Article:
Kyritsakas, G., Boxall, J. and Speight, V. (2024) A data-driven predictive model for disinfectant residual in drinking water storage tanks. AWWA Water Science, 6 (3). e1376. ISSN: 2577-8161

https://doi.org/10.1002/aws2.1376

Received: 19 January 2024 | Revised: 16 May 2024 | Accepted: 23 May 2024

DOI: 10.1002/aws2.1376

ORIGINAL RESEARCH

AWWA WATER SCIENCE

A data-driven predictive model for disinfectant residual in drinking water storage tanks

Grigorios Kyritsakas[1,2] | **Joby Boxall**[1] | **Vanessa Speight**[1]

[1]Sheffield Water Centre, University of Sheffield, Sheffield, UK

[2]Department of Sanitary Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Correspondence
Grigorios Kyritsakas, Sheffield Water Centre, University of Sheffield, Sheffield, South Yorkshire, UK.
Email: g.kyritsakas@sheffield.ac.uk

Funding information
EPSRC Centre for Doctoral Training in Engineering for the Water Sector, Grant/Award Number: STREAM IDC, EP/L015412/1; Scottish Water

Deputy Editor: Lauren Weinrich
Associate Editor: Juneseok Lee

Abstract

A data-driven approach is developed and proven for ranking the risk of low disinfection residual in water distribution storage tanks, 1 month ahead. The forecasting methodology uses water quality data collected from drinking water treatment plants, storage tank outlets, and rainfall data as inputs. This methodology was developed and tested with data from a water utility serving more than 5 million people. Results show high-risk category prediction accuracy of 75%–80%. Using a final year of unseen validation data, more than 90% of the storage tanks ranked in the top 20 by the forecasting methodology experienced low disinfectant residual in the following month. Storage tanks are critical water distribution system infrastructure that are currently managed reactively. The adoption of such readily transferable machine learning approaches enables direct proactive management strategies and efficient interventions that can help ensure drinking water quality.

KEYWORDS

disinfection residual, drinking water quality, ensemble decision trees, machine learning, monitoring, storage tanks

1 | INTRODUCTION

Disinfection of drinking water is typically the last stage of the treatment process before water reaches the drinking water distribution system (DWDS). Disinfection is critical for ensuring drinking water is free from bacterial contamination. A residual of the chosen disinfectant is usually retained in the final treated water to limit bacteriological regrowth and provide some protection against contamination events within the DWDS. The most commonly used disinfectant is free chlorine but there are systems where chloramines are used as the residual to reduce disinfection by-products (DBPs) formation (Mian et al., 2018). The concentration of disinfectant residual is generally highest in the water exiting the drinking water treatment plant (DWTP) and the disinfectant is consumed by reactions with water quality constituents as water travels through the DWDS. Disinfectant residual consumption is caused by chemical, biological, and physical reactions that occur in the bulk water and at the pipe wall, that couple with extended water residence time (or water age) to cause loss of residual (Al-Jasser, 2007; Vasconcelos et al., 1997).

Low disinfectant residual concentrations indicate excessive reactions over a short period of time or moderate reactions over an extended period have occurred and an increased risk of regrowth or from possible but rare contamination events. Water utilities are therefore required to monitor disinfectant residuals as part of regulatory compliance. This is typically achieved using

discrete samples taken from across the DWDS, from the DWTPs to storage tanks, and consumers' taps and although online monitoring is growing in application, it is still used less frequently than discrete sampling. In addition to disinfectant residual, other key water quality variables are monitored in different jurisdictions (e.g., iron, manganese, turbidity, coliform bacteria, heterotrophic plate counts), depending on the regulations.

Storage tanks, commonly known in the United Kingdom as Service Reservoirs (SRs), are fundamental components of the DWDS. They are used for reserving water when the water demand increases, for balancing the pressure fluctuations in the DWDS, for providing continuous supply to consumers when an event or a shutdown occur in some parts of the system, and for providing emergency demands for fire suppression (Brandt et al., 2017). The water residence time in SRs varies from a couple of hours to days depending on the size of the SR, the size of its related DWDS, and the associated demand. Increased SRs residence time can and does influence water quality, degrading the disinfection residual over time. In the worst case when microbial contaminants are also present in the storage tank, waterborne disease outbreaks and loss of life can result, for example, as happened in Gideon, MO (Clark et al., 1996; Craun & Calderon, 2001). Moreover, Ellis et al. (2018) found that the number of bacteriological failures at SR outlets was double the amount at DWTPs, indicating the elevated water quality risk that storage tanks can bring to the large populations that they serve. Therefore, consistent monitoring of the water quality exiting these infrastructure components is required to understand their performance and hence to maintain the water safety standards before reaching the consumers taps.

Water utilities use disinfectant residual sample data to provide assurance of the safety of drinking water and to prioritize interventions in the DWDS such as flushing or SR cleaning. However, this can only ever be a reactive approach and cannot preempt and prevent problems; once low chlorine events are detected, the risk of bacteriological contamination in the DWDS is already elevated. There is hence a need for the ability to predict low disinfectant residual events, particularly in SRs, for proactive protection of public health. There is a family of models that attempt to quantify our understanding of the mechanisms of chlorine decay and have accurately predicted water distribution system hydraulics and water quality for decades, provided they are well calibrated and maintained (Clark & Sivaganesan, 2002; Rossman et al., 1994; Speight & Boxall, 2015). These can be limited in applicability due to uncertainty and site-specific coefficients required, particularly by the uncertainty of reactions and interactions associated with the pipe wall (Vasconcelos et al., 1997). SR in contrast are dominated by bulk water reactions, with coefficients potentially estimated from jar tests. There is less understanding of the residence time distributions within SR and hence modeling water quality in storage is more complicated, sometimes requiring tools such as computational fluid dynamics modeling to be applied and well validated for a comprehensive prediction of disinfectant residual (Grayman et al., 2004). There is, however, sufficient historical water quality data available such that an alternative modeling approach to prediction of chlorine residuals could be usefully attempted by machine learning techniques.

Article Impact Statement

A machine learning model is developed for predicting low disinfection residual events in storage tanks 1 month in advance into future with 75-80% accuracy, using historical monitoring samples.

Machine learning (ML) is a type of artificial intelligence that develops algorithms based on statistical knowledge to understand patterns in data (MathWorks, 2016). ML algorithms use existing data (training dataset) to create and train models for either understanding data trends (unsupervised learning) or predicting trends based on new unseen data (supervised learning). Supervised learning techniques can be further split into regression and classification categories based on the type of the predictions required. Regression techniques are used to predict numerical values while classification techniques determine an output category (class), such as those used in email software to classify junk email. ML algorithms have been successfully applied in many different fields such as natural language processing, image recognition, finance, policing, and engineering (Jordan & Mitchell, 2015).

Data-driven models based on ML algorithms are trained using available data to connect inputs and outputs with no need for understanding or specifying the complex processes that occur between input and output. In drinking water quality applications, ML algorithms have been successfully used to predict high iron concentrations in DWDS (Kazemi et al., 2023; Mounce et al., 2017), short–term turbidity trends in water distribution trunk mains (Kazemi et al., 2018; Meyers et al., 2017), bacteriological contamination indicators in the DWDS (Mohammed et al., 2017), and factors that cause discoloration (Speight et al., 2019). For prediction of disinfectant residual, Gibbs et al. (2006) used artificial neural networks (ANNs) to understand the connections between chlorine decay at customers' taps and various

water quality parameters, and there have been some advances using various ML methods for the prediction of chlorine losses in the DWDS (Garcia et al., 2020; Kyritsakas et al., 2023; Xu et al., 2019). This growing body of knowledge demonstrates that data-driven models, especially those using ML methods, have the potential to transform the sparse water quality monitoring data collected from DWDS into useful information for water utilities. However, to the best of the authors' knowledge there is no data-driven research that has been conducted for investigating water quality deterioration in SRs, so this study represents an exploration of that gap.

This paper investigates the ability of ML algorithms to predict low chlorine events in SRs. A new methodology is proposed to inform utilities about SRs that are at high risk of experiencing low chlorine with a sufficient time to implement interventions. The outputs are SR risk rankings for the following month based on their relative probability of low chlorine events. This type of information enables water utilities to prioritize their interventions toward the high-risk SRs, transforming reactive management to a proactive practice.

