

5.3. Policy implications and scalability

These findings strengthen the feasibility of implementing artificial intelligence solutions in facilities with limited resources, such as Brazabrantes WTP, illustrating how accessible and adaptable technologies can simultaneously advance public health and environmental sustainability. The adoption of low-cost predictive models, aligned with the principles of transparency, interpretability, and equity, is consistent with G 20 commitments to ensure that digital transformation reduces regional disparities and promotes social inclusion. Designed to operate reliably under infrastructure and data constraints, the model is particularly suitable for small and medium-sized WTPs in resource-limited contexts, enhancing process control without the need for costly instrumentation or extensive infrastructure upgrades.

This study provides a practical basis for the formulation of public policies that aim to improve the sustainability of sanitation systems. Reductions in chemical waste, operational improvements, and compliance with international standards support the implementation of stricter environmental regulations, particularly regarding automated chemical dosing in WTPs. Government agencies are encouraged to introduce economic and fiscal incentives to promote the adoption of artificial intelligence tools in public water supply systems, especially in resource-limited municipalities. Furthermore, standardized protocols for data collection and storage are necessary to enable model replication and ensure consistency in monitoring, auditing, and decision-making processes. The adaptability of the framework to various operational contexts improves its replicability in public water systems, particularly in developing countries where financial constraints frequently hinder digital innovation.

The findings indicate that the proposed optimization model has considerable potential to inform public sanitation policies focusing on sustainability, operational efficiency, and equitable access to safe drinking water. These results are consistent with global sustainability agendas, particularly SDG 6 and target 6.3, by advancing pollution control, chemical optimization, and universal access to safe water. The model reflects emerging governance paradigms that emphasize efficiency, transparency, and adaptability. By integrating artificial intelligence with sustainability metrics and regulatory compliance, the framework supports the transition to digitally integrated, resilient, and sustainable sanitation systems, thus improving its scalability and relevance to policy. These attributes contribute to cost-effective, environmentally responsible and socially inclusive sanitation services. More details on scalability and governance implications are provided in Section S1 and Section S3 of the Supplementary Information.

5.4. Limitations and future research directions

Despite the documented benefits, a trade-off between computational efficiency and environmental impact was observed. Although the RF model achieved high predictive accuracy, it required approximately 30% more computational resources than XGBoost, which may restrict its applicability in settings with limited infrastructure or energy availability. In contrast, XGBoost combined superior predictive performance with lower energy demand, making it more suitable for sustainable implementation in small and medium-sized WTPs.

In addition to computational trade-offs, practical constraints can also restrict the broader adoption of the proposed optimization framework. A major challenge arises from the dependence on historical operational data, which are frequently incomplete, inconsistent, or noisy due to irregular sampling, equipment failures, or manual entry errors. The impact of missing data was quantitatively assessed during method development (Table 4). In our validation, removal of 20% of the data points led to a measurable decline in the performance of the model. The application of the STL gap filling technique proved important in addressing this issue, reducing the RMSE for critical parameters such as chlorine dosage (D_{Cl}) by 12. 4% compared to simpler imputation approaches. These results indicate that, although the model is sensitive to data gaps, appropriate preprocessing can effectively maintain predictive precision. In addition, the model generalization capability may be constrained by sensor quality and data irregularities inherent in real-world deployment scenarios.

However, performance in extreme data loss scenarios (> 30%) or during high-turbidity events remains a limitation, as reflected by the higher residual errors in Table 4 for parameters such as T_{ab}. Although the STL method alleviates part of this issue, the model still depends on a minimum threshold of data quality. This limitation is further compounded by the limited technological infrastructure in many facilities. In Brazil, for example, only about 20% of the WTPs maintain historical records with sufficient quality and resolution to support reliable model training. Moreover, even artificial intelligence models tailored for low-resource contexts require essential infrastructure, including calibrated sensors, reliable network connectivity, and basic local computing capacity.

