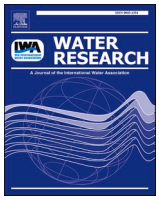
Water Research 240 (2023) 120075

Contents lists available at ScienceDirect

Water Research

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



Predicting microbial water quality in on-site water reuse systems with online sensors

Eva Reynaert [a,b,*], Philipp Steiner [a], Qixing Yu [a,c], Lukas D'Olif [a,b], Noah Joller [a,b], Mariane Y. Schneider [d,*], Eberhard Morgenroth [a,b]

[a] Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b] ETH Zürich, Institute of Environmental Engineering, 8093 Zürich, Switzerland
[c] Ecole Polytechnique Fédérale de Lausanne (EPFL), Section of Environmental Sciences and Engineering, 1015 Lausanne, Switzerland
[d] The University of Tokyo, Next Generation Artificial Intelligence Research Center & School of Information Science and Technology, 113-8656 Tokyo, Japan

ARTICLE INFO

Keywords:
Online monitoring
Machine learning
Membrane bioreactor
Chlorination
Virus removal
Bacterial regrowth

ABSTRACT

Widespread implementation of on-site water reuse is hindered by the limited availability of monitoring approaches that ensure microbial quality during operation. In this study, we developed a methodology for monitoring microbial water quality in on-site water reuse systems using inexpensive and commercially available online sensors. An extensive dataset containing sensor and microbial water quality data for six of the most critical types of disruptions in membrane bioreactors with chlorination was collected. We then tested the ability of three typological machine learning algorithms – logistic regression, support-vector machine, and random forest – to predict the microbial water quality as "safe" or "unsafe" for reuse. The main criteria for model optimization was to ensure a low false positive rate (FPR) – the percentage of safe predictions when the actual condition is unsafe – which is essential to protect users health. This resulted in enforcing a fixed FPR $\leq$ 2%. Maximizing the true positive rate (TPR) – the percentage of safe predictions when the actual condition is safe – was given second priority. Our results show that logistic-regression-based models using only two out of the six sensors (free chlorine and oxidation–reduction potential) achieved the highest TPR. Including sensor slopes as engineered features allowed to reach similar TPRs using only one sensor instead of two. Analysis of the occurrence of false predictions showed that these were mostly early alarms, a characteristic that could be regarded as an asset in alarm management. In conclusion, the simplest algorithm in combination with only one or two sensors performed best at predicting the microbial water quality. This result provides useful insights for water quality modeling or for applications where small datasets are a common challenge and a general advantage might be gained by using simpler models that reduce the risk of overfitting, allow better interpretability, and require less computational power.

Abbreviations

FC	free chlorine
FN	false negative
FP(R)	false positive (rate)
Logit	logistic regression
LRT	log-removal target
LRV	log-removal value
MBR	membrane bioreactor
ORP	oxidation–reduction potential
RF	random forest
SVM	support-vector machine
TN	true negative
TP(R)	true positive (rate)

1. Introduction

On-site water reuse can improve global access to clean water, sanitation, and hygiene (Rodriguez et al., 2020) and increase water use efficiency (Wilcox et al., 2016), but only if the reclaimed water is safe for the intended reuse application. Treatment technologies for on-site water reuse are increasingly becoming available. Membrane bioreactors

* Corresponding authors.
E-mail addresses: eva.reynaert@eawag.ch (E. Reynaert), mariane.schneider@alumni.ethz.ch (M.Y. Schneider).

https://doi.org/10.1016/j.watres.2023.120075
Received 14 November 2022; Received in revised form 24 March 2023; Accepted 11 May 2023
Available online 13 May 2023

(MBRs) produce higher effluent quality than conventional treatment processes and are considered a "best available" technology for the treatment of wastewater from as few as 20 people (Diaz-Elsayed et al., 2019; Lesjean et al., 2011). MBR-based treatment systems are often combined with chlorination to ensure high microbial water quality and to meet disinfection requirements set by regulatory agencies (Hirani et al., 2014).

MBR with chlorine disinfection (MBR+Cl_2) is a well-tried and robust combination of technologies for water reclamation with a low probability of treatment disruptions (Hirani et al., 2014). However, even a single day of disrupted operation can pose a risk to human health (Schoen et al., 2018). Therefore, monitoring systems must be established to ensure microbially safe water when hazardous events occur (Branch et al., 2016). Currently, no widely applicable online monitoring exists for MBR+Cl_2 systems, which constitutes a major bottleneck for their real-world application (Reynaert et al., 2021).

Microbial water quality targets depend on the specific reuse context, because MBR+Cl_2 systems can treat various types of water such as mixed wastewater or source-separated greywater for a range of reuse applications such as landscaping, toilet flushing, and showering at different scales, including household and building scale. Quantitative microbial risk assessment can be used to calculate treatment log-removal targets (LRTs) for pathogens for combinations of wastewater qualities, reuse applications, and reuse scales, which thus ensures that the risk to human health remains below a certain benchmark (WHO, 2016). Stochastic models have been applied to determine the monitoring frequencies required to prevent a significant increase in risk as a function of the LRT. These indicate that frequencies as low as 1 s are required to verify a LRT of 7 (Smeets, 2010). Manual sampling and laboratory-based analytical methods at such frequencies are impractical due to high costs. Consequently, online monitoring of the microbial water quality becomes indispensable for high-risk applications.

Soft sensors are software-based models, which are increasingly used to predict response variables that are difficult to measure with data that can be obtained by more easily applicable methods (Haimi et al., 2013). Soft sensors have been developed for monitoring the effluent quality of drinking water (Aliashrafi et al., 2021) and on-site wastewater treatment plants (Haimi et al., 2013; Schneider et al., 2019). So far, such soft-sensing approaches have primarily been used to predict physico-chemical water quality parameters or fecal indicator bacteria such as *E. coli*. For instance, Bedell et al. (2022) use fluorescence measurements with an ensemble learning method to detect and quantify *E. coli* in a drinking water supply. Similarly, Foschi et al. (2021) use a range of conventional measurements, including pH, conductivity, turbidity, and UV absorbance, and neural networks to predict *E. coli* in wastewater disinfection inflow.

In MBR+Cl_2 systems, fecal indicator bacteria are not the most critical group of enteric pathogens, because they are mostly retained by the membrane in contrast to enteric viruses, which are smaller and more resistant to many treatment processes (Zhu et al., 2020). Even if enteric pathogens are removed by the treatment process, bacteria, including opportunistic pathogens, can regrow in the treated water without a disinfection residual (Garner et al., 2019; Nocker et al., 2020).

The present study aims to develop a methodology that couples risk-based approaches with machine learning algorithms to issue an alarm if microbial water quality targets for minimum virus removal and maximum bacterial regrowth are not met. We compare the predictive power for virus removal and bacterial regrowth in MBR+Cl_2 systems of three typological machine learning algorithms: logistic regression, support vector machine, and random forest. We used data from inexpensive and commercially available online sensors: free chlorine (FC), oxidation–reduction potential (ORP), pH, turbidity, conductivity, and temperature. The models are trained to prioritize human health by minimizing predictions that water is safe for reuse when it is not, and we also make a detailed analysis of the consequences of this conservative approach for the frequency of false alarms.

In a related paper, Reynaert et al. (2023) provide sensor setpoints that allow tuning the chlorination to a certain water quality (e.g., a certain log-removal value for viruses). But these sensor setpoints are valid for standard operation, i.e., for situations during which the water reuse system are operating under non-disrupted conditions only. The setpoints can thus not be used to predict the water quality in real time. The present study aims to fill this gap and provide a real-time prediction of the microbial water quality under dynamic operation.

2. Materials and methods

2.1. MBR+Cl_2 system: Water Wall

The MBR+Cl_2 system used in this study is referred to as the Water Wall (Reynaert et al., 2020). The Water Wall (see Fig. 1) consists of two main components: the core treatment takes place in a biologically activated membrane bioreactor (BAMBi, Künzle et al., 2015), after which the water is polished and disinfected in the second component, a clean water tank.

The BAMBi reactor contains a standing sandwich membrane module (Microclear MCXL, Newterra, Langgöns, Germany) with a 150 kDa polyethersulfone ultrafiltration membrane (Microdyn-Nadir, Wiesbaden, Germany). Aeration is introduced directly below the membrane module. The reactor is operated in a gravity-driven membrane configuration in which the pressure on the membrane is supplied only by the water head (Peter-Varbanets et al., 2010). Water that passes through the membrane is collected in a permeate reservoir (10 cm polyvinyl chloride pipe connected to the membrane module permeate outlet, holding volume of 4 L), from where it is pumped through a granular activated carbon filter (Norit 830, ~1.5 mm grain diameter, Cabot, Boston, USA) to the clean water tank. In the clean water tank, a concentrated NaOCl solution (1750 mg Cl_2/L) is pumped into the tank at regular intervals (3 s on/250 s off) to reach a concentration of 1 mg/L of free chlorine. Mixing in the clean water tank is ensured through a submersed pump that operates for 30 s every 5 min. Water from the clean water tank is constantly recirculated through three sensor flow cells at a rate of 0.5 L/min. The tank volumes are 60 L water for the reactor and 25 L for the clean water tank, and average hydraulic residence times are 19 h in the

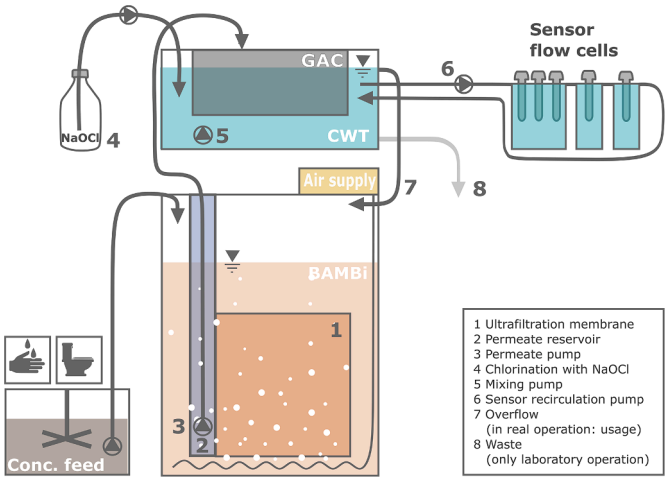


Fig. 1. Experimental setup for the biologically activated membrane bioreactor (BAMBi) configured with granular activated carbon (GAC) and chlorination post-treatment with a concentrated NaOCl solution. The clean water tank (CWT) is positioned above the BAMBi so that the overflow water from the CWT flows into the BAMBi. Water from the CWT is constantly pumped through flow cells containing five commercially available online sensors. In this laboratory setup, concentrated feed, representing handwashing or source-separated toilet flush water, was added to the BAMBi, with the same quantity of water being removed from the CWT as waste.

reactor and 5 h in the clean water tank.