2 | METHODOLOGY

A data-driven model was developed for the classification of SRs into "High-risk" or "Low-risk" classes using historical water quality data from SRs and DWTPs outlets. ML techniques were applied to data for a large water utility that operates multiple systems in the north of the United Kingdom. Ensemble decision tree ML algorithms were used for the classification and a comparison of results is made using performance metrics.

2.1 | Case study dataset and data preprocessing

The data used in this investigation was taken from a large corporate database of a utility located in the north of the United Kingdom. This utility serves water to more than 5.4 million people via complicated DWDS consisting of more than 250 DWTPs, more than 1000 SRs, and approximately 45,000 km of pipes. The disinfection type is mainly chlorination, but some systems have switched to chloramination due to DBPs presence. The dataset contained water quality data collected from the DWTPs and the SRs outlets, collected mostly for regulatory purposes during the period January 2012 to November 2021. Turbidity data measured in the raw water before reaching the DWTPs were also included. The dataset contained more than 450,000 SRs water quality samples, more than 330,000 DWTPs water quality samples, and more than 35,000 turbidity samples taken in the raw water before reaching the DWTPs.

In the United Kingdom, the regulations require four samples per month for every active SR to measure chlorine concentration (total and free), coliform bacteria, and heterotrophic plate counts (HPCs) at 22 and 37°C. There are no specific requirements regarding metals and other important water quality variables such as total organic carbon(TOC) and turbidity (DWI, 2020; DWQR, 2019). Detection of coliform bacteria in SRs water quality samples is rare (only 0.14% in this dataset) and therefore they were excluded from this analysis. Turbidity and the metals data at the SRs were also excluded as there was insufficient data. However, there were a large number of flow cytometry total and intact cells counts (FC_TCCs, FC_ICCs) and temperature data available, collected since January 2015. This data was included in the analysis. The DWTPs dataset variables that were selected for analysis were free chlorine, total chlorine, TOC, temperature, FC_TCCs, and FC_ICCs. Temperature, FC_TCCs, and FC_ICCs measurements in the DWTPs also started in January 2015.

Rainfall has been associated with bacteriological events in the DWDS (Curriero et al., 2001; Kumpel & Nelson, 2013). Therefore, in order to examine the influence of rainfall in the prediction model, rainfall data covering the complete area of supply for the examined period were collected from the Met Office gauging stations that were in a close proximity to the water utility's DWTPs and the SRs. (Met Office, 2021).

The raw dataset required preprocessing before being used in the predictive model. This involved three steps: association of the SRs with the other assets; identification of the low chlorine events; and selection of the temporal scale of the analysis.

2.1.1 | Association of service reservoirs with the other water infrastructure

The association between the SRs, the DWTPs that supplied them, and the raw water that fed the DWTPs was achieved using the water utilities asset management information. The association between each SR and DWTP and Met Office gauging station was achieved using a nearest neighbor Euclidian search. Given the geographically sparse nature of the case study utility's water systems (across a country scale), it is unlikely that a distance-based search would mismatch rainfall gauges with DWTPs or SRs to wrong systems so this approach was not verified beyond spot checks.

2.1.2 | Identification of low chlorine events

Low chlorine events were defined by samples where measured chlorine was below a certain threshold. This was different for chlorination and chloramination systems. A low chlorine event in the chlorination systems was defined as a sample where the free chlorine concentration was below 0.25 mg/L as the WHO guidelines and the utility's suggestion for chlorine concentration in the consumers taps is 0.2 mg/L (WHO, 2000). For the chloramination systems the minimum threshold was defined equal to 0.7 mg/L of total chlorine as per the utility's suggestion. Based on those two thresholds the low chlorine events were calculated separately for the chlorinated and the chloraminated systems.

2.1.3 | Selection of the temporal scale of the analysis

A monthly scale was selected reflecting the low frequency of low chlorine events and because a monthly prediction time-step gives just sufficient time for proactive interventions. All the variables used as inputs were therefore monthly averaged. The monthly sum of low chlorine events per SR was also calculated.

2.2 | Low chlorine event prediction model

The predictive model was designed using a supervised classification approach. Input and corresponding output class were required for training. The input data were the monthly averaged values of the water quality variables. The output class was the group that the SR belongs to, which could be either the no low chlorine event class (Low-Risk) or the event class (High-Risks). The high event class included all SRs that had at least one low chlorine event per month. The monthly scale design of the model implies that there was a monthly lag between the inputs (variables) and the outputs (classes). So, for example, for January 2012 inputs the corresponding output classes were those in February 2012.

The model used different ensemble decision trees as the ML algorithms for its prediction. The ensemble decision trees use multiple weak decision trees to improve the predictive model performance (Rokach, 2010). Ensemble decision trees were selected because of their "white-box" approach as they provide a human interpretable justification of the prediction results: a variable importance analysis and a class probability for each SR. Two main ensemble decision tree approaches were compared, random forest and boosting (Dietterich, 2000). The model used the inputs–outputs of the years 2012 to 2020 for training and then the year 2021 was used for testing its predictive performance. As the chlorinated and the chloraminated systems have different water quality behavior, the final input dataset was split into two subgroups, the chlorinated SRs and the chloraminated SRs. The model was trained and tested for these two groups separately. A simplified diagram of the model is shown in Figure 1.

2.3 | Input–output variables

The output variables for month n are the class that each SR belongs to, either High- or Low-Risk. The associated input variables are the monthly averaged values of the various water quality parameters and precipitation for the month n-1. In recognition that monthly average values obscure potentially useful information it was also decided to include and explore the influence on model accuracy of five further calculated variables. These were monthly standard deviation of free chlorine and the total chlorine per SR, the age of water exiting the SRs (given by the water utility, based on design flows and SR dimensions, as the sum of cascading SRs retention time), and the average total and free chlorine per SR in the previous year (n-12). For each output variable (high- or low-risk class) at month n, 21 different input variables were used. The input and output variables for a given month n are presented in Table 1.

A schematic that shows the inputs and outputs for a given month n is presented in Figure 2.

2.4 | Ensemble decision tree methods

The ensemble decision tree methods tested and compared for their performance in this work were classification random forest (CRF), the main sub-space algorithm, AdaBoost the main boosting algorithm, and two other boosting algorithms, LogitBoost and RusBoost (Breiman, 2001; Friedman et al., 2000; Seiffert et al., 2008).

CRF consists of a number of independent weak decision tree classifiers which contribute equally to the final decision (in classification each tree has one vote). In CRF, the split of a randomly selected sample of data at each weak tree node is made by one of the variables of a small randomly selected sample of the total variables. Boosting trees is an ensemble method that also creates a strong classifier from a number of weak decision trees but in this case the weak classifiers do not contribute equally to the final decision and are dependent to the previous ones. In AdaBoost (adaptive boosting), each new generated tree aims to improve the errors made by

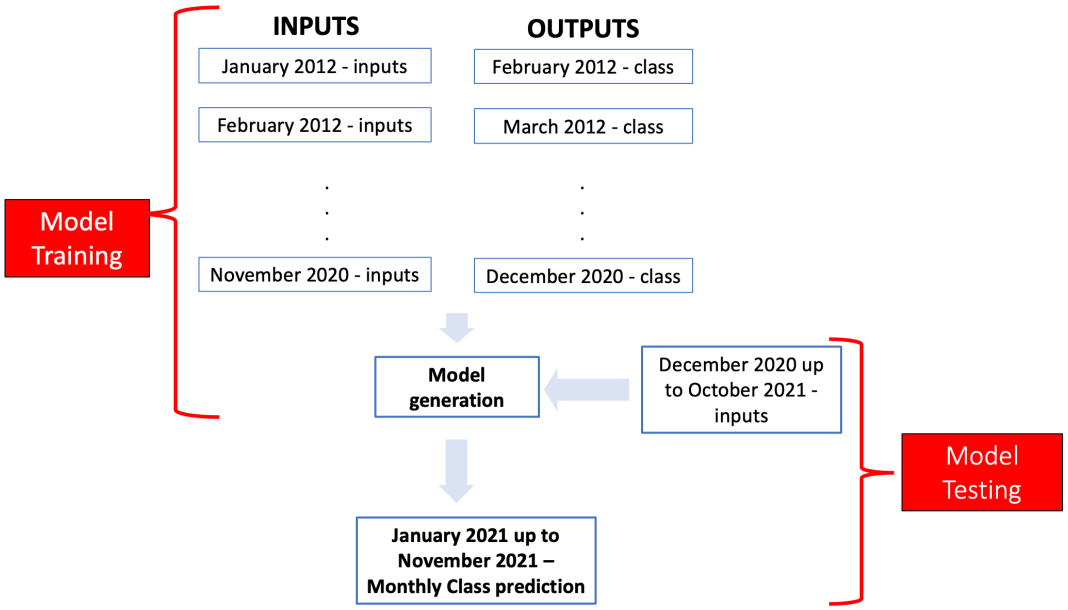

FIGURE 1 A schematic simplified diagram of the model implementation.