An important limitation concerns the applicability of the model to groundwater systems, since it was developed specifically for surface water sources. Groundwater was excluded due to fundamental operational differences that hinder direct model transfer: (1) the markedly reduced seasonal variability in flow rate and quality parameters (e.g., turbidity, organic matter), which compromises the core STL decomposition strategy, (2) distinct disinfection by-product (DBP) associated with lower concentrations of natural organic matter (NOM), which alter the kinetics of chlorine decay and require alternative predictor variables, and (3) an infrastructural mismatch, since groundwater systems often operate through direct well-to-distribution networks without contact tanks, thus eliminating hydraulic retention time, a critical input variable.

Although adaptation is technically feasible, it would require retraining with groundwater-specific datasets and re-calibration of threshold logic, as direct model transfer often leads to significant performance loss [85,86]. As a result, the current implementation restricts immediate scalability to larger or groundwater-based systems with distinct characteristics, and its generalizability in diverse operational contexts remains an open area for future research.

Future research should prioritize strategies to address these limitations. To overcome data and infrastructure constraints, the integration of IoT devices and active learning approaches could enable adaptive updates from incoming data and strengthen resilience under real-world conditions. Federated learning is also recommended to facilitate decentralized training without centralized data aggregation, thereby lowering energy demand, preserving data sovereignty, and fostering collaboration between WTPs. Regarding model transferability, future work should investigate domain-adaptive machine learning architectures (e.g., adversarial training) to narrow the performance gap between surface water and groundwater systems.

Additional research avenues include real-time deployment with smart sensors and autonomous control mechanisms. Large-scale evaluations across diverse regions and hydrological contexts could provide greater support for model scalability. The integration of artificial intelligence with renewable energy sources, such as solar photovoltaics, may further reduce the carbon footprint of water treatment operations. In addition, incremental learning techniques could allow continuous updates to the model without the need for complete retraining, thereby improving adaptability over time.

The proposed framework demonstrates strong potential to improve both operational efficiency and environmental sustainability in small-scale WTPs. Its design aligns with the practical constraints commonly faced in developing regions. By integrating artificial intelligence with STL, the model sustains reliable performance even when historical data are incomplete. To fully realize the benefits of AI-driven optimization, the implementation of digital monitoring infrastructure, particularly IoT-enabled sensors, is essential. Pilot deployments in facilities with diverse treatment configurations and water sources are necessary to advance SDG 6 by promoting safe, equitable and sustainable water treatment in resource-constrained environments.

6. Conclusion

This study developed and validated an artificial intelligence–supported forecasting framework for chlorine and fluoride dosing in small-scale water treatment plants, achieving consistent improvements over conventional approaches. The framework integrates STL decomposition with XGBoost to optimize dosing under conditions of data uncertainty or fragmentation, without the need for extensive instrumentation. The results confirmed the main objectives of the study, demonstrating substantial gains in operational efficiency and environmental sustainability. Key findings include a 15%–18% reduction in

reagent consumption, approximately 99% compliance with WHO standards, and a 17%–25% decrease in the formation of toxic disinfection by-products (THM and HAA). By maintaining high predictive accuracy with low computational demand, the framework proved particularly suitable for resource-limited contexts, supporting practical advances toward the targets of SDG 6.

The findings were integrated into a broader perspective through a discussion of practical and managerial implications, environmental and sustainability benefits, public policy implications, and limitations that inform future research directions. This comprehensive analysis highlights the potential for the model to be implemented in resource-limited contexts, its ability to generate significant benefits through optimized dosing, and its alignment with policies that support large-scale adoption, while also acknowledging areas for further improvement.

Despite the promising results, future research should focus on addressing the identified limitations. Priority directions include: (1) extending the framework to groundwater sources through domain adaptation techniques to account for fundamental physicochemical differences, (2) implementing federated learning architectures to allow secure and decentralized model training across multiple plants, and (3) integrating the model with IoT-based sensors and renewable energy sources to support autonomous, low-carbon operation. Large-scale validation in diverse geographical and operational contexts will be essential to confirm scalability and resilience. In conclusion, this study provides a practical, data-driven pathway for the gradual digital modernization of water treatment systems. It also offers a replicable framework to strengthen public health and environmental protection in developing regions, aligning technological innovation with urgent real-world needs.