A total of 3.75 L/day of concentrated feed is pumped into the reactor in 50 feedings evenly distributed throughout the day. This daily feed represents the loading that would be introduced by a total of 75 L of water of real hand washing or source-separated toilet flush water, equivalent to the usage of a 10-person household. The same amount of water is removed from the system to maintain a constant volume.

2.2. Standard operation of the Water Wall

Two full-scale Water Walls were operated in this study: one mimicked the recycling of source-separated toilet flush water separated from the majority of urine and feces (WW_{TF}) and the other handwashing greywater (WW_{HW}). The composition of the 20 × concentrated feed solutions is presented in the Supplementary Information (SI 1). The Water Walls were operated for several days under stable conditions as described above, with constant feed and 1 mg/L of chlorine in the treated water, before they were subject to the disruptions described in section 2.1.2.

2.3. Disruptions of operation

Failure mode and effects analysis is a systematic method for identifying the possible failures that pose the most significant overall risk to a process. We used this type of analysis to estimate four human-health-relevant failures per year and identified the most problematic failure modes (more information in SI 2). The following disruptions were experimentally simulated and are ordered according to their risk priority number (occurrence × severity × effort for remediation):

1. "Aeration off": Breakdown of the aeration for example due to pump breakdown or clogging of the aeration tubes, leads to ammonia in the permeate due to incomplete nitrification, which consumes the free chlorine in the clean water tank. The aeration pump was manually deactivated to simulate the breakdown of the aeration.
2. "Chlorine off": Breakdown of the chlorination, for example due to pump breakdown, tube clogging, or no refill of chlorine solution, leads to a decrease of chlorine concentrations in the treated water. The pump was manually deactivated to simulate a breakdown of the chlorine pump.
3. "Power off": A power outage can cause elevated concentrations of organics and ammonia in the permeate, which consumes the free chlorine once the power is back and permeate is pumped into the clean water tank. The aeration pump and the permeate pump were manually deactivated to simulate a power outage.
4. "High usage": High usage of the systems, for instance during a party, can cause elevated concentrations of organics and ammonia in the permeate. High usage of the system was simulated through a 15-fold increase of the concentrated feed.
5. "Membrane damage": Membrane damage can lead to direct passage of contaminants into the treated water. Membrane damage was simulated by pumping liquid from the reactor into the clean water tank at a flow rate of 1.6 L/h, simulating a 0.5 mm-diameter hole in the membrane.
6. "Toxic substance": A spill of a toxic substance into the biological treatment tank can be harmful to the biomass in the reactor and can result in increased concentrations of organics and ammonia in the permeate. This disruption was simulated by the instantaneous addition of 1 L of a cleaning substance (0.5% sodium hypochlorite) into the reactor.

During the disruptions, frequent samples were taken to assess the microbial water quality (see Section 2.2). Sampling started before the start of the disruption and continued until the microbial water quality was stable, i.e. there was no further deterioration of the water quality. For all disruptions, it took several hours after the start of the disruption for the water quality to deteriorate due to the robustness of the Water Wall technology. Disruptions were mimicked between December 2021 and March 2022. After each disruption, the Water Walls were given at least three days to recover before the next disruption was simulated (timeline of experiments presented in SI 3).

3. Data

3.1. Online sensors

We investigated established commercially available sensors with promising mechanistic relationships with microbial water quality and excluded sensors that we considered too costly for small-scale applications. Five sensors were installed in sensor flow cells to monitor the water quality: ORP, FC, pH and temperature, turbidity, and conductivity (Table 1). Promising new sensing approaches, such as online ATP measurements or online flow cytometry, were excluded in this study focusing on on-site reuse systems, due to high costs and requirement for qualified operating personnel and consumables. Reference measurements were taken with the recommended buffer solutions for ORP (220 mV) and pH (pH 4 and 7). The pH sensors were calibrated whenever the drift was larger than 0.2. The FC sensor was calibrated with reference-free chlorine measurements (Hach DPD test kits, 0–2 mg/L free chlorine, Hach, Loveland, USA) at the flow cell. Turbidity, conductivity, and temperature sensors were not calibrated during the experiments. Sensor measurements were automatically logged at 5-min intervals.

3.2. Microbial water quality

Microbial water quality was evaluated for the removal of enteric pathogens and regrowth of pathogens in the treated water.

We used the bacteriophage MS2 as an indicator of the removal of enteric viruses. A concentrated solution of MS2 was spiked into the feed. The log-removal value (LRV) was calculated as

Table 1
Specifications and expected links to microbial water quality of the sensors installed in the Water Wall. All sensors were purchased from Endress+Hauser, Reinach, Switzerland.

Measurement	Sensor specification	Measurement principle	Mechanistic relationship with the microbial water quality
Conductivity	Condumax CLS21D	Electric current carried by charged ions	Information on changes in the water composition
Free chlorine (FC)	Digital free chlorine sensor Memosens CCS51D	Closed, membrane-covered measuring cell; reduction of free chlorine at the cathode	Direct measurement of free chlorine concentration
Oxidation-reduction potential (ORP)	Ceragel CPS72D	Ceramic diaphragm double chamber and double gel reference-platinum ring	Measurement of the oxidative capacity of all chlorine species
pH	Orbisint CPS11D	Gel compact electrode with PTFE ring diaphragm	Information on speciation and thus disinfection potential of free chlorine
Temperature	Orbisint CPS11D	Change in electrical resistance	Information on speciation and efficacy of chlorine; influence on regrowth of bacteria
Turbidity	Turbimax CUS52D	Nephelometric turbidity sensor (90° scattering) according to ISO7027	Turbidity can be linked to bacteria concentrations

$$LRV_{MS2} = -\log_{10}\frac{C_{CWT}}{C_{feed}/f_{conc}} \quad (1)$$

where C_{CWT} is the MS2 concentration in the clean water tank, C_{feed} the MS2 concentration in the concentrated feed, and f_{conc} is the concentration factor of the feed. The double agar layer method was used to enumerate MS2 as described in detail in Reynaert et al. (2023).

Regrowth was quantified as the total concentration of bacterial cells with an intact membrane and measured using flow cytometry as described in Reynaert et al. (2023). The regrowth of human enteric pathogenic viruses in the treated water is not anticipated due to the absence of host cells (Zhu et al., 2020).

3.3. Development of prediction models

The overall goal of this study was to predict whether microbial water quality targets are met using machine learning algorithms. In machine learning, simplifying assumptions made by a model to make the target function easier to learn result in a model bias. The amount that the estimate of the target function will change if different training data is used is termed the variance. The model bias can only be decreased at the cost of a higher variance, or vice versa, resulting in a bias–variance tradeoff (Briscoe and Feldman, 2011). In this study, we selected three commonly used algorithms with differing bias–variance tradeoffs:

- Logistic regression (logit) has high bias and low variance. It models the probability of the water quality meeting the microbial target as a logistic (sigmoid) function of a linear combination of one or more features. Logit represents the most basic regression method for binary classification and was used as a benchmark method against which the other algorithms were compared. For models using only one feature as input, the logicistic regression comes down to fixing a threshold for this feature.
- Support-vector machine (SVM) has intermediate bias and intermediate variance. It separates the water quality into two classes. Unlike logit, the separation does not need to be linear. SVM creates a hyperplane (decision boundary) by applying transformations to the input data if the data is not linearly separable (kernel trick).
- Random forest (RF) has low bias and high variance. It is an ensemble learning method that randomly sets up a large number of decision trees made of sequential binary decisions. The final prediction is then calculated as the average from the final predictions of all trees.

The development of prediction models using these three algorithms is illustrated in Fig. 2 and described in the following subsections. All models were implemented using the scikit-learn 1.1.2 package (Pedregosa et al., 2011) for Python 3 (Van Rossum and Drake, 2009). The code can be downloaded from https://doi.org/10.25678/000885.

3.4. Microbial water quality targets

To assess the microbial quality of reclaimed water, water reuse frameworks specify various LRTs for the removal of enteric pathogens depending on the contamination of the wastewater and the reuse application. Monitoring needs to be optimized for specific water quality targets (Reynaert et al., 2021). To demonstrate the applicability of the approach to various water quality requirements, we present results for two virus LRTs (LRT = 5 and 6) and two maximum concentrations of intact cells (ICC, $\log_{10}$ICC = 4 and 5), from which the microbial water quality was classified into two categories: "safe," meeting the target, and "unsafe" not meeting the target. For LRT, the lower value represents a less stringent target. LRT = 5 can represent the virus removal required when reusing greywater from a 1000-people collection system for toilet flushing, and LRT = 6 can represent the removal required when reusing the same water for indoor reuse (Schoen et al., 2017). For ICC, the lower value represents a stricter quality target. We cannot define risk-based targets here, as the ICC reflects the entire bacterial community and is not representative of pathogens. However, it is possible to define typical ICC values for a specific system (Van Nevel et al., 2017).