the previous tree by introducing weights in the misclassified data of the random sampling. Therefore, each new tree is dependent on the previous one. The final decision is made using the weighted contribution of each different tree. RusBoost (Random under-sampling) uses the AdaBoost algorithm in combination with a random under-sampling method to create more balanced datasets. Finally, LogitBoost (adaptive logistic boosting) use the Adaboost algorithm, but then applies the logistic regression cost function instead of minimizing the exponential loss. RF and Rusboost were successfully applied in two different research papers for the prediction of iron exceedance in district distribution areas (Kazemi et al., 2023; Mounce et al., 2017).

The predictive model was developed in MATLAB 2021b using the statistics and machine learning toolbox. The number of trees and the minimum number of observations per tree leaf for both the RF and the boosting methods were set to 1000 and 1, respectively. In RF, the number of randomly selected variables per node were set equal to square root of the total number of available variables as Breiman (2001), that is, when all 21 variables were used for training, the model randomly selects five variables from which the split decision in the node is made. In boosting, the learning rate was set equal to 0.1.

2.5 | Data bias and data augmentation methods

The dataset was heavily unbalanced toward the Low-Risk class for both the chlorination and the chloramination SRs as Table 2 indicates. In general, data bias could skew the training of the model and produce inaccurate predictions. To tackle this problem two types of methods were explored, oversampling the minority class (High-Risk) and under-sampling the majority class (Low-Risk). For the former approach, synthetic data (artificial data) of the Low-Risk class were required. Two different oversampling methods were used, the synthetic minority oversampling technique (SMOTE) and the adaptive synthetic method (ADASYN). SMOTE generates synthetic data by calculating the distance of the vector between a random sample of the minority class and some of its neighbor samples and then by multiplying the result with a random weight (Chawla et al., 2002). ADASYN follows the SMOTE approach but instead of selecting the minority sample randomly, the minority sample is selected based on the number of majority class samples in its neighbor, with high priority given to those surrounded by more majority class samples (He et al., 2008). As regard the under-sampling approach the method selected in this work was the random removal of samples belonging in the majority class.

The predictive model was initially tested using the four aforementioned ML algorithms (tests 1–4). Then, each augmentation method was used to reduce the Low-Risk/High-Risk bias and the predictive model was retrained using the updated dataset (tests 5–16). The level of bias reduction was matched to ML algorithms, in particular for RusBoost as it is designed for use with unbalanced datasets.

TABLE 1 Input and outputs of the predictive model for month n.

Variable category	Variable (unit)	Abbreviation	Source	Month
Input variables	Age of water exiting the SR (hrs)	Water age	Age of water exiting the SRs as calculated by the utility	*n*-1
	Average free chlorine concentration per SR (mg/l)	FreeCl	Water quality samples in the SR inlet (average of the measured samples)	*n*-1
	Free chlorine standard deviation per SR (mg/l)	FreeClstd		*n*-1
	Average total chlorine concentration per SR (mg/l)	TotalCl		*n*-1
	Total chlorine standard deviation (mg/l)	TotalClstd		*n*-1
	Average HPCs at 22°C (number of colonies)	HPC22		*n*-1
	Average HPCs at 37°C (number of colonies)	HPC37		*n*-1
	Average flow cytometry ICCs (number/ml)	FC_ICCs		*n*-1
	Average flow cytometry TCCs (number/ml)	FC_TCCs		*n*-1
	Average water temperature (°C)	Temperature		*n*-1
	Average free chlorine at the associated DWTP (mg/l)	FreeCl_DWTP	Water quality samples taken in the associated to the SR DWTP (average of the measured samples)	*n*-1
	Average total chlorine at the associated DWTP (mg/l)	TotalCl_DWTP		*n*-1
	Average flow cytometry ICCs at the associated DWTP (number/ml)	ICCs_DWTP		*n*-1
	Average flow cytometry TCCs at the associated DWTP (number/ml)	TCCs_DWTP		*n*-1
	Average water temperature at the associated DWTP (°C)	Temperature_DWTP		*n*-1
	Average TOC at the associated DWTP (mg/l)	TOC_DWTP		*n*-1
	Average turbidity at the source water associated with the SR (NTU)	Turbidity_Raw	Turbidity measured in the source water ssociated with the SR	*n*-1
	Daily average precipitation (mm)	SRdailyprecipitation	Calculated by the authors using the available monthly rainfall data in the area of the SR	*n*-1
	Daily average precipitation at the associated DWTP (mm)	DWTPdailyprecipitation	Calculated by the authors using the available monthly rainfall data in the area of the DWTP associated SR	*n*-1
	Average free chlorine concentration (mg/l)	FreeCl_1	Water quality samples taken in the SR a year before the investigation month *n*	*n*-12
	Average total chlorine concentration(mg/l)	TotalCl_1		*n*-12
Output variables	High-risk SR (−)	High-Risk	SR with at least one low-risk chlorine event[a]	*n*
	Low-risk SR (−)	Low-Risk	SR with at least one low-risk chlorine event	*n*

[a]Low0risk event: Sample with free chlorine<0.2 mg/L for chlorination and 0.7 mg/L for chloramination.

2.6 | Performance metrics

Metrics were used to evaluate the performance of the different ML algorithms. The simplest performance metric is Accuracy; however, this was not applied due to the disproportional number of Low-Risk samples compared to the High-Risk ones. Instead, three performance metrics were selected to cover different aspects of the model

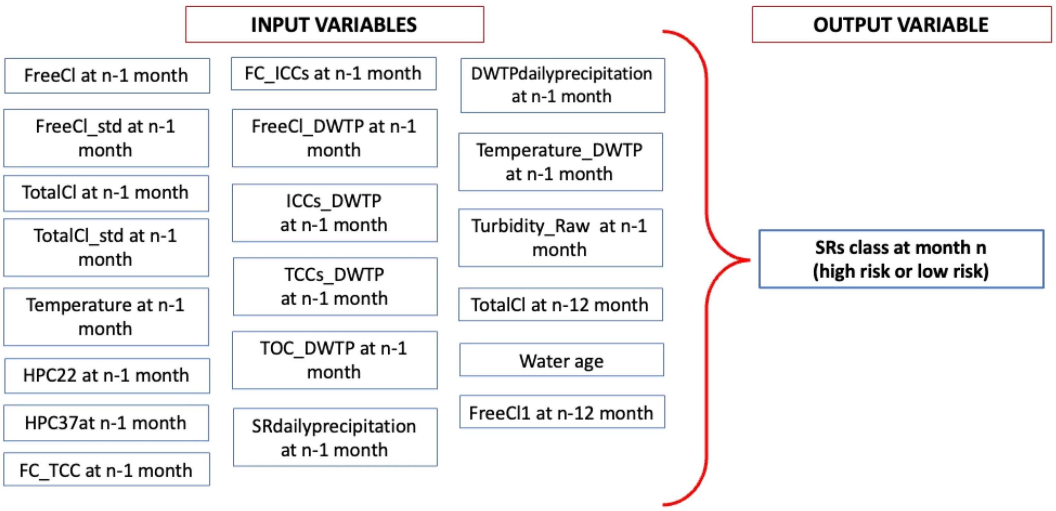

FIGURE 2 Schematic for inputs and outputs for a certain month n.