CRediT authorship contribution statement

Diego Takashi Sato: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Orlando M. Oliveira Belo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antonio P. Castro:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Viviane M. Gomes Pacheco:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cloves Gonçalves Rodrigues:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antonio Paulo Coimbra:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Wesley Pacheco Calixto:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

This work was supported by the Fundação para a Ciência e a Tecnologia (FCT, Portugal), under the project UIDB/00048/2020. Additional support was provided by the National Council for Scientific and Technological Development (CNPq, Brazil), through a Research Productivity Fellowship (Grant No. 301644/2022-5), and by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES, Brazil), through Postdoctoral Fellowships (Grant No. 88887.985910/2024-00).

Data and code availability

All data and source code used in this study are publicly available on Zenodo at the following DOI: https://doi.org/10.5281/zenodo.15770106. The repository includes a full set of pre-processing scripts, model training routines, and evaluation procedures to ensure complete transparency and reproducibility. These resources comply with the data sharing policy of *Journal of Water Process Engineering*, supporting the development of open, replicable and accessible environmental decision support systems.

Declaration of competing interest

We, the undersigned authors of the submitted manuscript, declare that we have no financial interests or conflicts of interest that could potentially influence the objectivity, integrity, or impartiality of our research findings. Specifically:

1. Financial Support: The research conducted and the preparation of this manuscript received no external financial support, grants, or funding from any public or private entity.
2. Patents: We confirm that there are no patents associated with the research work presented in this manuscript, and no patent applications have been submitted during the course of this study.
3. Salary Reimbursement: The authors involved in this research project have not received any salary, fees, or reimbursements related to the publication of this work. The research was conducted as part of our academic and professional activities, and no financial compensation has been sought or received.

The authors hereby affirm that this Declaration of Interest accurately reflects their financial and non- financial relationships, and they acknowledge their responsibility to promptly inform the editorial board of any changes in our circumstances that may impact this declaration during the review process.

Acknowledgments

The authors acknowledge the financial support provided by the Fundação para a Ciência e a Tecnologia (FCT/Portugal), I.P., under project UIDB/00048/2020 (DOI: 10.54499/UIDB/00048/2020) and the National Council for Scientific and Technological Development (CNPq/Brazil) for supporting this research through a Research Productivity Fellowship (Grant No. 301644/2022-5). Furthermore, the authors express their gratitude to the Coordination for the Improvement of Higher Education Personnel (CAPES/Brazil) for the postdoctoral fellowship support (Grant No. 88887.985910/2024-00). The authors also express their gratitude to Saneago S.A. for providing the data and to LaMCAD/UFG for supporting the research with computational resources.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jwpe.2025.108736.

Data and code availability

All data and source code used in this study are publicly available on Zenodo at the following DOI: https://doi.org/10.5281/zenodo.15770106. The repository includes a full set of pre-processing scripts, model training routines, and evaluation procedures to ensure complete transparency and reproducibility. These resources comply with the data sharing policy of *Journal of Water Process Engineering*, supporting the development of open, replicable and accessible environmental decision support systems.

References

[1] Xing-Fang Li, William A. Mitch, Drinking water disinfection byproducts (DBPs) and human health effects: multidisciplinary challenges and opportunities, Environ. Sci. Technol. 52 (4) (2018) 1681–1689.

[2] Sundas Kali, Marina Khan, Muhammad Sheraz Ghaffar, Sajida Rasheed, Amir Waseem, Muhammad Mazhar Iqbal, Muhammad Bilal khan Niazi, Mazhar Iqbal Zafar, Occurrence, influencing factors, toxicity, regulations, and abatement approaches for disinfection by-products in chlorinated drinking water: A comprehensive review, Environ. Pollut. 281 (2021) 116950.

[3] Rupal Sinha, Ashok Kumar Gupta, Partha Sarathi Ghosal, A review on trihalomethanes and haloacetic acids in drinking water: Global status, health impact, insights of control and removal technologies, J. Environ. Chem. Eng. 9 (6) (2021) 106511.

[4] Andreea Florina Gilca, Carmen Teodosiu, Silvia Fiore, Corina Petronela Musteret, Emerging disinfection byproducts: A review on their occurrence and control in drinking water treatment processes, Chemosphere 259 (2020) 127476.