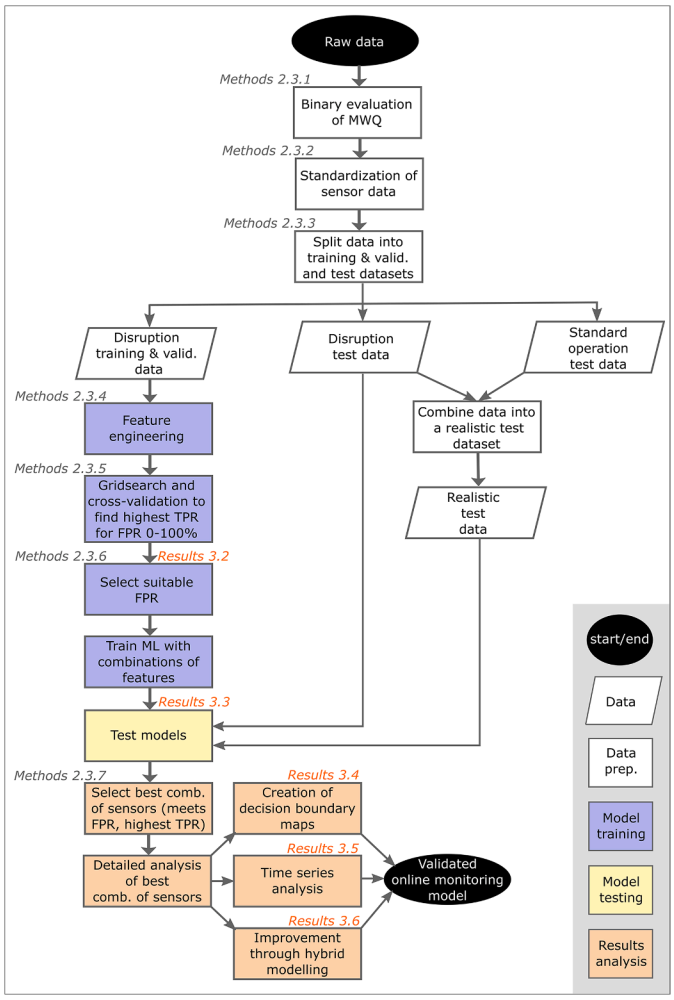


Fig. 2. Flowchart for developing machine learning (ML) models that predict the microbial water quality (MWQ) for one ML algorithm, one microbial indicator, and one MWQ target. FPR: false positive rate. TPR: true positive rate.

3.5. Standardization of sensor data

Before being used as input to the machine learning algorithms, the sensor measurements were standardized to zero mean and one standard deviation.

3.6. Training, validation, and testing

The disruption dataset was balanced, meaning that the numbers of safe and unsafe data points were similar. The percentages of safe data points were 58% (LRT MS2 $\geq$ 6), 60% (LRT MS2 $\geq$ 5), 74% ($\log_{10}$ICC $\leq$ 5), and 51% ($\log_{10}$ICC $\leq$ 4).

The data was divided into a training & validation dataset and a test dataset, as presented in Table 2.

For testing, we used either the disruption data alone or an assembled test dataset representative of realistic operation:

- Disruption test data: This dataset was balanced with respect to the numbers of safe or unsafe data points. Testing on the disruption test data provided information on the model performance during failures of the Water Wall and served to verify if all occurrences when the

Table 2
Overview of disruptions used for the training and validation of the machine learning models and for the testing. Water Walls (WW) treating source-separate toilet flush (TF) water or handwashing water (HW). t: duration of the disruption, n: number of data points for removal of MS2 and intact cell concentration (ICC), op: operation.

Training & validation dataset					Test dataset				
Disruption	**WW**	**t [h]**	**n (MS2)**	**n (ICC)**	**Disruption**	**WW**	**t [h]**	**n (MS2)**	**n (ICC)**
Aeration off 1	TF	24	15	14	Chlorine off 5	TF	28	12	12
Aeration off 2	TF	24	9	9	Chlorine off 6	TF	27	23	21
Aeration off 3	HW	24	15	14	Chlorine off 7	HW	27	19	21
Aeration off 4	HW	24	14	14	Chlorine off 8	HW	26	16	18
Chlorine off 1	TF	26	16	18	Chlorine off 9	HW	7	10	15
Chlorine off 2	TF	25	15	19	Chlorine off 10	HW	12	12	25
Chlorine off 3	HW	30	22	22	Chlorine off & restart 1	TF	16	12	12
Chlorine off 4	HW	30	29	27	Chlorine off & restart 2	HW	16	12	12
Chlorine restart 1	TF	n/a	11	15	High usage 3	HW	6	9	10
Chlorine restart 2	TF	n/a	7	8	High usage 4	HW	6	13	14
Chlorine restart 3	HW	n/a	10	11	Power outage 5	HW	12	12	13
Chlorine restart 4	HW	n/a	8	10	Power outage 6	HW	12	17	16
Memb. damage 1	HW	36	15	15	Toxic subst. 1	HW	n/a	15	11
Memb. damage 2	HW	48	24	24	**Total: 13 disruptions**			**182**	**200**
High usage 1	TF	6	9	10					
High usage 2	TF	6	12	14					
Power outage 1	TF	12	11	11					
Power outage 2	TF	12	13	13	Standard op. data 1	TF	12	12	12
Power outage 3	HW	36	28	28	Standard op. data 2	HW	12	12	12
Total: 19 disruptions			**283**	**296**	**Total: standard op. data**			**24**	**24**

water was not safe for reuse were correctly detected (user perspective).

- Realistic test data: The disruption dataset is not representative of realistic operation, where the water is safe over 99% of the time. Therefore, we assembled a dataset from sensor data from standard undisrupted operation combined with four disruptions per year (see Section 2.1.2). From baseline measurements of the microbial water quality, the water was assumed to always be safe in standard operation. The 13 disruptions included in the test dataset allowed standard operation to be simulated for 3.25 years. Testing with the realistic dataset served to verify that the percentage of false alarms remained low in long-term operation (operator perspective).

3.7. Feature engineering and hybrid modeling

Principal component analysis (PCA) was used to reduce the dimensionality of the sensor measurements (Khalid et al., 2014). The optimal number of principal components was chosen using cross-validation during the parameter optimization of the machine learning models (see Section 2.3.5).

Additionally to sensor measurements, the rate of change of sensor measurements over time, termed sensor slope, was used to include mechanistic knowledge about the system dynamics. The temporal dynamics of LRV MS2 and $\log_{10}$ICC differ depending on the system state. Deterioration of the water quality is a relatively slow process, as it takes time for MS2 to enter the clean water tank, and thus for LRV MS2 to decrease, and for bacterial cells to grow in the clean water tank, and thus for $\log_{10}$ICC to increase. For instance, it will take at least one HRT for the water in the clean water tank to be replaced, and thus for the LRV MS2 to decrease to 0. In contrast, the water quality can be improved almost instantly if sufficient chlorine is added to disinfect it. Combining this process knowledge with the machine learning approach was a first test whether hybrid modeling is promising for MBR+Cl_2 systems. Hybrid modeling here means the combination of data-driven and mechanistic models (Schneider et al., 2022).

3.8. Model optimization

Model performance was evaluated using the confusion matrix presented in Table 3.

The true positive rate (TPR) and false positive rate (FPR) were calculated using the following equations:

Table 3
Confusion matrix.

		Actual condition Positive (safe)	Negative (unsafe)
Predicted condition	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Note that the use of "positive" to describe safe water and "negative" to describe unsafe water differs from the terminology used in medical testing, where a positive outcome typically refers to a detection.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

Optimal model hyperparameters, confidence thresholds, and number of components from PCA were identified using a grid search with reasonable parameter values. The assumption underlying this parameter optimization was that a low FPR is more important, because these are false predictions that will put users' health at risk. The TPR was given second priority, thus reducing the number of FNs and associated false alarms.

The models were trained and validated with leave-one-out cross-validation, in which replicates of one disruption type for each reactor, together termed a disruption set, were removed from the training dataset at each iteration and used to validate the model. Leave-one-out cross-validation is a resampling technique appropriate for relatively small datasets when a highly accurate estimate of model performance is required (Wong, 2015).

3.9. Selection of FPR (operating point)

We computed optimization curves representing the highest TPR achieved for a certain FPR, and evaluated the gain in TPR per increase in FPR. Based on the optimization curves, we selected an FPR that was enforced during model training and validation.

3.10. Model evaluation

The FPR and TPR were reported on both the training and validation dataset and the test datasets comprising disruption test data and realistic

test data. For the best-performing models, the detailed analysis included (1) decision boundary maps to visualize the test datasets and for easier interpretability of model predictions, and (2) time series of predictions for a closer evaluation of the occurrence of FPs and FNs. Note that the FPR on the test datasets is not suitable to compare the actual performance of the different models, but rather to verify that the FPR selected as an operating point (see Section 2.3.6) is achievable.

4. Results

4.1. Experimental data: effect of disruptions on the microbial water quality

Understanding the relations of sensor measurements to water quality is critical to the success of predictive models. One major outcome of this study is a large dataset consisting of 32 disruptions (> 500 data points) with detailed information on changes in microbial water quality (LRV of MS2 and $\log_{10}$ICC) and sensor measurements. Fig. 3 shows the standard operation and typical effects of the six types of disruption simulated and the restart of chlorination on microbial water quality and on the sensor measurements. The complete dataset for all 32 simulated disruptions to two reactors is available at https://doi.org/10.25678/000885.

Fig. 3 shows that ORP values and FC concentrations both consistently decreased with deteriorations in the microbial water quality. Turbidity measurements also varied during some of the disruptions (e.g., membrane damage) but not for all, and not in a way that was consistent with the changes in microbial water quality. Conductivity only changed during the toxic substance scenario. Finally, neither the pH nor the temperature varied more in disruptive operation than during the standard operation.

Without disruptions, the microbial water quality was stable throughout the testing, with LRV MS2 > 6.4 and $\log_{10}$ICC < 2.7 (Fig. 3, standard operation).