TABLE 2 Number and percentage of events in the monthly scaled dataset for January 2012–November 2021.

Disinfection type	No. of samples	No. of events	Percentage of events	Low-risk/High-risk
Chlorination	68,357	9824	15.2%	6.0
Chloramination	41,106	8430	20.5%	3.9

behavior: true positive rate (TPR), precision, and Matthews correlation coefficient (MCC). TPR is the rate of the correctly predicted events (true positives) and precision is the ratio of correct positives over the total number of predicted positives (true positives and false positives). These two metrics were used to evaluate the ability of predicting High-Risk SRs and the ability to generate a smaller number of false positives. MCC is a metric that quantifies the overall performance and balance of the model as it uses all four probable prediction results (true positives, true negatives, false positives, and false negatives) for its calculation (Baldi et al., 2000). MCC values lie between −1 and 1 with −1 indicating a complete disagreement between observations and predictions and 1 a completely agreement between observations and predictions. The formulas for these metrics are as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (1)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{X}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{X}\,\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (3)$$

where TP, TN, FP, FN are the True Positives (correctly predicted events), True Negatives (correctly predicted nonevents), False Positives (incorrectly predicted nonevents), False Negatives (incorrectly predicted events), respectively. Given that each metric provides information about a different aspect of model performance, the best models were selected using the average value of the three metrics. Using the average, we can see which model predicts the highest number of high-risk class SRs (high TPR) without creating a large number of false positives (high precision) and with good overall balanced performance (high MCC).

2.7 | Variables importance

To examine the importance of the various input variables on the model's performance, two different approaches were applied. Firstly, the models with higher than 0.7 performance metrics average were selected and rerun using only the variables indicated by the ML analysis as being the most important ones as indicated by the form of the decision trees (i.e., selected for inclusion in the top levels of the tree and thus scoring higher in the ensemble decision tree variable importance test during the training period). Secondly, the importance of each unique variable in the top two overall models was investigated by retraining the model iteratively without using one variable at a time, then comparing the MCC of the new test with the MCC of its corresponding model when all the variables were used.

TABLE 3 Prediction accuracy in the chlorinated systems for the year 2021.

Chlorinated systems						
			Performance metrics			
Test	ML model	Augmentation method	TPR	MCC	Precision	Average
Cl.1	RUSBoost	No	0.76	0.74	0.77	0.76
Cl.2	CRF		0.54	0.67	0.89	0.70
Cl.3	ADABoost		0.77	0.79	0.85	0.80
Cl.4	LogitBoost		0.55	0.67	0.89	0.70
Cl.5	RUSBoost	SMOTE	0.76	0.75	0.79	0.77
Cl.6	CRF		0.79	0.79	0.83	0.80
Cl.7	ADABoost		0.85	0.52	0.39	0.59
Cl.8	LogitBoost		0.8	0.75	0.76	0.77
Cl.9	RUSBoost	ADASYN	0.76	0.73	0.75	0.75
Cl.10	CRF		0.79	0.77	0.79	0.78
Cl.11	ADABoost		0.87	0.65	0.55	0.69
Cl.12	LogitBoost		0.78	0.75	0.76	0.76
Cl.13	RUSBoost	UNDER	0.78	0.54	0.46	0.59
Cl.14	CRF		0.83	0.72	0.68	0.74
Cl.15	ADABoost		0.91	0.46	0.32	0.56
Cl.16	LogitBoost		0.83	0.66	0.59	0.69

3 | RESULTS

3.1 | Model results

In the chlorinated systems, SMOTE and ADASYN were used to generate synthetic data equal to four times the amount of the minority class (Low-Risk class) data when ADABoost, LogitBoost, and CRF were used as ML algorithms, and equal to two times the Low-Class data when RusBoost was used. This is because RusBoost includes under-sampling method in its algorithm and, therefore, it should be applied in unbalanced datasets. In the chloraminated systems, there are more SRs included in the Low-Risk class and, thus, SMOTE and ADASYN were used to generate synthetic data equal to two times the minority class data. The random under-sampling approach was used to generate a completely balanced dataset to train ADABoost, LogitBoost, and CRF algorithms and to generate a dataset with a Low-class to High-class risk ratio equal to 2 to train the RusBoost algorithm.

The performance metrics of the 16 different initial tests in each one of the disinfection systems are presented in Tables 3 and 4. The model tests are named using the acronym Cl and Clm for the chlorination and the chloramination systems, respectively. For the chlorinated SRs, the best tests, based on the average performance metric (=0.8), were Cl.3 that uses AdaBoost with no augmentation method and Cl.6 that used CRF and SMOTE as an augmentation method. The former test has a higher MCC and Precision that indicates this is a more stable model and the latter test has a higher TPR performance. For the chloraminated systems, the best test was Clm.10 with a performance metrics average of 0.75. Overall, the tests in the chlorinated systems produced better results compared with those in the chloraminated systems as indicated by both the MCC and performance metrics.

In both types of systems, the AdaBoost tests with augmentation methods produced the highest TPR results reaching 91% TPR (Cl.7, Cl.11, Cl.15, Clm.7, Clm.11, Clm.15), however, these tests also had low MCC (0.46–0.66) and precision (0.32–0.6) values indicating that they also produced many false positives that reduced the model's stability. The RusBoost without augmentation (Cl.1 and Clm.1) performed better than all the other models in chloramination SRs and better than all except AdaBoost when no augmentation was used. However, RusBoost performance had a different behavior in the two disinfection systems when augmentation methods were introduced as its performance was increased in the chlorinated systems (Cl.5, Cl.9, Cl.13) and decreased in the chloraminated systems (Clm.5, Clm.9, Clm.13). CRF and LogitBoost had the lowest TPR performance in both disinfection systems when no augmentation method was used, but when SMOTE and ADASYN were introduced,

TABLE 4 Prediction accuracy in the chloraminated systems for the year 2021.

Chloraminated systems						
			Performance metrics			
Test	ML model	Augmentation method	TPR	MCC	Precision	Average
Clm.1	RUSBoost	No	0.73	0.7	0.73	0.72
Clm.2	CRF		0.62	0.69	0.83	0.71
Clm.3	ADABoost		0.64	0.67	0.77	0.69
Clm.4	LogitBoost		0.6	0.65	0.79	0.68
Clm.5	RUSBoost	SMOTE	0.73	0.69	0.71	0.71
Clm.6	CRF		0.74	0.71	0.75	0.73
Clm.7	ADABoost		0.81	0.48	0.38	0.56
Clm.8	LogitBoost		0.73	0.52	0.46	0.57
Clm.9	RUSBoost	ADASYN	0.7	0.68	0.73	0.70
Clm.10	CRF		0.76	0.73	0.76	0.75
Clm.11	ADABoost		0.84	0.49	0.39	0.57
Clm.12	LogitBoost		0.74	0.53	0.48	0.58
Clm.13	RUSBoost	UNDER	0.76	0.69	0.69	0.71
Clm.14	CRF		0.83	0.7	0.65	0.73
Clm.15	ADABoost		0.83	0.66	0.6	0.70
Clm.16	LogitBoost		0.79	0.61	0.55	0.65

their TPR performance was increased that also improved their average performance between 4% to 8%. However, these two algorithms had different behavior when the under-sampling method was used as the CRF increased its average performance in both systems and the LogitBoost decreased its own one.

3.2 | Variable importance

A critical element of this study was to understand the need for monitoring of different water quality parameters and thus an analysis of variable importance was performed. In the chlorinated systems, three CRF models (Cl.6, Cl.10, Cl.14), three RusBoost models (Cl.1, Cl.5, Cl.9), one AdaBoost model (Cl.3), and two LogitBoost models (Cl.8 and Cl.12) were selected to be retrained using only their most important variables as indicated by the form of the decision trees. In the chloraminated systems, the selected top performing models were three CRF models (Clm.6, Clm.10, and Clm.14) and three RusBoost models (Clm.1, Clm.5, Clm.13). The variables that were important and their relative importance was different in each of these models. Free chlorine was consistently one of the most important variables in the chlorinated system models and total chlorine for the chloraminated systems. Examples of the variables' importance in two different tests, test Cl.3 and test Cl.6, are shown in Figure 3. This graph shows that the contribution of monthly average free chlorine (FreeCl) in the Cl.3 model was more significant than any other variable. Comparison of the two graphs in Figure 3 clearly shows how the variables' contribution and importance differ from model to model. For example, in Cl.3 the Temperature_DWTP variable is the second most important variable but in Cl.6 this specific variable is one of the least important contributors.