[5] Aniruddha Bhalchandra Pandit, Jyoti Kishen Kumar, Drinking Water Treatment for Developing Countries: Physical, Chemical and Biological Pollutants, Royal Society of Chemistry, 2019.

[6] Sheila Olmstead, Jiameng Zheng, Water pollution control in developing countries: Policy instruments and empirical evidence, Rev. Environ. Econ. Policy 15 (2) (2021) 261–280.

[7] Jusman Rahim, Lilin Budiati, et al., Drinking water treatment system and the challenges faced by developing countries, Malays. J. Med. Heal. Sci. 20 (2) (2024) 293–299.

[8] Rucha Vaidya, Kavita Verma, Mohan Kumar, Chanakya Hoysall, Lakshminarayana Rao, Assessing wastewater management challenges in developing countries: a case study of India, current status and future scope, Environ. Dev. Sustain. 26 (8) (2024) 19369–19396.

[9] R. Makungo, J.O. Odiyo, N. Tshidzumba, Performance of small water treatment plants: The case study of Mutshedzi Water Treatment Plant, Phys. Chem. Earth (ISSN: 1474-7065) 36 (14–15) (2011) 1151–1158, http://dx.doi.org/10.1016/j.pce.2011.07.073, 11th WaterNet/WARFSA/GWP-SA Annual Symposium, Victoria Falls, ZIMBABWE, OCT 27-29, 2010.

[10] Derrick Dadebo, Denis Obura, Nathan Etyang, David Kimera, Economic and social perspectives of implementing artificial intelligence in drinking water treatment systems for predicting coagulant dosage: A transition toward sustainability, Groundw. Sustain. Dev. 23 (2023) 100987.

[11] Amit Bhati, Kamal Kant Hiran, Ajay Kumar Vyas, Maad M. Mijwil, Mohammad Aljanabi, Ahmed Sayed M. Metwally, Md Fayz Al-Asad, Mohd Khalid Awang, Hijaz Ahmad, Low cost artificial intelligence internet of things based water quality monitoring for rural areas, Internet Things 27 (2024) 101255.

[12] Michael Ayorinde Dada, Michael Tega Majemite, Alexander Obaigbena, Onyeka Henry Daraojimba, Johnson Sunday Oliha, Zamathula Queen Sikhakhane Nwokediegwu, Review of smart water management: IoT and AI in water and wastewater treatment, World J. Adv. Res. Rev. 21 (1) (2024) 1373–1382.

[13] Claudia Cossio, Jenny Norrman, Jennifer McConville, Alvaro Mercado, Sebastien Rauch, Indicators for sustainability assessment of small-scale wastewater treatment plants in low and lower-middle income countries, Environ. Sustain. Indic. 6 (2020) 100028.

[14] Natalia Pichel, Marta Vivar, Manuel Fuentes, The problem of drinking water access: A review of disinfection technologies with an emphasis on solar treatment methods, Chemosphere 218 (2019) 1014–1030.

[15] Jan Ott, World bank world development reports, in: Encyclopedia of Quality of Life and Well-Being Research, Springer, 2024, pp. 7858–7859.

[16] Wenyu Yang, Christian Schmidt, Shixue Wu, Ziyong Zhao, Ruifei Li, Zhenyu Wang, Haijun Wang, Pei Hua, Peter Krebs, Jin Zhang, Exacerbated anthropogenic water pollution under climate change and urbanization, Water Res. 280 (2025) 123449.

[17] Weixin Zhao, Yanan Hou, Liangliang Wei, Wei Wei, Kefeng Zhang, Haoran Duan, Bing-Jie Ni, Chlorination-induced spread of antibiotic resistance genes in drinking water systems, Water Res. (2025) 123092.

[18] Umesh Ghimire, Gideon Sarpong, Veera Gnaneswar Gude, Transitioning wastewater treatment plants toward circular economy and energy sustainability, ACS Omega 6 (18) (2021) 11794–11803.

[19] Mehri Solaimany-Aminabad, Afshin Maleki, Mahdi Hadi, Application of artificial neural network (ANN) for the prediction of water treatment plant influent characteristics, J. Adv. Environ. Heal. Res. 1 (2) (2013) 89–100.