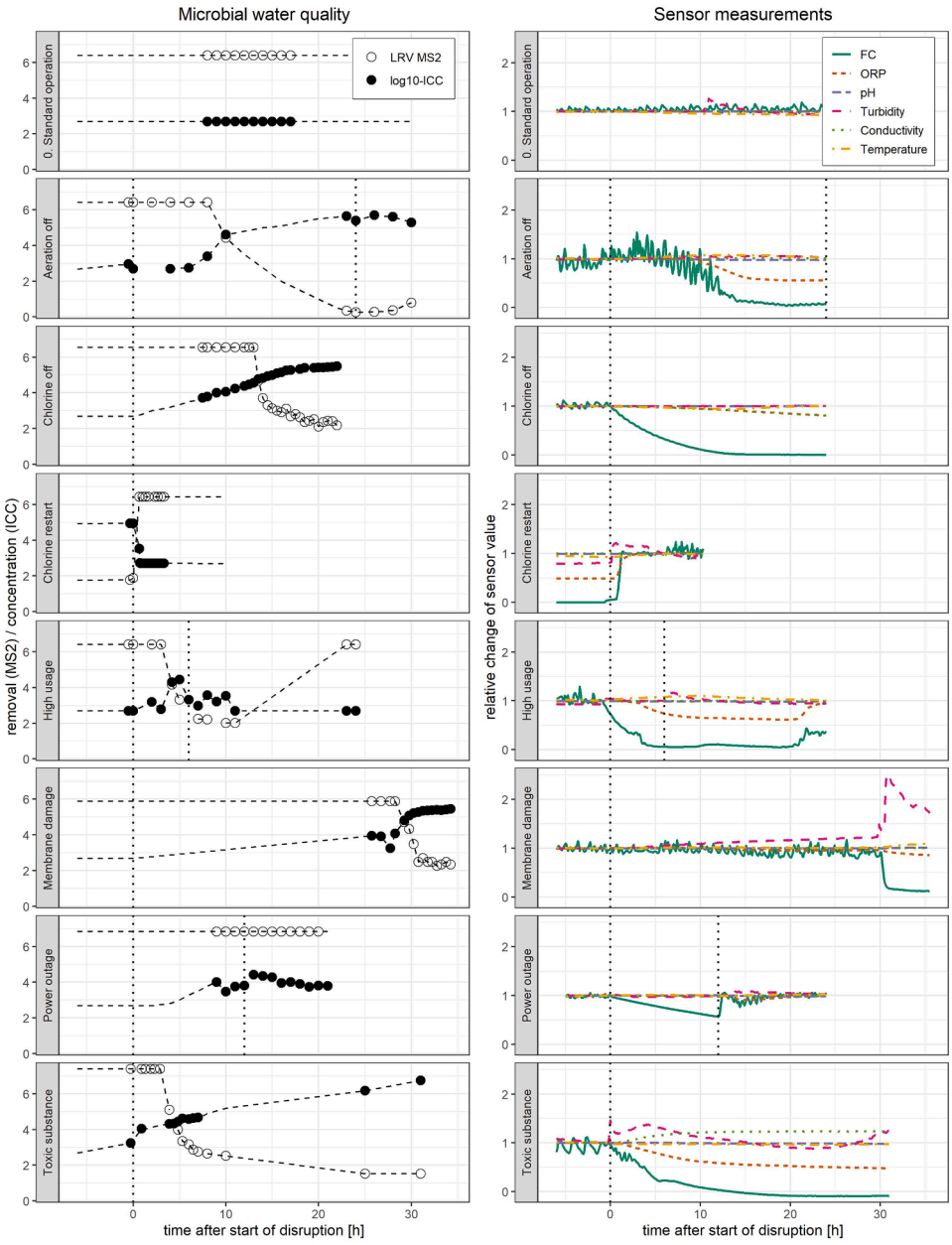


Fig. 3. Effect of treatment disruptions on the microbial water quality (log-removal value LRV of MS2, log10-value of ICC) and sensor measurements (normalized to 1 compared to $t = 0$, except chlorine restart: normalized to 1 for $t =$ 6 h). Dashed lines on microbial data represent expected interpolation. Vertical dotted lines represent the start of the disruption and, when applicable, the end of the disruption. For plots with only one vertical line, Water Walls were exposed to continued disruption.

4.2. Selection of suitable FPR (operating point)

Fig. 4 shows the optimization curves that were used to select a suitable FPR. The optimization curve represents the optimized TPR for a certain FPR, determined from a grid search of possible model parameters. Each point on an optimization curve thus represents different hyperparameters and confidence thresholds from the other points. These optimization curves differ from commonly-used receiver operating characteristic (ROC) curves that are computed for one model, i.e. one set of hyperparameters. FPR and TPR results are reported on the training and validation dataset using leave-one-out validation. Models can use all possible combinations of sensor measurements as input. For instance, for an SVM algorithm trained to predict meeting a LRT MS2 of 6 using measurements from six sensors, the optimized TPR is 46% if no FP are allowed, 58% if an FPR of 1% is allowed, and 68% if an FPR of 2% is allowed (Fig. 4.A). For a FPR above 2%, the TPR only increases slowly, meaning that we have to allow a significantly higher FPR to increase the TPR.

The optimization curves serve two purposes:

1. Selection of operating points: When selecting operating points, the goal was to keep the FPR as low as possible. Fig. 4 indicates that selecting a FPR of 0% is not an appropriate choice, as this is associated with very low TPRs. Allowing the FPR to increase to 1% or 2% is associated with significantly higher TPRs. For a FPR above 2%, the gain in TPR per increase in FPR is much lower. A FPR of 2% was thus selected as a sensible choice for training the monitoring algorithms.
2. Performance comparison of logit, SVM, and RF: Fig. 4 shows that logit and SVM often overlap. Therefore, they perform similarly, whereas RF is associated with lower TPRs.

Consequently, further model analysis was conducted for operating points at FPR = 2%.

4.3. Comparison of TPRs from different models

Table 4 shows the FPRs and TPRs of models using single sensors, two sensors, and all sensors as features. All models were trained to meet an FPR $\leq$ 2%. The single-sensor models with the highest TPR were ORP and FC, and two-sensor combinations with the highest TPR were ORP+FC. Results are reported for the training and validation data, the disruption test data, and the realistic test data. To be classified as performing well (highlighted in green), models needed to satisfy two conditions: (i) FPR $\leq$ 2% in the disruption test datasets and (ii) classification of the standard operation data as safe in the realistic test dataset. The FPR on the test datasets cannot be used to compare the performance of different models, as long as the condition from (i) is met (see Section 2.3.7). Four key results emerge from Table 4:

1. Logit and SVM were associated with higher TPRs than RF. Although RF performed relatively well on the training and validation data, the RF models seem to overfit: the performance significantly decreased

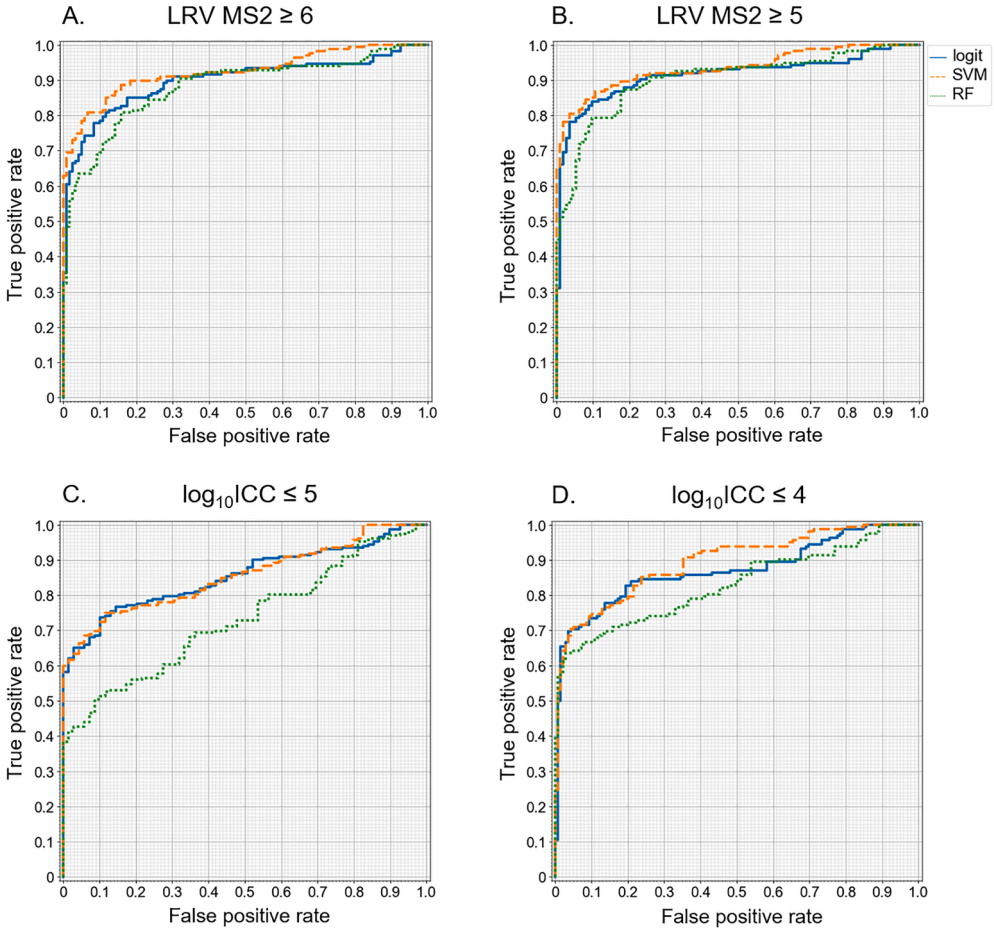


Fig. 4. Optimization curves for best-performing models (highest true positive rate for a certain false positive rate) using input from all six sensors. Each point on an optimization curve represents different hyperparameters and confidence thresholds from the other points. False positive: prediction is safe when the actual condition is unsafe. Logit: logistic regression. SVM: support vector machine. RF: random forest.