The retraining of these models was made using only the variables that had a significant contribution relative to other variables. Thus, the number of variables differ in the new tests. Our approach was to include a minimum of four important variables in the second batch of tests even if the importance of one variable was far higher than the other variables. We set the maximum number of variables equal to 10 if the importance of these variables was equivalent. The new tests' results are presented in Table 5 (chlorinated systems) and Table 6 (chloraminated systems). In the variables' column, the variables used for the training are shown with their order of importance in the initial tests (where all variables were used).

These new tests had worse average performance results in comparison to their equivalent initial ones (all variables used) except for test Cl.25 that performed better than Cl.14 and Clm.18 that had equal performance with test Clm.6. The worst performance drop appeared in the

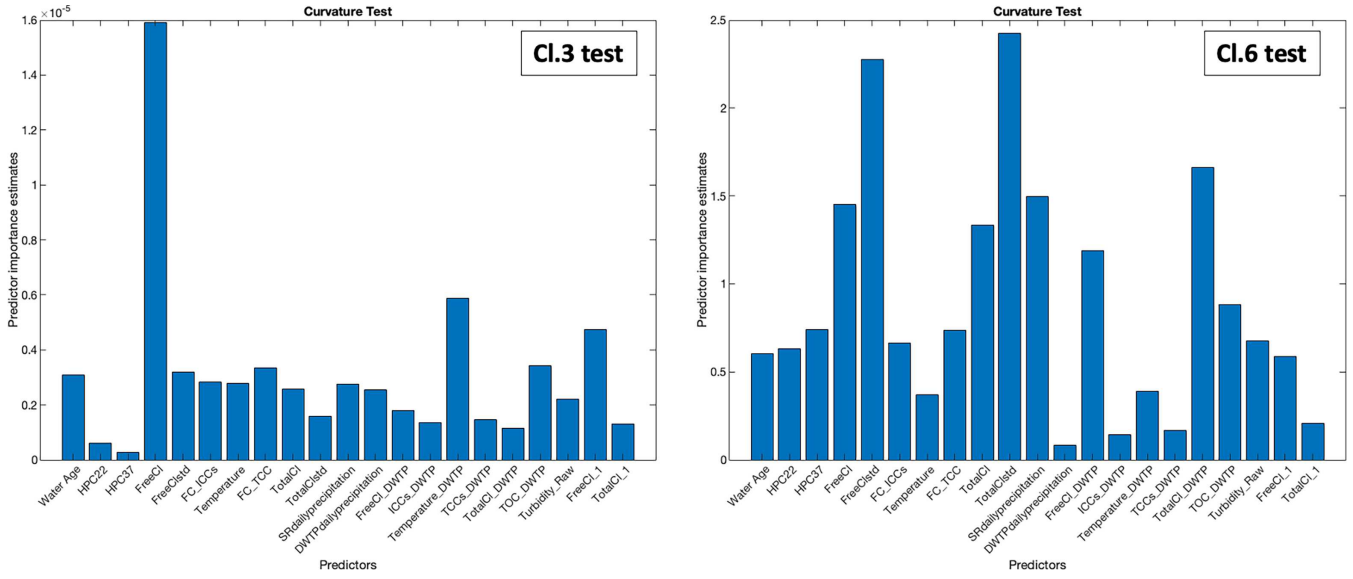

FIGURE 3 Variables' importance for tests Cl.3 (left) and Cl.6 (right). The importance estimates are calculated using the decrease in node impurity at each tree.

TABLE 5 Prediction accuracy with reduced number of variables for the year 2021 in the chlorinated systems.

Chlorinated systems							
				Performance metrics			
Test	Relevant initial test	ML model	Variables	TPR	MCC	Precision	Average
Cl.17	Cl.1	RUSBoost	Cl/Clstd/Temperature_DWTP/Water age	0.78	0.67	0.64	0.70
Cl.18	Cl.3	ADABoost	Cl/Temperature_DWTP/Cl_1/TOC_DWTP/FC_TCC	0.76	0.78	0.84	0.79
Cl.19	Cl.5	RUSBoost	Cl/Cl_1/Turbidity_raw/Water Age	0.74	0.67	0.67	0.69
Cl.20	Cl.6	CRF	Cl/Clstd/TotCl_std/ SRdailyprecipitation/TotCl_DWTP	0.79	0.72	0.72	0.74
Cl.21	Cl.8	LogitBoost	TOC_DWTP/Turbidity_Raw/SRdailyprecipitation/ FreeCl_DWTP	0.76	0.29	0.22	0.42
Cl.22	Cl.9	RUSBoost	Cl/Cl_1/Turbidity_raw/Water Age	0.74	0.65	0.63	0.67
Cl.23	Cl.10	CRF	Cl/Clstd/ TotClstd/SRdailyprecipitation/TotCl_DWTP	0.8	0.64	0.59	0.68
Cl.24	Cl.12	LogitBoost	Turbidity_Raw/TOC_DWTP/SRdailyprecipitation/ FreeCl_DWTP/ TotalCl_DWTP	0.76	0.27	0.2	0.41
Cl.25	Cl.14	CRF	Cl/TotClstd/ Clstd/TotCl_DWTP/TOC_DWTP/ TotCl/ Cl_1/DWTPdailyprecipittion	0.82	0.75	0.73	0.77

LogitBoost models where their performance was reduced by up to 26% (Cl.8-Cl.21, Cl.12-Cl.24). This drop could be, potentially, explained by the fact that in both LogitBoost tests the free chlorine variable (Cl) is absent. However, in most of the cases the performance drop was insignificant which indicates that the model could be applied with less input variables, and consequently with less computational effort for data preprocessing and training, and still produce accurate results (Cl.18, Cl.20, Clm.19, Clm.20, Clm.22).

In the chlorination systems, the above results indicated that the Adaboost without any data augmentation tests (Cl.3 becoming Cl.18) had the best performance. It provided the best performance without requiring any additional computational cost for balancing the training dataset, which is the case for test Cl.6 where SMOTE augmentation is required. In terms of computational cost and data preprocessing time, Cl.18 could be considered the best model as it provided insignificantly worse results (an average of just 1%) comparing to Cl.3 but uses only

TABLE 6 Prediction accuracy with reduced number of variables for the year 2021 in the chloraminated systems.

Chloraminated systems							
				Performance metrics			
Test	Relevant initial test	ML model	Vaiables	TPR	MCC	Precision	Average
Clm.17	CLm.1	RUSBoost	TotCl/ TotCl_1/Temperature/Temperature_DWTP/ Turbidity_Raw	0.74	0.69	0.71	0.71
Clm.18	Clm.6	CRF	TotCl/TotCl_DWTP/TOC_DWTP/TotCl_1/ DWTPdailyprecipitation/TCC_DWTP/Water age	0.82	0.7	0.66	0.73
Clm.19	Clm.10	CRF	TotCl/TotCl_DWTP/TCC_DWTP/TotCl_1/ DWTPdailyprecipitation/TOC_DWTP/Water age	0.83	0.7	0.63	0.72
Clm.20	Clm.14	CRF	TotCl/TotCl_1/TotClstd/DWTPdailyprecipitation/Water Age/TOC_DWTP/Srdailyprecipitation	0.84	0.68	0.62	0.71
Clm.21	Clm.5	RUSBoost	TotCl/TotClstd/Temperature	0.75	0.63	0.6	0.66
Clm.22	Clm.13	RUSBoost	TotCl/ TotalCl_1/Turbidity_Raw/Temperature_DWTP/ Temperature/ Water Age	0.75	0.67	0.67	0.70

TABLE 7 Importance of each variable in the predicting model best tests for chlorinated (left) and chloraminated systems (right).