[20] André Felipe Librantz, Fábio Cosme Rodrigues dos Santos, Cleber Gustavo Dias, Artificial neural networks to control chlorine dosing in a water treatment plant, Acta Sci. Technol. (ISSN: 1807-8664) 40 (1) (2018) e37275.

[21] Yanyang Zhang, Xiang Gao, Kate Smith, Goulven Inial, Shuming Liu, Lenny B. Conil, Bingcai Pan, Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network, Water Res. 164 (2019) 114888.

[22] Ruixing Huang, Chengxue Ma, Jun Ma, Xiaoliu Huangfu, Qiang He, Machine learning in natural and engineered water systems, Water Res. 205 (2021) 117666.

[23] Maria Teresa Gaudio, Gerardo Coppola, Lorenzo Zangari, Stefano Curcio, Sergio Greco, Sudip Chakraborty, Artificial intelligence-based optimization of industrial membrane processes, Earth Syst. Environ. 5 (2) (2021) 385–398.

[24] Robin J. Slaughter, Martin Watts, J. Allister Vale, Jacob R. Grieve, Leo J. Schep, The clinical toxicology of sodium hypochlorite, Clin. Toxicol. (ISSN: 1556-3650) 57 (5) (2019) 303–311, http://dx.doi.org/10.1080/15563650.2018.1543889.

[25] Majid Gholami Shirkoohi, Rajeshwar Dayal Tyagi, Peter A Vanrolleghem, Patrick Drogui, Artificial intelligence techniques in electrochemical processes for water and wastewater treatment: a review, J. Environ. Heal. Sci. Eng. 20 (2) (2022) 1089–1109.

[26] Matthew Lowe, Ruwen Qin, Xinwei Mao, A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring, Water 14 (9) (2022) 1384.

[27] Guangtao Fu, Yiwen Jin, Siao Sun, Zhiguo Yuan, David Butler, The role of deep learning in urban water management: A critical review, Water Res. 223 (2022) 118973.

[28] Alka S. Kote, Dnyaneshwar V. Wadkar, Modeling of chlorine and coagulant dose in a water treatment plant by artificial neural networks, Eng. Technol. Appl. Sci. Res. 9 (3) (2019) 4176–4181.

[29] Adriano Bressane, Ana Paula Garcia Goulart, Carrie Peres Melo, Isadora Gurjon Gomes, Anna Isabel Silva Loureiro, Rogério Galante Negri, Rodrigo Moruzzi, Adriano Gonçalves dos Reis, Jorge Kennety Silva Formiga, Gustavo Henrique Ribeiro da Silva, et al., A non-hybrid data-driven fuzzy inference system for coagulant dosage in drinking water treatment plant: machine-learning for accurate real-time prediction, Water 15 (6) (2023) 1126.

[30] Md Khan, Nazrul Islam, Jia Uddin, Sifatul Islam, Mostofa Nasir, Water quality prediction and classification based on principal component regression and gradient boosting classifier approach, J. King Saud Univ. - Comput. Inf. Sci. 34 (2021).

[31] A. Pinto, A. Fernandes, H. Vicente, J. Neves, Optimizing water treatment systems using artificial intelligence based tools, WIT Trans. Ecol. Environ. (ISSN: 1743-3541) 125 (2009) 185–194.

[32] Dongsheng Wang, Yan Wang, Rui Zhou, Yong Cao, Fuchun Jiang, Xue Zhang, Jinghua Li, Water plant optimization control system based on machine learning, Desalination Water Treat. (ISSN: 1944-3994) 222 (2021) 168–181, http://dx.doi.org/10.5004/dwt.2021.27056.

[33] Ramon Pérez, Albert Martínez-Torrents, Manuel Martínez, Sergi Grau, Laura Vinardell, Ricard Tomàs, Xavier Martínez-Lladó, Irene Jubany, Chlorine concentration modelling and supervision in water distribution systems, Sensors 22 (15) (2022) 5578.

[34] Iman Jafari, Rongmo Luo, Emily Ng, Felipe Corral Jr., Yixiong Chua, Szu Hui Ng, Jiangyong Hu, Robust prediction of residual chlorine decay in a drinking water distribution system integrating water quality sensing and predictive tools, ACS ES T Water 4 (12) (2024) 5506–5521.