Table 4
True positive rate (TPR) and false positive rate (FPR) for models trained with one sensor (oxidation–reduction potential ORP or free chlorine concentration FC), two sensors, and all sensors. TPR and FPR are reported for the training and validation dataset, the disruption test dataset, and the realistic dataset (disruption test data + standard operation data). Selected operating point for training and validation: FPR = 0.02. Green and bold: well-performing models with FPR ≤ 0.02 in the realistic test dataset and correctly classification of all standard operation data as safe. Yellow: FPR ≤ 0.02, but false classifications of the standard operation data. Orange and italics: FPR exceeds 0.02. LRV MS2: log-removal value of MS2. log10ICC: log10-value of the intact cell concentration (ICC).

Microbial water quality target	Algorithm		# Sensors = 1 ORP			# Sensors = 1 FC			# Sensors = 2 ORP+FC (combination with highest TPR in training&valid.)			# Sensors = all ORP+ FC+Turbidity+ Conductivity+pH+ Temperature		
			Training & valid.	Disruption test data	Realistic test data	Training & valid.	Disruption test data	Realistic test data	Training & valid.	Disruption test data	Realistic test data	Training & valid.	Disruption test data	Realistic test data
LRV MS2 ≥ 6	Logit	FPR	0.017	0	**0**	0.017	0	0	0.017	0	**0**	0.017	0.013	**0.013**
		TPR	0.617	0.590	**0.998**	0.269	0.152	0.552	0.737	0.819	**0.999**	0.641	0.695	**0.998**
	SVM	FPR	0.017	0	**0**	0.017	0	0	0.017	0	**0**	0.008	0.178	*0.178*
		TPR	0.611	0.533	**0.998**	0.275	0.609	0.717	0.731	0.819	**0.999**	0.695	0.838	*0.998*
	RF	FPR	0.01	0	**0**	0.008	0	0	0.017	0	0	0.017	0.063	*0.063*
		TPR	0.563	0.520	**0.998**	0.102	0	0	0.671	0.743	0.949	0.557	0.438	*0.870*
LRV MS2 ≥ 5	Logit	FPR	0.018	0	**0**	0.018	0	0	0.018	0	**0**	0.018	0	**0**
		TPR	0.586	0.569	**0.998**	0.621	0.606	0.998	0.741	0.817	**0.999**	0.695	0.825	**0.999**
	SVM	FPR	0.018	0	**0**	0.018	0	**0**	0.018	0	**0**	0.018	0.2	*0.2*
		TPR	0.563	0.569	**0.998**	0.632	0.679	**0.998**	0.741	0.807	**0.999**	0.782	0.997	*0.999*
	RF	FPR	0.009	0	**0**	0.018	0	0	0.018	0	**0**	0.018	0	0.0
		TPR	0.54	0.504	**0.997**	0.626	0.651	0.994	0.713	0.807	**0.999**	0.534	0.580	0.919
$\log_{10}$ICC ≤ 5	Logit	FPR	0.014	0.017	**0.017**	0.014	0	**0**	0.014	0	**0**	0.014	0	**0**
		TPR	0.608	0.644	**0.999**	0.496	0.519	**0.997**	0.690	0.644	**0.998**	0.621	0.630	**0.998**
	SVM	FPR	0.014	0	**0**	0.014	0	**0**	0.014	0	**0**	0.014	0.051	*0.051*
		TPR	0.586	0.596	**0.997**	0.478	0.489	**0.997**	0.591	0.615	**0.998**	0.616	0.993	*0.997*
	RF	FPR	0.014	0.0	0	0.014	0	0	0.014	0	0	0.014	0	0
		TPR	0.569	0.0	0	0.517	0	0	0.569	0.600	0.957	0.414	0.319	0.839
$\log_{10}$ICC ≤ 4	Logit	FPR	0.014	0.049	*0.049*	0.014	0	0	0.014	0.029	*0.029*	0.014	0.088	*0.088*
		TPR	0.654	0.604	*0.998*	0.265	0.154	0.521	0.463	0.407	*0.991*	0.654	0.571	*0.918*
	SVM	FPR	0.014	0.049	*0.049*	0.007	0	0	0.014	0.078	*0.078*	0.014	0.088	*0.088*
		TPR	0.642	0.593	*0.998*	0.259	0.121	0.403	0.630	0.604	*0.998*	0.605	0.560	*0.930*
	RF	FPR	0.014	0.039	0.039	0.014	0	0	0.014	0.068	0.068	0.014	0.127	0.127
		TPR	0.605	0.593	0.998	0.025	0	0	0.580	0.549	0.865	0.574	0.253	0.898

on the test datasets, with the FPR frequently exceeding 2%, and the standard operation data being misclassified. Comparing the less complex logit with the more complex SVM, we can observe that in this study, based on an inherently small dataset, the model complexity has neither a positive nor a negative effect on the TPR.
2. For most models considered, the TPR was > 99% when calculated for the realistic test dataset. This result shows that most models correctly classified the standard operation data as safe (exception: see point 3 below).
3. The monitoring algorithm development requires correctly classifying the standard operation data as safe. None of the models tested performed well for a $\log_{10}$-ICC of 4, where the models incorrectly classified the standard operation data as unsafe, resulting in a low TPR on the realistic test dataset.
4. Compared to single-sensor models, the combination of information from ORP and FC improved the models' predictive capacity for both aspects: keeping the FPR below 2% in the test dataset and correctly classifying the standard operation data as safe. However, including more sensors did not generally improve the TPR: we even observed a tipping point beyond which including all sensors reduced performance compared to the ORP+FC-based models.

Logit-based models using ORP+FC as input perform as well as SVM but are less computationally intensive. Therefore, a detailed evaluation of logit-based models using ORP+FC is presented in the following sections.[1]

4.4. Decision boundary maps

High performance is required not only on the training and validation dataset but also on the test dataset. Table 4 has already indicated that the models can achieve a low FPR and high TPR in the test datasets as well. Fig. 5 enables more detailed evaluation from decision boundary maps. The decision boundary maps show the probability of water being safe (gray shaded area) and the decision boundaries selected for the logit-based models using ORP+FC as input. The maps visualize how the models classify the water quality as safe (area right of the decision boundary) or unsafe (area left of the decision boundary), and allow evaluating how well the models are able to separate the two water quality classes. In Fig. 5A1-C1 (LRT MS2 6 and 5, $\log_{10}$-target of 5 for ICC), all cases where the water was unsafe in the disruption test dataset are classified correctly (no FPs, represented as red triangles), but some occurrences of water being safe are classified incorrectly. These incorrect classifications illustrate the consequence of selecting conservative decision boundaries. In contrast, there were three FPs for a $\log_{10}$-ICC of 4 (Fig. 5D1). The decision boundary would have had to be considerably more conservative to avoid these FPs at the cost of a low TPR. Interestingly, the best-performing SVM models also had linear decision

[1] Note: Figure 4 in Section 3.2 presents optimization curves for models using six sensors. The optimization curve presented in SI 5 confirms that a FPR of 2% for the training and validation stage is also a sensible choice for models based only on ORP+FC.

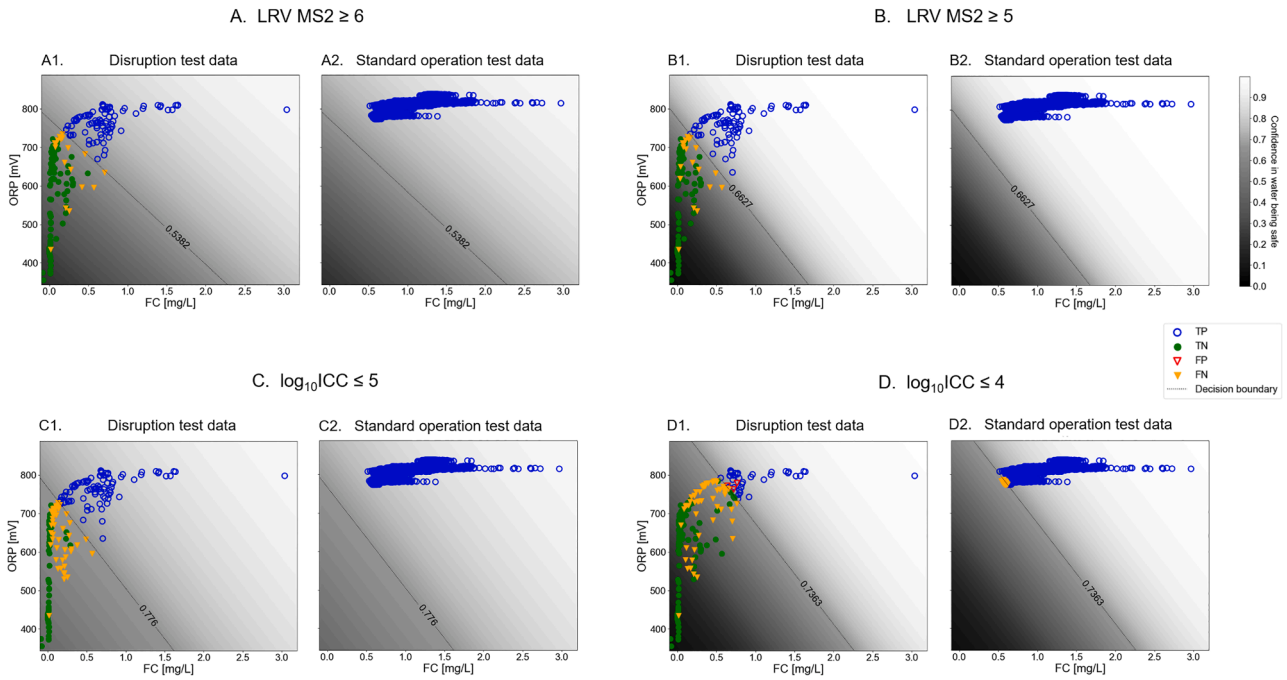


Fig. 5. Decision boundary maps for the disruption test dataset (A1-D1) and the standard operation test dataset (A2-D2). Logistic-regression-based models using two sensors (ORP+FC). The decision boundaries were set to keep the false positive rate below 2%. TP: true positive. TN: true negative. FP: false positive. FN: false negative.

boundaries, which explains why both SVM and logit have similar performance (results shown in SI 6).