Chlorinated sytems			Chloraminated sytems		
Excluded variable	Cl.3 MCC decrease	Cl.18	Excluded variable	Clm.6 MCC decrease	Clm.18
Water age	0.005	-	Water age	−0.028	−0.028
HPC22	0.005	-	HPC22	−0.038	-
HPC37	−0.002	-	HPC37	−0.032	-
FreeCl	−0.126	−0.167	FreeCl	−0.037	-
FreeClstd	0.007	-	FreeClstd	−0.033	-
FC_ICCs	−0.008	-	FC_ICCs	−0.029	-
Temperature	0.009	-	Temperature	−0.035	-
FC_TCC	0.004	-	FC_TCC	−0.027	-
TotalCl	−0.008	-	TotalCl	−0.100	−0.123
TotalClstd	0.003	-	TotalClstd	−0.032	-
SRdailyprecipitation	0.008	-	SRdailyprecipitation	−0.029	-
DWTPdailyprecipitation	−0.019	-	DWTPdailyprecipitation	−0.027	−0.021
FreeCl_DWTP	0.011	-	FreeCl_DWTP	−0.032	-
ICCs_DWTP	0.013	-	ICCs_DWTP	−0.029	-
Temperature_DWTP	0.003	0.007	Temperature_DWTP	−0.031	-
TCCs_DWTP	−0.001	−0.028	TCCs_DWTP	−0.033	−0.037
TotalCl_DWTP	0.001	-	TotalCl_DWTP	−0.031	−0.030
TOC_DWTP	0.007	−0.006	TOC_DWTP	−0.035	−0.035
Turbidity_Raw	0.016	-	Turbidity_Raw	−0.016	-
FreeCl_1	0.007	−0.033	FreeCl_1	−0.038	-
TotalCl_1	0.006	-	TotalCl_1	−0.049	−0.049

five input variables in the training period. For the chloramination systems, the best test was Clm.10 with a performance average of 0.75. There were three other models with a performance metrics average of 0.73 (Clm.6, Clm.14, Clm.18) with Clm.18 being considered better than the other two as it required fewer input variables.

A sensitivity analysis was conducted based on the best four tests (Cl.3, Cl.18, Clm.6, Clm.18) by retraining them with one permutated variable at a time. For each new model, the difference between the MCC value of each retrained model and the MCC value of the initial test was calculated. Thus, the larger the negative MCC difference the higher the variable's contribution in the model. A positive difference indicates that this variable has negative influence and should be removed from the input dataset. The results of this process are shown in Table 7.

For all the CRF tests (chloraminated systems Clm.6, Clm.18) each unique variable has some contribution in the model's prediction, as Table 7 shows. However, for the Adaboost model Cl.3 there were variables that their absence increased the model's MCC indicating that they could be excluded from the training input. These variables were the turbidity in the raw water (1.6% improvement), the free chlorine in the DWTPs (1.1% improvement), and the ICCs in the DWTPs (1.3% improvement). In chlorinated systems, free chlorine was the most important variable for all the tests as its absence reduced the MCC performance by 13% for Cl.3 and 17% for Cl.18. In chloraminated systems, total chlorine was the most important variable as its absence reduced the MCC performance by 10% for Clm.6 and 12% for Clm.18. Overall, the average MCC drop in the CRF tests was 4% and in the AdaBoost tests was 1%. This indicates that apart from the chlorine variables (free for chlorination, total for chloramination) the models could still produce reliable results when one of the other input variables is not available.

3.3 | Refining the chlorine model

The variables' importance test for the Cl.3 indicated that AdaBoost with no augmentation could improve its performance if one of the FreeCl_DWTP, the ICC_DWTP, or the Turbidity_Raw is removed. Therefore, the performance results of these three tests are presented in Table 8. In the same table, the performance results of another test where all three variables were removed (Cl.29 test) is presented.

Table 8 indicates that the overall average performance was improved when one of these variables was permutated. In addition, these three tests improved each one of the performance metrics with the Cl.26 test having the best overall performance improvement. However, there was no performance improvement when all three of them were permutated (Cl.29). Cl.26 was the best out of all these tests and overall it was the best out of all the predictive model tests implemented in the chlorinated systems SRs.

3.4 | Validation of model results using risk ranking

To verify the model's performance, data for January through November 2021 was used to predict SR low residual events for the following month, and this result was compared to actual system measurements. Table 9 shows the number of the correctly predicted Low-Risk SRs and High-Risk SRs for the best two models for both disinfection types. Note that this table shows the overall results for all the SRs, that is, there are 11 different class predictions for each SR, one for each month.

Ensemble decision trees produce in their outputs a probability of an SR being in one of the two classes based on the number of trees in the ensemble that belongs to one of them. Therefore, the larger the number of trees classifying an SR in the High-Risk class the higher the probability of low residual for that SR. The risk ranking for each SR per month was implemented using the best predictive models for each disinfection system which were, as Table 9 demonstrates, Cl.26 for the chlorination SRs and Clm.10 for the chloramination SRs. Table 10 shows the number of the top-20 highest risk ranked SRs, as ranked by the predictive model, that actually experienced low residual at that month during the investigation period. We should bear in mind here, that the top-20 SRs' risk ranking list contains different SRs at each month of the investigation period.

The model performance (Table 10) was impacted by months with relatively few (lower than 20) low residual events, such as March, April, and June. The arbitrary selection of 20 top high-risk SRs might not be appropriate for triggering interventions in all cases and could perhaps be modified to better fit the actual occurrence of low residual events over time.

3.5 | Comparing the best performing models with the last months measurement approach

A final comparison between the best predictive models and the "last-month" approach that is commonly used by water utilities to prioritize interventions in SRs has been made. For this comparison, an assumption that an SR that has failed in the last month will also fail in the following month has been made (i.e., an SR that fails in the month of February will also fail in the following month) and the same performance metrics were used. The results of the "last month" for both the chlorination and the chloramination SRs are presented in the following table (Table 11). These results clearly indicate that this approach performance is worse than the performance of

TABLE 8 AdaBoost model with reduced number of variables.

Chlorinated systems									
						Performance metrics			
Test	ML model	Augmentation method	Total number of training samples	High risk class percentage	Number of parameters	Parameters	TPR	MCC	Precision	Average
Cl.26	AdaBoost	No	57,947	15.7%	20	All except Turbidity_Raw	0.78	0.81	0.87	0.82
Cl.27	AdaBoost	No	57,947	15.7%	20	All except ICCs_WTW	0.78	0.8	0.86	0.81
Cl.28	AdaBoost	No	57,947	15.7%	20	All except FreeCl_WTW	0.78	0.8	0.86	0.81
Cl.29	AdaBoost	No	57,947	15.7%	18	All except Turbidity_Raw, ICCs_WTW, FreeCl_WTW	0.77	0.79	0.84	0.80
Cl.30	AdaBoost	No	57,947	15.7%	19	All except Turbidity_Raw, ICCs_WTW	0.78	0.79	0.84	0.80

all the different tests of the model presented in Sections 3.1 and 3.2.

4 | DISCUSSION

4.1 | Machine learning to predict low residual events in service reservoirs

The ML based predictive models for both the chlorination and the chloramination SRs provided high levels of performance with only a few of the tests having unacceptable average performance metrics. The best chlorination model was found to be one that used the AdaBoost ML algorithm with 20 out of the 21 available variables (only turbidity in the raw water was excluded), without the use of an augmentation method to account for data bias. The best chloramination model was the one that used a CRF ML algorithm with all 21 available variables and the ADASYN augmentation method.

Overall CRF based models outperformed the other algorithms. There was just one CRF test with an average performance below 0.7. Regarding the other models, RusBoost was the second-best overall algorithm, AdaBoost was the third, and LogitBoost the last. AdaBoost precision performance was significantly dropped (up to 50%) when augmentation methods were used, a finding which indicates that, when the test dataset is unbalanced, generating a fully balanced training dataset using augmentation methods misleads AdaBoost algorithm training into predicting many false positives. A similar behavior was noticed for the LogitBoost algorithm which also increased its false positive predictions when the training imbalance was reduced using augmentation methods. In addition, this algorithm's performance dropped significantly when the input variables were reduced and for this reason LogitBoost was the worse algorithm for this water quality problem.