[35] Machodi Mathaba, JeanClaude Banza, A comprehensive review on artificial intelligence in water treatment for optimization. Clean water now and the future, J. Environ. Sci. Heal. Part A 58 (14) (2023) 1047–1060.

[36] Sheetal Kumari, Jyoti Chowdhry, Manoj Chandra Garg, AI-enhanced adsorption modeling: challenges, applications, and bibliographic analysis, J. Environ. Manag. 351 (2024) 119968.

[37] Catherine E Richards, Asaf Tzachor, Shahar Avin, Richard Fenner, Rewards, risks and responsible deployment of artificial intelligence in water systems, Nat. Water 1 (5) (2023) 422–432.

[38] Hanyuan Zhang, Wenxin Yang, Weilin Yi, Jit Bing Lim, Zenghui An, Chengdong Li, Imbalanced data based fault diagnosis of the chiller via integrating a new resampling technique with an improved ensemble extreme learning machine, J. Build. Eng. 70 (2023) 106338.

[39] Wenxin Yang, Hanyuan Zhang, Jit Bing Lim, Yuyu Zhang, Huanhuan Meng, A new chiller fault diagnosis method under the imbalanced data environment via combining an improved generative adversarial network with an enhanced deep extreme learning machine, Eng. Appl. Artif. Intell. 137 (2024) 109218.

[40] Hanyuan Zhang, Yuyu Zhang, Huanhuan Meng, Jit Bing Lim, Wenxin Yang, A novel global modelling strategy integrated dynamic kernel canonical variate analysis for the air handling unit fault detection via considering the two-directional dynamics, J. Build. Eng. 96 (2024) 110402.

[41] Renfei He, Limao Zhang, Alvin Wei Ze Chew, Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning, Knowl.-Based Syst. 251 (2022) 109125.

[42] Zhaoyang Xiong, Xingyang Liu, Thomas Igou, Zhanchao Li, Yongsheng Chen, Using hybrid machine learning to predict wastewater effluent quality and ensure treatment plant stability, Water 17 (13) (2025) 1851.

[43] Antonio Duarte Marcos Junior, Cleiton da Silva Silveira, José Micael Ferreira da Costa, Suellen Teixeira Nobre Gonçalves, Combining traditional hydrological models and machine learning for streamflow prediction, RBRH 29 (2024) e11.

[44] Hui Guo, Zhiyuan Chen, Fang Yenn Teo, Intelligent water quality prediction system with a hybrid CNN–LSTM model, Water Pr. Technol. 19 (11) (2024) 4538–4555.

[45] Dulce Brigite Ocampo-Rodríguez, Gabriela Alejandra Vázquez-Rodríguez, Sylvia Martínez-Hernández, Ulises Iturbe-Acosta, Claudia Coronel-Olivares, Water disinfection: a review of conventional and advanced treatments with chlorine and peracetic acid, IngenieríA Del Agua 26 (3) (2022) 185–204.