Minimizing FPR at the cost of the TPR of the disruption dataset, relies on the algorithm always classifying the water correctly as safe in standard, undisrupted operation. From the high TPR on the realistic dataset, we can see that this is the case for all scenarios except for a $\log_{10}$-ICC of 4 (see Table 4). Fig. 5A2-D2 provide more detail on the classification for the standard operation. All points in standard operation are located far from the decision boundary in plots A2-C2, implying a low probability of misclassification. In contrast, plot D2 ($\log_{10}$-ICC of 4) shows that the data from the standard operation is close to the decision boundary, with a few occurrences where the model misclassifies the standard operation data (FNs represented as orange triangles).

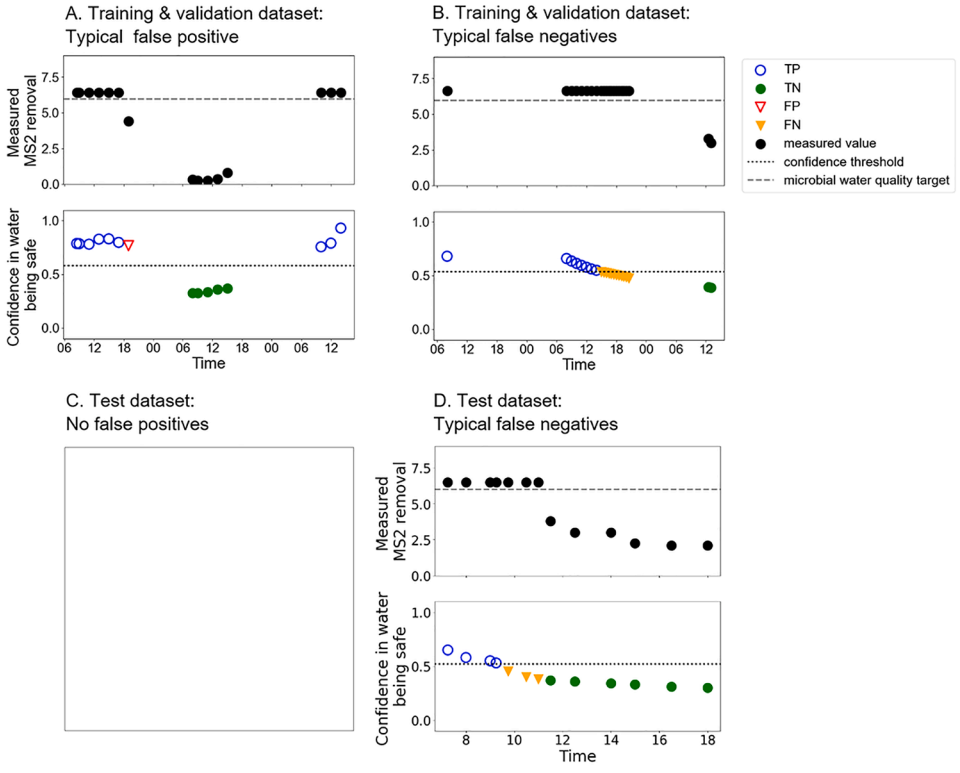


Fig. 6. Typical examples of prediction errors for logit-based models developed for microbial water quality for log-removal target of MS2 = 6, based on two sensors (ORP+FC). A: aeration off scenario for a Water Wall treating toilet flush water (WW_{TF}) (replicate 1). B: chlorine off scenario for a Water Wall treating handwashing water (WW_{HW}) (replicate 4). C: no false positives in the test dataset. D: chlorine off scenario for a Water Wall treating toilet flush water (WW_{TF}) (replicate 4). TP: true positive. TN: true negative. FP: false positive. FN: false negative.

4.5. Time series of predictions

Table 4 indicates that we can achieve TPRs between 64% and 82% in the disruption test datasets (logit using ORP+FC as inputs) while meeting the selected FPR of 2%. However, even low rates of FNs can lead to high operation costs caused by false alarms and a loss of trust in the system by the user. It is thus important to understand more precisely when FPs and FNs occur to determine a management strategy.

Fig. 6A and Fig. 6B show examples of prediction errors for an FP and FNs during training and validation of a logit-based model using ORP+FC as input. Fig. 6A indicates that the confidence in the water being safe is around the same for the TPs as for the FP (~0.75). Optimizing the model (meeting a MS2 LRT of 6) to have zero FPs instead of one through an increase of the decision boundary (dashed line) would massively decrease the TPR due to an increase of FNs. Results are reported on the training & validation dataset because there are no FPs in the test dataset for the two-sensor models predicting MS2 LRT. Conversely, Fig. 6B. (training & validation dataset) and Fig. 6D. (test dataset) show that FNs occur before the water quality falls below the treatment target, and thus represent early alarms. Unlike false alarms that occur randomly, an appropriate alarm management can mitigate these early alarms, for instance by using them for early interventions. This implies that the early alarms can become an asset when handled properly. The time series of predictions for both MBR+Cl_2 systems and all simulated disruptions can be found in SI 4.

4.6. Hybrid models

The data-driven approach presented in the previous sections was combined with mechanistic knowledge in a hybrid approach by including the sensor slopes. Sensor slopes allow the models to distinguish between deterioration and improvement of water quality. Table 5 presents the TPRs and changes in TPR compared to the results from the data-driven model in Table 4 for logit-based hybrid models, including ORP slopes (slope over 90 min) and/or FC slopes (slope over 210 min). Results for all models, including SVM and RF, are presented in SI 7. The time intervals used for the slope calculation were chosen to maximize the TPR.

Table 5 shows that the TPR can be increased by up to 25 percentage points on the training and validation dataset, often even with a lower FPR (see SI 7). For hybrid models using ORP+ORP slope, the performance on the test dataset can almost match the performance of the ORP+FC-based data-driven models.

However, Table 5 also shows that this increase in TPR cannot always be transferred to the disruption test dataset. Although there is an increase in TPR for models using ORP+ORP slope and FC+FC slope, the TPR of the models using both ORP+FC and their respective slopes has a substantially lower TPR on the test dataset than the purely data-driven model. One reason for this decrease in TPR may be an overfit of the model during the training and validation stage.

5. Discussion

5.1. Selection of machine learning algorithms: no tradeoff between bias and variance

Although the dataset generated for this study (~500 data points from 32 disruption events) can be considered large for microbial water quality measurements, it is not comparable to typical volumes and velocities of "big data" (Zhou et al., 2017). In their review on data-driven modeling of drinking water treatment, Aliashrafi et al. (2021) show that neural networks are most commonly used to model water quality, followed by SVM, decision trees, and RF. Logistic regression has not been widely used, because most studies focus on regression rather than classification. In this study, we did not include neural networks because their training data requirement is substantially larger than those of SVM and RF (Osisanwo et al., 2017).

Our results show that more easily interpretable logit-based models performed on par with the more complex black box model (SVM), and RF-based models had lower TPRs than both. Consequently, this dataset produced no tradeoff between bias and variance: logit-based models with low bias also had the lowest variance. The similar performances of logit and SVM might be due to the limited size of the dataset, because overfitting issues can arise when applying more complex machine learning algorithms to small datasets. Small datasets are a common challenge in water quality modeling (Aliashrafi et al., 2021). An additional challenge is the variability of water quality measurements inherent in microbial indicator measurements such as MS2 (Levy et al.,

Table 5
True positive rate (TPR) for models trained with one sensor, two sensors, or all sensors, including ORP and FC slopes (hybrid models). The TPR is reported for the training and validation dataset, the disruption dataset, and the realistic dataset. TPR diff refers to the difference in TPR compared to the purely data-driven models from Table 4. Selected operating point for training and validation: false positive rate (FPR) = 2%. Color code: green = increase of TPR; yellow = no change; red = decrease of TPR.

Microbial water quality target		# Sensors = 1						# Sensors = 2		
		ORP+ORP slope			FC+FC slope			ORP+ORP slope+FC+FC slope		
		Training & valid.	Disrupt. test data	Realistic test data	Training & valid.	Disrupt. test data	Realistic test data	Training & valid.	Disrupt. test data	Realistic test data
LRT MS2 ≥ 6	TPR	0.874	0.724	0.999	0.353	0.286	0.506	0.796	0.705	0.999
	TPR diff	+0.257	+0.134	+0.001	+0.084	+0.134	-0.046	+0.059	-0.114	0.000
LRT MS2 ≥ 5	TPR	0.839	0.697	0.998	0.621	0.651	0.998	0.833	0.706	0.998
	TPR diff	+0.253	+0.128	0.000	0.000	+0.045	0.000	+0.092	-0.111	-0.001
log_{10}ICC ≤ 5	TPR	0.698	0.637	0.998	0.487	0.519	0.997	0.724	0.681	0.998
	TPR diff	+0.090	-0.007	-0.001	-0.009	0.000	0.000	+0.034	+0.037	0.000
log_{10}ICC ≤ 4	TPR	0.735	0.626	0.998	0.327	0.176	0.581	0.698	0.626	0.998
	TPR diff	+0.081	+0.022	0.000	+0.062	+0.022	+0.060	+0.235	+0.219	+0.007

2012). Therefore, a general advantage might be gained by using simpler models that reduce the risk of overfitting, allow better interpretability, and require less computational power. This is especially true for controlled environments such as the MBR+Cl_2 systems investigated in this study, where we do not expect complex underlying patterns that can be encountered in natural systems.

In MBR+Cl_2 systems, the inactivation of viruses and prevention of bacterial regrowth is mostly achieved with chlorination. Thus, we expected that ORP and FC sensors, which measure the disinfection capacity and the concentration of chlorine, respectively, would be promising predictors of the microbial water quality (Reynaert et al., 2023). In this study, we observed that models using two sensors (ORP+FC) or even only one sensor (ORP+ORP slope) performed better on the test dataset than models including all sensors. Possible reasons for the lower performance of models using all sensors may include (1) the low enforced FPR, as the first classification errors occurred very soon when using all sensors, whereas higher FPRs models including all sensors had TPRs similar to the two-sensor models, and (2) the small size of the dataset, which increases the risk of overfitting when using input from multiple sensors.