Water distribution systems are complex reactors for water quality with a number of competing and ill-defined processes taking place concurrently. A prediction with accuracy of approximately 80% is significantly better than most other predictive approaches to modeling water quality in water storage tanks.

4.2 | Input variables

The importance of the variables was different for each model, with free chlorine and total chlorine measured in the storage tank in the previous month being consistently one of the most significant variables in the chlorination and chloramination systems, respectively. Intact cell

TABLE 9 Best models service reservoir predictions for the testing period (January 2021–November 2021).

Chlorinated systems			Chloraminated systems		
Test	Correctly predicted high-risk SRs/Total number of high risk SRs	Correctly predicted low-risk SRs/Total number of low-risk SRs	Test	Correctly predicted high-risk SRs/Total number of high-risk SRs	Correctly predicted low-risk SRs/Total number of low-risk SRs
Cl.26	475/606	5602/5676	Clm.10	333/441	3527/3635
Cl.18	457/606	5593/5676	Clm.18	361/441	3450/3635

TABLE 10 Number of service reservoirs (SRs) with low residual events belonging in the predicted models' top-20 high risk SRs per month, during the investigation period (January 2021–November 2021).

Month	Cl.26	Clm.10
January	20	17
February	18	12[a]
March	0[b]	0[b]
April	20	17
May	20	20
June	20	17
July	20	18
August	20	20
September	20	20
October	20	20
November	20	20

[a]February 2021 had 13 low chloramination events.
[b]March 2021 had zero low residual events in both systems.

TABLE 11 Last month approach performance metrics.

Disinfection	TPR	Precision	MCC	Average
Chlorination SRs	1	0.25	0.42	0.56
Chloramination SRs	1	0.3	0.48	0.59

counts, raw water turbidity, and disinfectant residual at the WTP all contributed to improving the prediction, but their impact was less significant and the exclusion of one or more of these additional would still result in a valid and useful model for this particular dataset. This finding emphasizes the need to perform disinfectant residual monitoring at storage tanks with even weekly grab samples being sufficient to support this type of predictive model.

The tests exploring permutations of the input variables indicated that all the variables have some contribution to the model's decision trees and, thus, it could be suggested that the more variables are included, the better the model's performance could be. Given the number of different potential causes of low chlorine events in a specific SR, including for example low chlorine in DWTP finished water, changes in water age, and elevated temperature, the performance of the ML approach for prediction utility-wide in over 1000 SRs is remarkable. The ability of the top performing ML models to make quite accurate predictions despite lacking other information about the cause (or requirement for the user to presuppose a cause) is a key strength of the approach. It is likely that different model permutations are predicting certain types of failures better than others, which could also explain why certain parameters (e.g., raw water turbidity) that could plausibly have an impact have not played a significant role in these models because such root causes of SR low chlorine events are not as common in the dataset.

Future work could therefore extend the variables considered here, for example, to include TOC and metals data that were not available for this research, or to consider subsets of the SR assets grouped geographically. This future work should not necessarily be restricted to only data collected and managed by the water utility. Rainfall data was found to contribute to all models, but with significant variation across the different model formulations. Overfitting to input parameters is also a possibility in these types of ML modeling approaches, thus the importance analysis is a critical step to include in any such analysis.

The analysis showed that flow cytometry was a useful input variable. But this is not a parameter required by regulations so may not be widely available at DWTP or SRs. It is important to note that usefully accurate predictions were not dependent on the availability of flow cytometry parameters.

4.3 | Practical implications

One of the advantages of ensemble decision trees is that they offer human interpretable insight. Each weak decision tree of the ensemble algorithm can be extracted as an image. This example shows the split criteria of the data in the first two leaves of this decision tree. Operators and decision-makers can examine such information over

a number of the weak decision trees to gain an understanding of the reasons that the predictive model made a certain classification decision. The risk ranking list that the predictive model generates could be used to direct the water utilities interventions in the SRs that consistently fail. However, it is important to recognize that these types of ML models do not provide details about the root cause of each individual predicted SR low chlorine event. A further analysis of the causes of disinfectant residual deterioration would be required to identify the factors that that caused it, or development of a model with root cause identification as its goal.

This research does not definitively recommend one particular model for all utilities to predict low residual events as the results indicated that there will always be a trade-off between increasing the TPR performance and decreasing the MCC index and precision by generating false positives. The optimal trade between these metrics is heavily dependent on the requirements of the proactive management approach that water utilities would like to follow and should be decided on a managerial level, based on the available financial sources for interventions.

Collecting and integrating the various data to create the final dataset used for the predictive model required significant effort and collaboration between the authors and the utility. However, once the dataset was completed, the required time to train the model with a different ML algorithm was less than 10 min. Hence, the investment and change required to unlock the potential of these types of data-driven applications, to better manage drinking water quality and more, is in better storage, linkage, and accessibility of existing data.

Classification approaches, like risk ranking, are a useful outcome for utilities in the management of water quality. Based on weekly grab samples from tanks, the ability to precisely predict a numerical disinfectant residual in the following month would be limited. However, the utility does not need a precise numerical prediction of disinfectant residual to take action, an indication of high risk for low residual is sufficient. A list of the top 20 highest risk SRs is helpful in prioritizing maintenance activities. Many utilities use historical events or previous month's measurements as the prioritization criteria for current interventions. However, it has been demonstrated that machine learning methods like the one used in this study can perform better than the historical event approach (Kyritsakas et al., 2023).

This paper presents a data-driven predictive model that uses only available water quality monitoring data so there was no requirement for hydraulic or other mechanistic models or the calibration, etc. associated with their use. These machine learning approaches (random forest and boosting selected here) can be transferred and applied to other utilities if there are sufficient discrete water quality monitoring data over a time scale suitable to capture past events, which are critical for learning aspects of the approach. Similarly, these types of machine learning approaches have been demonstrated as valuable for prediction and risk ranking of other drinking water quality parameters beyond disinfectant residual. There is significant potential for machine learning methods to mine useful and actionable information from historical drinking water quality data.

5 | CONCLUSIONS

This paper demonstrated the ability of data-driven methodologies to predict SRs disinfection residual risk class 1 month in advance. Water quality data collected from different parts of the DWDS and rainfall data were utilized as inputs in the model, and different machine learning methodologies and formulations were applied to decrease the imbalance of the dataset (most of the SRs belonged in the low-risk class) and to identify the key variables that yield better predictions. Based on the results the key findings are:

- The predictive model reached an overall average (of True positive rate, Precision, and Matthews correlation coefficient) performance of 0.82 for the chlorination SRs and 0.76 for the chlorination SRs.
- The importance of different input variables, and their combination, was complex and varied across the different model formulations explored. The selection methodology identified the input variables required to produce results with minimal performance drop and decreased computational time.
- The input variables sensitivity analysis indicated that free chlorine and total chlorine are consistently one of the most important input variables in the chlorinated and the chloraminated systems, respectively.

The month ahead risk ranking outputs of the model are well suited to enabling proactive management by providing sufficient early warning to arrange for additional sampling, flushing, cleaning, or other interventions for the highest ranked SRs. The Machine Learning approaches applied are of the “open box” type, hence the form of the decision trees can be interrogated to gain deeper understanding of the contributing factors and mechanisms that might have contributed to water quality deterioration through the parameter importance analysis, further targeting strategic management and decision-making. The data-driven nature of the approach means that the methodology is generic; it could be readily

applied to the SRs of other water utilities or different areas of the DWDS depending on data availability. Methodologies like the one presented are an important first step on the pathway toward the Digital Water era.

AUTHOR CONTRIBUTIONS

Grigorios Kyritsakas: Conceptualization; data curation; visualization; methodology; writing – original draft. **Joby Boxall:** Conceptualization; supervision; writing – review and editing. **Vanessa Speight:** Conceptualization; supervision; funding acquisition; writing – review and editing.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Claire Thom, Graeme Moore at Scottish Water for the data collection, their input, and assistance. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) license to any author accepted manuscript versions arising.