[46] Jeffrey W.A. Charrois, Private drinking water supplies: challenges for public health, Can. Med. Assossiation J. 182 (10) (2010) 1061–1064.
[47] Surendra Roy, Gurcharan Dass, Fluoride contamination in drinking water–a review, Resour. Env. 3 (3) (2013) 53–58.
[48] Annepu Arudra, Dankan Gowda, Parashuram Shankar Vadar, Mahadeo Ramchandra Jadhav, Mandeep Kaur, et al., Predictive modeling of dental health outcomes based on fluoride concentrations using AI, in: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking, SMART GENCON, IEEE, 2023, pp. 1–7.
[49] K.G. Aparna, R. Swarnalatha, et al., Optimizing wastewater treatment plant operational efficiency through integrating machine learning predictive models and advanced control strategies, Process. Saf. Environ. Prot. 188 (2024) 995–1008.
[50] Subin Lin, Jiwoong Kim, Chuanbo Hua, Seoktae Kang, Mi-Hyun Park, Comparing artificial and deep neural network models for prediction of coagulant amount and settled water turbidity: lessons learned from big data in water treatment operations, J. Water Process. Eng. 54 (2023) 103949.
[51] Tiexiang Mo, Shanshan Li, Guodong Li, An interpretable machine learning model for predicting cavity water depth and cavity length based on XGBoost-SHAP, J. Hydroinformatics 25 (4) (2023) 1488–1500.
[52] Junio S. Bulhoes, Cristiane L. Martins, Marcia D. Oliveira, Debora F. Calheiros, Wesley P. Calixto, Indirect prediction system for variables that have gaps in their time series, Chaos Solitons Fractals 131 (2020) 109509.
[53] Zuokun Ouyang, Meryem Jabloun, Philippe Ravier, STLformer: exploit STL decomposition and rank correlation for time series forecasting, in: 2023 31st European Signal Processing Conference, EUSIPCO, IEEE, Helsinki, Finland, 2023, pp. 1405–1409.
[54] Rohit Nishant, Mike Kennedy, Jacqueline Corbett, Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda, Int. J. Inf. Manage. 53 (2020) 102104.
[55] Anton E. Lawson, The generality of hypothetico-deductive reasoning: Making scientific thinking explicit, Am. Biol. Teach. 62 (7) (2000) 482–495.
[56] Chonghua Xue, Ying Yu, Xin Huang, Comparison of organic matter properties and disinfection by-product formation between the typical groundwater and surface water, Water 14 (9) (2022) 1418.
[57] Yukun Hou, Wenhai Chu, Meng Ma, Carbonaceous and nitrogenous disinfection by-product formation in the surface and ground water treatment plants using Yellow River as water source, J. Environ. Sci. 24 (7) (2012) 1204–1209.
[58] R.K. Padhi, K.K. Satpathy, S. Subramanian, Impact of groundwater surface storage on chlorination and disinfection by-product formation, J. Water Heal. 13 (3) (2015) 838–847.
[59] Sowmya Chandrasekaran, Martin Zaefferer, Steffen Moritz, Jörg Stork, Martina Friese, Andreas Fischbach, Thomas Bartz-Beielstein, Data preprocessing: A new algorithm for univariate imputation designed specifically for industrial needs, in: Proceedings, vpl. 26, 2016, pp. 77–95.
[60] Gobiinda G. Chowdhury, How to improve the sustainability of digital libraries and information services? J. Assoc. Inf. Sci. Technol. 67 (10) (2016) 2379–2391.
[61] Emma Strubell, Ananya Ganesh, Andrew McCallum, Energy and policy considerations for modern deep learning research, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (09) 2020, pp. 13693–13696.
[62] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, Jeff Dean, Carbon emissions and large neural network training, 2021, arXiv preprint arXiv:2104.10350.
[63] David B Miklos, Christian Remy, Martin Jekel, Karl G Linden, Jörg E Drewes, Uwe Hübner, Evaluation of advanced oxidation processes for water and wastewater treatment–A critical review, Water Res. 139 (2018) 118–131.
[64] Daniel Dianchen Gang, Robert L. Segar Jr., Thomas E. Clevenger, Shankha K. Banerji, Using chlorine demand to predict TTHM and HAA9 formation, J.-Am. Water Work. Assoc. 94 (10) (2002) 76–86.
[65] Sajith Madhawa Premarathna, George Kastl, Ian Fisher, Arumugam Sathasivan, Model for halo-acetic acids formation in bulk water of water supply systems, Sci. Total Environ. 857 (2023) 159267.