The outputs from machine learning models need to be interpretable. Humans need to understand the reasoning behind the model outcome if they are to trust it, and the use of blackbox models in health-relevant applications is linked with concerns around accountability, safety, and liability (Aliashrafi et al., 2021). However, increased interpretability can come at the cost of lower accuracy, a tradeoff that has been observed, for instance, in a study aiming to predict *E. coli* concentrations in agricultural water (Weller et al., 2021). The present study's identification of an easily interpretable algorithm in combination with only one or two sensors in a known mechanistic relationship yielding the highest TPRs is a very positive outcome: it shows that we can have both high accuracy and good interpretability with the same models. As an additional advantage, logit is less computationally intensive and can easily be implemented in practical applications.

5.2. Conservative training: no tradeoff between detecting all failures and false alarms

Detecting all failures is crucial to limiting microbial risk, but avoiding false alarms is also important for operation costs (Storey et al., 2011) and, presumably, user acceptance of water reuse systems. While there are no studies on the effect of false alarms on user acceptance, studies have reported technology failure as a cause for a detrimental effect on the public acceptance of small-scale water reuse systems (Domènech and Saurí, 2010). Our results show that we can reach TPRs of between 70% and 80% when enforcing a conservative FPR of 2%. An important question is whether this TPR is sufficiently high to keep operational costs and the deterioration of user acceptance caused by false alarms reasonably low. Our results indicate that this is the case for two reasons:

1. FNs did not occur randomly but were usually early alarms. Appropriate alarm management can mitigate these early alarms and even use the early alarms as an asset. For instance, reactive systems could increase chlorination when a deterioration of the water quality is detected, either preventing a health-critical deterioration or providing additional time for intervention.
2. Because MBR+Cl_2-based water reuse systems such as the Water Wall provide robust treatment, they are in standard, undisrupted operation most of the time, with water quality meeting the microbial quality targets (Schoen et al., 2018). The TPR for realistic operation, standard operation with four disruptions per year, was thus $> 99\%$ in most cases. These results also highlight the importance of training and validating models on a disruption dataset and using the FPR and TPR to evaluate model performance. Because of the high number of cases classified as safe in the realistic operation dataset, a trivial classifier that predicts every case as safe could still have achieved a very high overall accuracy but missed all the cases when the water did not meet the microbial quality targets (Monard and Batista, 2002).

It is possible that model performance could be further improved through an optimized sampling strategy. In this study, it was not always possible to capture the exact moment of change in microbial water quality, leading to a datasets that mostly consists of very good or very poor water quality but misses intermediate water qualities. This is for example visible in Fig. 6.A, where the single false positive prediction was for an intermediate water quality. The prediction models developed in this study could be used to inform researchers or practicioners when changes in microbial water quality are most likely and additional samples should be taken.

5.3. Prediction models have to be trained for specific microbial water quality targets

In this study, we trained models for two virus LRTs and two maximum concentrations of intact cells. For actual deployment, the models must be trained for system-specific microbial water quality targets. For virus removal, the LRTs will depend on the wastewater composition, reuse application, and scale of the water reuse system and can be taken from water reuse frameworks (if the specific wastewater compositions and reuse applications are covered) or calculated using quantitative microbial risk assessment (WHO, 2016). Water reuse frameworks typically include no requirements for the ICC, as the ICC is not representative of pathogens. However, several studies have suggested that the ICC can be used to measure process performance (Cheswick et al., 2019; Van Nevel et al., 2017). Here, baseline measurements of ICC during standard operation can provide information on the variability of anticipated ICC levels and help select appropriate targets.

In the present study, we observed that it is possible to train the models for a range of microbial water quality targets. However, the models could no longer distinguish between safe and unsafe states for the more stringent water quality target for ICC ($\log_{10}$ICC ≤ 4). This result highlights one requirement for the prediction: the quality target cannot be too close to the standard operation data, or standard operation is classified wrongly too often. Monitoring models can be trained for a range of microbial water quality targets, but water quality targets cannot be arbitrarily chosen for a given system. For the $\log_{10}$ICC ≤ 4 target, the distinction between safe and unsafe system states would have been difficult even for humans, as ICC values sometimes rose above the set threshold in standard operation.

6. Limitations and future directions

6.1. Generalizability

Many machine learning models can yield good results with training data but perform poorly on new or unseen samples, because the ability of conventional models to extrapolate to data outside of the training range is limited. In this study, we tested two setups of the Water Wall treating different wastewater inputs. Our results indicate that the algorithms can be applied to different water types. We divided the disruption dataset into a training and validation dataset and a test dataset. Consequently, by including only one disruption type in the test dataset, we were able to verify that the models can predict unseen experiments. However, the experimental setup used here does not enable us to assess the generalizability of the models to completely new reactors.

The best-performing models rely solely on ORP and FC measurements that have mechanistic relationships with the microbial water quality, and we therefore expect the models to be generalizable to new MBR+Cl_2 systems. In this study, including sensor slopes did not

significantly increase the TPR for models using ORP+FC measurements as inputs. We term the inclusion of sensor slopes hybrid modeling, as these engineered features allowed the models to differentiate between two system states, namely deterioration (typically a slow process) and improvement (typically a fast process) of the microbial water quality. However, we hypothesize that hybrid modeling might also increase transferability to new systems, because other MBR+Cl_2 systems may have other baseline values during standard operation, especially for ORP. This could not be tested in the present study, as both Water Walls had similar ORP levels. Schneider et al. (2020) showed the feasibility of monitoring the physicochemical water quality of on-site wastewater treatment plants using engineered features such as inflection points of sensor measurements without plant-specific retuning of the soft sensors.

Assessing the generalizability of these models to new MBR+Cl_2 systems and ensuring their applicability to real-life contexts will require them to be validated in operational MBR+Cl_2 water reuse systems. One particular challenge will be verifying virus removal, as the spiking of MS2 is not possible in such systems. Here, using indicators such as coliphage that are naturally present could be an alternative but will likely need to be coupled with concentration methods due to low concentrations in the wastewater that do not allow the verification of high log-removals.

6.2. *Increasing the redundancy of the monitoring system*

Our results quantify how a decrease in the number of sensors (FC+ORP vs. ORP alone) increases the rate of false negative predictions, but they also show that these FNs are effectively early alarms. For the actual deployment of such monitoring approaches, selecting the most suitable models and developing alarm management strategies will depend on the monitoring and maintenance scheme envisioned (Bedell et al., 2022). This study shows that it is possible to achieve the highest TPRs with only a few sensors (e.g., ORP+FC, or ORP alone). This is beneficial in terms of cost and model interpretability, but predictions rely more heavily on single sensors, with unclear consequences for long-term implementation where sensors might drift or even fail at some point. In on-site water reuse systems, regular checks and maintenance by humans are a challenge (Schneider et al., 2019), but several solutions may be implemented for increased robustness in real-life deployments:

1. Add monitoring redundancy by implementing several identical sensors. This is a particularly interesting option for ORP-based models because ORP sensors are relatively inexpensive, making this a financially more viable option than FC+ORP-based models. Monitoring redundancy can be helpful when reconstructing sensors: where a sensor validation system detects faulty sensors, the remaining measurements can be used to reconstruct the faulty sensors, thus allowing more robust process monitoring (Yoo et al., 2008).
2. Include plausibility checks by including anomaly detection strategies, for instance by using active learning with human expert labeling (Russo et al., 2020) or automatically evaluating credible values and their mutual compatibility (Isermann, 2011). For instance, in monitoring systems using ORP and FC sensors, the Nernst equation, which permits the calculation of the reduction potential of a reaction, can be used to verify the compatibility between the two measurements.
3. Introduce deliberate system dynamics to test that sensors are reacting as expected, for instance by registering a daily chlorine peak. This strategy has been suggested by Thürlimann et al. (2019). They achieved long-term stability of a urine nitrification process that relied on pH control by introducing regular intended changes in pH and using information about the sensor's reaction to correct for sensor drift.

7. Conclusion

- When working with microbial water quality data, the inherently small size of such laboratory-analysis-derived datasets will constrain the selection of suitable machine learning algorithms. In this study, we show that simple models using logistic regression predicted the microbial water quality in MBR+Cl_2 water reuse systems as well as did more complex black box models, with the additional advantages of better interpretability and, potentially, transferability to other systems.
- The main goal of the monitoring models developed here was to ensure user safety by limiting false positive predictions that water is safe when actually it is unsafe. We observed that the monitoring models are well-suited for early warning systems, as the resulting false negative predictions that water is unsafe when actually it is were not only false alarms but also early warnings.
- Monitoring algorithms need to be trained for specific applications. In this study, we successfully trained and tested models for microbial water quality targets corresponding to different reuse applications. However, the results also illustrate the limitation of this approach, with false classifications occurring when the microbial water quality targets are too close to the standard operation data.
- For some models, the TPR was improved by including sensor slopes, an engineered feature selected through mechanistic understanding of the system. Although this study does not systematically investigate the application of engineered features, the results highlight the potential of hybrid models that include system understanding to improve prediction accuracies and reduce costs by requiring fewer sensors.

CRediT authorship contribution statement

Eva Reynaert: Conceptualization, Methodology, Data collection, Formal analysis, Visualization, Writing – original draft, Writing – review & editing; **Philipp Steiner:** Methodology, Formal analysis, Visualization, Writing – review & editing; **Qixing Yu:** Data collection, Formal analysis; **Lukas D'Olif:** Data collection, Formal analysis; **Noah Joller:** Data collection, Formal analysis; **Mariane Y. Schneider:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing; **Eberhard Morgenroth:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

All data and code is available on https://doi.org/10.25678/000885.

Acknowledgments

We thank ETH Zurich and Eawag for providing internal funding for this research, and the Japan Society for the Promotion of Science (JSPS) for providing funding to MYS (grant P20763). We also thank Richard Fankhauser, Marco Kipf, and Martin Breitenstein for support with hardware; Karin Rottermann and Sylvia Richter for performing the chemical analyses; Flavia Gretener for supporting initial data collection; Andreas Scheidegger and Andreas Frömelt for helpful discussions on statistical analyses and machine learning methods; and Simon Milligan for language advice.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2023.120075.