FUNDING INFORMATION

This work was funded by the EPSRC Centre for Doctoral Training in Engineering for the Water Sector (STREAM IDC, EP/L015412/1) and Scottish Water.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data analysed in this study is subject to the following licenses/restrictions: These data belong to a water utility and cannot be shared publicly. Requests to access these datasets should be directed to g.kyritsakas@sheffield.ac.uk.

ORCID

Grigorios Kyritsakas https://orcid.org/0000-0003-0945-3754
Joby Boxall https://orcid.org/0000-0002-4681-6895
Vanessa Speight https://orcid.org/0000-0001-7780-7863

REFERENCES

Al-Jasser, A. O. (2007). Chlorine decay in drinking-water transmission and distribution systems: Pipe service age effect. *Water Research*, *41*(2), 387–396. https://doi.org/10.1016/j.watres.2006.08.032

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, *16*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Brandt, M. J., Michael Johnson, K., Elphinston, A. J., & Ratnayak, D. D. (2017). *Twort's water supply* (7th ed.). Butterworth-Heinemann. ISDN: 978–0–08-100025-0.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.3390/rs10060911

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Clark, R. M., Geldreich, E. E., Fox, K. R., Rice, E. W., Johnson, C. H., Goodrich, J. A., Barnick, J. A., Abdesaken, F., Hill, J. E., & Angulo, F. J. (1996). A waterborne salmonella typhimurium outbreak in Gideon, Missouri: Results from a field investigation. *International Journal of Environmental Health Research*, *6*(3), 187–193. https://doi.org/10.1080/09603129609356889

Clark, R. M., & Sivaganesan, M. (2002). Predicting chlorine residuals in drinking water: Second order model. *Journal of Water Resources Planning and Management*, *128*(2), 152–161. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(152)

Craun, G. F., & Calderon, R. L. (2001). Waterborne disease outbreaks caused by distribution system deficiencies. *Journal - American Water Works Association*, *93*(9), 64–75. https://doi.org/10.1002/j.1551-8833.2001.tb09287.x

Curriero, F. C., Patz, J. A., Rose, J. B., & Lele, S. (2001). The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994. *American Journal of Public Health*, *91*(8), 1194–1199. https://doi.org/10.2105/AJPH.91.8.1194

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees. *Machine Learning*, *40*, 139–157. https://doi.org/10.1023/A:1007607513941

DWI. (2020). Drinking Water 2020: The Chief Inspector's Report for Drinking Water in England.

DWQR. (2019). Drinking water quality in Scotland 2018: Public water supply.

Ellis, K., Gowdy, C., Jakomis, N., Ryan, B., Thom, C., Biggs, C. A., & Speight, V. (2018). Understanding the costs of investigating coliform and *E. coli* detections during routine drinking water quality monitoring. *Urban Water Journal*, *15*(2), 101–108. https://doi.org/10.1080/1573062X.2017.1398762

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *28*(2), 337–374. https://doi.org/10.1214/aos/1016218223

Garcia, D., Puig, V., & Quevedo, J. (2020). Prognosis of water quality sensors using advanced data analytics: Application to the Barcelona drinking water network. *Sensors (Switzerland)*, *20*(5), 1342. https://doi.org/10.3390/s20051342

Gibbs, M. S., Morgan, N., Maier, H. R., Dandy, G. C., Nixon, J. B., & Holmes, M. (2006). Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. *Mathematical and Computer Modelling*, *44*(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Grayman, W. M., Rossman, L. A., Deininger, R. A., Smith, C. D., Arnold, C. N., & Smith, J. F. (2004). Mixing and aging of water in distribution system storage facilities. *JAWWA*, *96*(9), 70–80. https://doi.org/10.1002/j.1551-8833.2004.tb10704.x

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322–1328). IEEE.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kazemi, E., Kyritsakas, G., Husband, S., Flavell, K., Speight, V., & Boxall, J. (2023). Predicting iron exceedance risk in drinking water distribution systems using machine learning. *IOP Conference Series: Earth and Environmental Science*, *1136*, 012047. https://doi.org/10.1088/1755-1315/1136/1/012047

Kazemi, E., Mounce, S., Husband, S., & Boxall, J. (2018). Predicting turbidity in water distribution trunk mains using nonlinear autoregressive exogenous artificial neural networks. In *Proceeding of 13th international conference on Hydroinformatics*. Palermo, Italy: IWA.

Kumpel, E., & Nelson, K. L. (2013). Comparing microbial water quality in an intermittent and continuous piped water supply. *Water Research*, *47*(14), 5176–5188. https://doi.org/10.1016/j.watres.2013.05.058

Kyritsakas, G., Speight, V., & Boxall, J. (2023). A data-driven model for the prediction of chlorine losses in water distribution trunk mains. *IOP Conference Series: Earth and Environmental Science*, *1136*, 012048. https://doi.org/10.1088/1755-1315/1136/1/012048

MathWorks. (2016). What is machine learning? *Machine Learning with MATLAB*, *12*, 1–10. https://doi.org/10.1111/j.2041-210X.2010.00056.x

Met Office. (2021). MIDAS Open: UK Daily Rainfall Data, V202007. Centre for Environmental Data Analysis. 2021 https://doi.org/10.5285/ec9e894089434b03bd9532d7b343ec4b

Meyers, G., Kapelan, Z., & Keedwell, E. (2017). Short-term forecasting of turbidity in trunk main networks. *Water Research*, *124*, 67–76. https://doi.org/10.1016/j.watres.2017.07.035

Mian, H. R., Guangji, H., Hewage, K., Rodriguez, M. J., & Sadiq, R. (2018). Prioritization of unregulated disinfection by-products in drinking water distribution Systems for Human Health Risk Mitigation: A critical review. *Water Research*, *147*, 112–131. https://doi.org/10.1016/j.watres.2018.09.054

Mohammed, H., Hameed, I. A., & Seidu, R. (2017). Random Forest tree for predicting fecal indicator organisms in drinking water supply. In *International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, Krakow, Poland (pp. 1–6). IEEE. https://doi.org/10.1109/BESC.2017.8256398

Mounce, S. R., Ellis, K., Edwards, J. M., Speight, V. L., Jakomis, N., & Boxall, J. B. (2017). Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems. *Water Resources Management*, *31*(5), 1575–1589. https://doi.org/10.1007/s11269-017-1595-8

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Rossman, L. A., Clark, R. M., & Grayman, W. M. (1994). Modeling chlorine residuals in drinking-water distribution systems. *Journal of Environmental Engineering, ASCE*, *120*(4), 803–820. https://doi.org/10.1061/(ASCE)0733-9372(1994)120:4(803)

Seiffert, C., Khoshgoftaar, T. M., Van Hulse J., & Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. In *19th International Conference on Pattern Recognition*, Tampa, FL, USA (pp. 1–4). IEEE. https://doi.org/10.1109/ICPR.2008.4761297

Speight, V., & Boxall, J. (2015). Current perspectives on disinfectant modelling. *Procedia Engineering*, *119*, 434–441. https://doi.org/10.1016/j.proeng.2015.08.906

Speight, V., Mounce, S., & Boxall, J. B. (2019). Identification of the causes of drinking water Discolouration from machine learning analysis of historical datasets. *Environmental Science: Water Research & Technology*, *5*(4), 747–755. https://doi.org/10.1039/c8ew00733k

Vasconcelos, J. J., Rossman, L. A., Grayman, W. M., Boulos, P. F., & Clark, R. M. (1997). Kinetics of chlorine decay. *Journal - American Water Works Association*, *89*, 54–65. https://doi.org/10.1002/j.1551-8833.1997.tb08259.x

World Health Organization. Water, Sanitation and Health Team. (2000). *WHO guidelines for drinking water quality: Training pack*. World Health Organization.

Xu, X., Liu, Y., Liu, S., Li, J., Guo, G., & Smith, K. (2019). Real-time detection of potable-reclaimed water pipe cross-connection events by conventional water quality sensors using machine learning methods. *Journal of Environmental Management*, *238*(March), 201–209. https://doi.org/10.1016/j.jenvman.2019.02.110

How to cite this article: Kyritsakas, G., Boxall, J., & Speight, V. (2024). A data-driven predictive model for disinfectant residual in drinking water storage tanks. *AWWA Water Science*, *6*(3), e1376. https://doi.org/10.1002/aws2.1376