[66] John Bratby, Coagulation and Flocculation in Water and Wastewater Treatment, IWA publishing, 2016.
[67] John C. Crittenden, R. Rhodes Trussell, David W. Hand, Kerry J. Howe, George Tchobanoglous, MWH's water treatment: principles and design, John Wiley & Sons, 2012.
[68] Diego Takashi Sato, Wesley Pacheco Calixto, Data and code for Hybrid machine learning model for disinfectant dosing in small-scale water treatment under data scarcity, 2025, http://dx.doi.org/10.5281/zenodo.15770106, Version 1.0. Zenodo.
[69] Henry H. Perritt, Law and the Information Superhighway, Wolters Kluwer, New York, NY, USA, 2001.
[70] J. Fawell, K. Bailey, J. Chilton, E. Dahi, L. Fewtrell, Y. Magara, Fluoride in Drinking-Water, vol. 1, World Health Organization, London, UK, 2006.
[71] World Health Organization, Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda, vol. 4, World Health Organization, Geneve, Switzerland, 2022.
[72] World Health Organization, Guidelines for Drinking-Water Quality, World Health Organization, Geneve, Switzerland, 2002.
[73] Jaime Garibay-Rodriguez, Vicente Rico-Ramirez, Jose M. Ponce-Ortega, Optimal water management in macroscopic systems under economic penalty scenarios, AIChE J. 63 (8) (2017) 3419–3441.
[74] Yiyi Chen, Xian Zhang, George Grekousis, Yuling Huang, Fanglin Hua, Zehan Pan, Ye Liu, Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (XGBoost) approach, J. Clean. Prod. 413 (2023) 137432.
[75] Jungsu Park, Woo Hyoung Lee, Keug Tae Kim, Cheol Young Park, Sanghun Lee, Tae-Young Heo, Interpretation of ensemble learning to predict water quality using explainable artificial intelligence, Sci. Total Environ. 832 (2022) 155070.
[76] Sina Moradi, Amr Omar, Zhuoyu Zhou, Anthony Agostino, Ziba Gandomkar, Heriberto Bustamante, Kaye Power, Rita Henderson, Greg Leslie, Forecasting and optimizing dual media filter performance via machine learning, Water Res. 235 (2023) 119874.
[77] Abdulrhman Fahmi Alali, Heavy metals removal from wastewater using nanoporous adsorbent: Separation analysis via machine learning model, Case Stud. Therm. Eng. 59 (2024) 104501.
[78] Sabanaz Peerzade, Pooja Kamat, Enhancing water quality prediction: a machine learning approach across diverse water environments, Water Qual. Res. J. 60 (1) (2025) 298–317.
[79] Mushtaque Ahmed Rahu, Abdul Fattah Chandio, Khursheed Aurangzeb, Sarang Karim, Musaed Alhussein, Muhammad Shahid Anwar, Toward design of internet of things and machine learning-enabled frameworks for analysis and prediction of water quality, IEEE Access 11 (2023) 101055–101086.
[80] Maria Alice Prado Cechinel, Juliana Neves, João Vitor Rios Fuck, Rodrigo Campos de Andrade, Nicolas Spogis, Humberto Gracher Riella, Natan Padoin, Cintia Soares, Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis, J. Water Process. Eng. 58 (2024) 104758.
[81] David Costa, Yared Bayissa, Kargean Vianna Barbosa, Mariana Dias Villas-Boas, Arun Bawa, Jader Lugon Junior, Antônio J. Silva Neto, Raghavan Srinivasan, Water quality estimates using machine learning techniques in an experimental watershed, J. Hydroinformatics 26 (11) (2024) 2798–2814.
[82] Ujala Ejaz, Shujaul Mulk Khan, Sadia Jehangir, Zeeshan Ahmad, Abdullah Abdullah, Majid Iqbal, Noreen Khalid, Aisha Nazir, Jens-Christian Svenning, Monitoring the industrial waste polluted stream-integrated analytics and machine learning for water quality index assessment, J. Clean. Prod. 450 (2024) 141877.
[83] Musiri Kailasanathan Nallakaruppan, E Gangadevi, M. Lawanya Shri, Balamurugan Balusamy, Sweta Bhattacharya, Shitharth Selvarajan, Reliable water quality prediction and parametric analysis using explainable AI models, Sci. Rep. 14 (1) (2024) 7520.
[84] Steve E. Hrudey, Chlorination disinfection by-products, public health risk tradeoffs and me, Water Res. 43 (8) (2009) 2057–2092.
[85] Shakhawat Chowdhury, Effects of seawater intrusion on the formation of disinfection byproducts in drinking water, Sci. Total Environ. 827 (2022) 154398.
[86] Hailong Cao, Xianjun Xie, Ziyi Xiao, Wenjing Liu, Transferability of machine learning models for geogenic contaminated groundwaters, Environ. Sci. Technol. 58 (20) (2024) 8783–8791.