References

Aliashrafi, A., Zhang, Y., Groenewegen, H., Peleato, N.M., 2021. A Review of Data-Driven Modelling in Drinking Water Treatment. Rev. Environ. Sci. Bio/Technol. 20 (4), 985–1009.

Bedell, E., Harmon, O., Fankhauser, K., Shivers, Z., Thomas, E., 2022. A Continuous, in-Situ, near-Time Fluorescence Sensor Coupled with a Machine Learning Model for Detection of Fecal Contamination Risk in Drinking Water: design, Characterization and Field Validation. Water Res., 118644

Branch, A., Trinh, T., Carvajal, G., Leslie, G., Coleman, H.M., Stuetz, R.M., Drewes, J.E., Khan, S.J., Le-Clech, P., 2016. Hazardous Events in Membrane Bioreactors–Part 3: impacts on Microorganism Log Removal Efficiencies. J. Memb. Sci. 497, 514–523.

Briscoe, E., Feldman, J., 2011. Conceptual Complexity and the Bias/Variance Tradeoff. Cognition 118 (1), 2–16.

Cheswick, R., Cartmell, E., Lee, S., Upton, A., Weir, P., Moore, G., Nocker, A., Jefferson, B., Jarvis, P., 2019. Comparing Flow Cytometry with Culture-Based Methods for Microbial Monitoring and as a Diagnostic Tool for Assessing Drinking Water Treatment Processes. Environ. Int. 130, 104893.

Diaz-Elsayed, N., Rezaei, N., Guo, T., Mohebbi, S., Zhang, Q., 2019. Wastewater-Based Resource Recovery Technologies across Scale: a Review. Resour. Conserv. Recycl. 145, 94–112.

Domènech, L., Saurí, D., 2010. Socio-Technical Transitions in Water Scarcity Contexts: public Acceptance of Greywater Reuse Technologies in the Metropolitan Area of Barcelona. Resour. Conserv. Recycl. 55 (1), 53–62.

Foschi, J., Turolla, A., Antonelli, M., 2021. Soft Sensor Predictor of E. Coli Concentration Based on Conventional Monitoring Parameters for Wastewater Disinfection Control. Water Res. 191, 116806.

Garner, E., Inyang, M., Garvey, E., Parks, J., Glover, C., Grimaldi, A., Dickenson, E., Sutherland, J., Salveson, A., Edwards, M.A., 2019. Impact of Blending for Direct Potable Reuse on Premise Plumbing Microbial Ecology and Regrowth of Opportunistic Pathogens and Antibiotic Resistant Bacteria. Water Res. 151, 75–86.

Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-Derived Soft-Sensors for Biological Wastewater Treatment Plants: an Overview. Environ. Model. Softw. 47, 88–107.

Hirani, Z.M., Bukhari, Z., Oppenheimer, J., Jjemba, P., LeChevallier, M.W., Jacangelo, J. G., 2014. Impact of Mbr Cleaning and Breaching on Passage of Selected Microorganisms and Subsequent Inactivation by Free Chlorine. Water Res. 57, 313–324.

Isermann, R., 2011. Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-Tolerant Systems. Springer Science & Business Media.

Khalid, S., Khalil, T., Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. 2014 science and information conference, pp. 372–378. IEEE.

Künzle, R., Pronk, W., Morgenroth, E., Larsen, T.A., 2015. An Energy-Efficient Membrane Bioreactor for on-Site Treatment and Recovery of Wastewater. J. Water, Sanit. Hyg. Dev. 5 (3), 448–455.

Lesjean, B., Tazi-Pain, A., Thaure, D., Moeslang, H., Buisson, H., 2011. Ten Persistent Myths and the Realities of Membrane Bioreactor Technology for Municipal Applications. Water Sci. Technol. 63 (1), 32–39.

Levy, K., Nelson, K.L., Hubbard, A., Eisenberg, J.N., 2012. Rethinking Indicators of Microbial Drinking Water Quality for Health Studies in Tropical Developing Countries: case Study in Northern Coastal Ecuador. Am. J. Trop. Med. Hyg. 86 (3), 499.

Monard, M.C., Batista, G., 2002. Learning with Skewed Class Distributions. Advances in Logic. Artif. Intellig. Robot. 85, 173–180.

Nocker, A., Schulte-Illingheim, L., Müller, H., Rohn, A., Zimmermann, B., Gaba, A., Nahrstedt, A., Mohammadi, H., Tiemann, Y., Krömer, K., 2020. Microbiological Changes Along a Modular Wastewater Reuse Treatment Process with a Special Focus on Bacterial Regrowth. J. Water Reuse Desalin. 10 (4), 380–393.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., Akinjobi, J., 2017. Supervised Machine Learning Algorithms: classification and Comparison. Int. J. Comp. Trend Tech. (IJCTT) 48 (3), 128–138.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-Learn: machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peter-Varbanets, M., Hammes, F., Vital, M., Pronk, W., 2010. Stabilization of Flux During Dead-End Ultra-Low Pressure Ultrafiltration. Water Res. 44 (12), 3607–3616.

Reynaert, E., Greenwood, E., Ndwandwe, B., Riechmann, M., Udert, K., Morgenroth, E., 2020. Practical Implementation of True on-Site Water Recycling Systems for Hand Washing and Toilet Flushing. Water Res. X, 100051.

Reynaert, E., Gretener, F., Julian, T.R., Morgenroth, E., 2023. Sensor Setpoints That Ensure Compliance with Microbial Water Quality Targets for Membrane Bioreactor and Chlorination Treatment in on-Site Water Reuse Systems. Water Res. X 18, 100164.

Reynaert, E., Hess, A., Morgenroth, E., 2021. Making Waves: why Water Reuse Frameworks Need to Co-Evolve with Emerging Small-Scale Technologies Water Research. Water Res. X, 100094.

Rodriguez, D.J., Serrano, H.A., Delgado, A., Nolasco, D., Saltiel, G., 2020. From Waste to Resource: Shifting paradigms for smarter wastewater interventions in Latin America and the Caribbean. World Bank.

Russo, S., Lürig, M., Hao, W., Matthews, B., Villez, K., 2020. Active Learning for Anomaly Detection in Environmental Data. Environ. Model. Softw. 134, 104869.

Schneider, M.Y., Carbajal, J.P., Furrer, V., Sterkele, B., Maurer, M., Villez, K., 2019. Beyond Signal Quality: the Value of Unmaintained Ph, Dissolved Oxygen, and Oxidation-Reduction Potential Sensors for Remote Performance Monitoring of on-Site Sequencing Batch Reactors. Water Research.

Schneider, M.Y., Furrer, V., Sprenger, E., Carbajal, J.P., Villez, K., Maurer, M., 2020. Benchmarking Soft Sensors for Remote Monitoring of on-Site Wastewater Treatment Plants. Environ. Sci. Technol. 54 (17), 10840–10849.

Schneider, M.Y., Quaghebeur, W., Borzooei, S., Froemelt, A., Li, F., Saagi, R., Wade, M.J., Zhu, J.-.J., Torfs, E., 2022. Hybrid Modelling of Water Resource Recovery Facilities: status and Opportunities. Water Sci. Technol. 85 (9), 2503–2524.

Schoen, M.E., Ashbolt, N.J., Jahne, M.A., Garland, J., 2017. Risk-Based Enteric Pathogen Reduction Targets for Non-Potable and Direct Potable Use of Roof Runoff, Stormwater, and Greywater. Microbial Risk Analysis 5, 32–43.

Schoen, M.E., Jahne, M.A., Garland, J., 2018. Human Health Impact of Non-Potable Reuse of Distributed Wastewater and Greywater Treated by Membrane Bioreactors. Microbial Risk Analysis 9, 72–81.

Smeets, P.W., 2010. Stochastic Modelling of Drinking Water Treatment in Quantitative Microbial Risk Assessment. IWA Publishing.

Storey, M.V., Van der Gaag, B., Burns, B.P., 2011. Advances in on-Line Drinking Water Quality Monitoring and Early Warning Systems. Water research 45 (2), 741–747.

Thürlimann, C.M., Udert, K.M., Morgenroth, E., Villez, K., 2019. Stabilizing Control of a Urine Nitrification Process in the Presence of Sensor Drift. Water Res. 165, 114958.

Van Nevel, S., Koetzsch, S., Proctor, C.R., Besmer, M.D., Prest, E.I., Vrouwenvelder, J.S., Knezev, A., Boon, N., Hammes, F., 2017. Flow Cytometric Bacterial Cell Counts Challenge Conventional Heterotrophic Plate Counts for Routine Microbiological Drinking Water Monitoring. Water Res. 113, 191–206.

Van Rossum, G., Drake, F.L., 2009. Python/C Api Manual-Python 3.

Weller, D.L., Love, T.M., Wiedmann, M., 2021. Interpretability Versus Accuracy: a Comparison of Machine Learning Models Built Using Different Algorithms, Performance Measures, and Features to Predict E. Coli Levels in Agricultural Water. Front. Artif. Intell. 4, 628441.

WHO, 2016. Quantitative microbial risk assessment: application for water safety management. Geneva, Switzerland: World Health Organization.

Wilcox, J., Nasiri, F., Bell, S., Rahaman, M.S., 2016. Urban Water Reuse: a Triple Bottom Line Assessment Framework and Review. Sustain. Cities Soc. 27, 448–456.

Wong, T.-.T., 2015. Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation. Pattern Recognit. 48 (9), 2839–2846.

Yoo, C.K., Villez, K., Van Hulle, S.W., Vanrolleghem, P.A., 2008. Enhanced Process Monitoring for Wastewater Treatment Systems. Environmetrics 19 (6), 602–617.

Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., 2017. Machine Learning on Big Data: opportunities and Challenges. Neurocomputing 237, 350–361.

Zhu, Y., Chen, R., Li, Y.-.Y., Sano, D., 2020. Virus Removal by Membrane Bioreactors: a Review of Mechanism Investigation and Modeling Efforts. Water Res., 116522