Contents lists available at ScienceDirect

Science of the Total Environment

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

Classification machine learning to detect de facto reuse and cyanobacteria at a drinking water intake

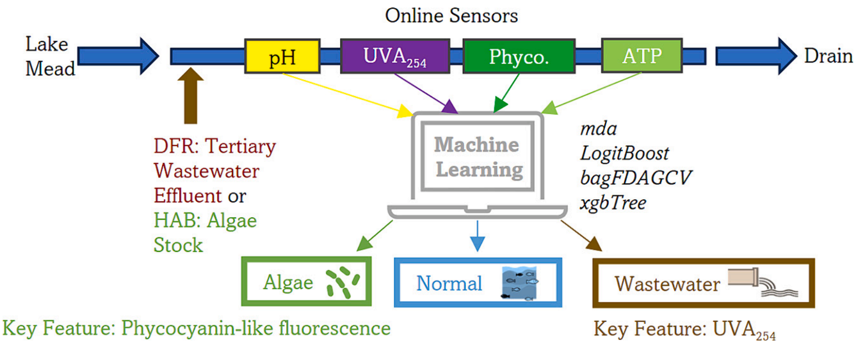
Emily Clements [a,1], Kyle A. Thompson [a,b,1], Deena Hannoun [a], Eric R.V. Dickenson [a,*]

[a] Southern Nevada Water Authority, 1299 Burkholder Blvd., Henderson, NV 89015, USA
[b] Carollo Engineers, Inc., 10900 Stonelake Blvd Bldg 2 Ste 126, Austin, TX 78759, USA

HIGHLIGHTS

- Supervised machine learning can detect very low levels of algae and recycled water.
- Mixture discriminant analysis had a testing set accuracy >98 % for the 3 classes.
- Phycocyanin-like fluorescence helped identify algal bloom events.
- UVA_{254} helped identify higher levels of wastewater effluent.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Jurgen Mahlknecht

Keywords:
Harmful algal bloom
Lake Mead
Phycocyanin
Mixture discriminant analysis
UVA_{254}

ABSTRACT

Harmful algal blooms (HABs) or higher levels of de facto water reuse (DFR) can increase the levels of certain contaminants at drinking water intakes. Therefore, the goal of this study was to use multi-class supervised machine learning (SML) classification with data collected from six online instruments measuring fourteen total water quality parameters to detect cyanobacteria (corresponding to approximately 950 cells/mL, 2900 cells/mL, and 8600 cells/mL) or DFR (0.5, 1 and 2 % of wastewater effluent) events in the raw water entering an intake. Among 56 screened models from the *caret* package in R, four (*mda*, *LogitBoost*, *bagFDAGCV*, and *xgbTree*) were selected for optimization. *mda* had the greatest testing set accuracy, 98.09 %, after optimization with 7 false alerts. Some of the most important water parameters for the different models were phycocyanin-like fluorescence, UVA_{254}, and pH. SML could detect algae blending events (estimated <9000 cells/mL) due in part to the phycocyanin-like fluorescence sensor. UVA_{254} helped identify higher concentrations of DFR. These results show that multi-class SML classification could be used at drinking water intakes in conjunction with online instrumentation to detect and differentiate HABs and DFR events. This could be used to create alert systems for the water utilities at the intake, rather than the finished water, so any adjustment to the treatment process could be implemented.

[*] Corresponding author.
E-mail address: eric.dickenson@snwa.com (E.R.V. Dickenson).
[1] These authors contributed equally.

https://doi.org/10.1016/j.scitotenv.2024.174690
Received 22 April 2024; Received in revised form 25 June 2024; Accepted 8 July 2024
Available online 9 July 2024

1. Introduction

Water quality events such as harmful algal blooms (HABs) and high levels of de facto wastewater reuse (DFR) are potentially hazardous to the public. Cyanobacteria (i.e., blue-green algae) are photosynthetic prokaryotes commonly associated with HABs as some strains are capable of producing taste and odor compounds or toxins (Liu et al., 2024; Shen et al., 2023). For example, under some environmental conditions some strains of the species *Microcystis aeruginosa* are capable of producing the toxin Microcystin-LR (Ouahid et al., 2005), a possible carcinogen in humans (Pan et al., 2021) and a potent hepatotoxin (Weng et al., 2007). This is merely one of many toxins that can be produced by algal species (Centers for Disease Control and Prevention, 2016). Concentrations of cyanobacteria can increase rapidly under certain environmental conditions, including high temperatures, high nutrients, and slow water movement. If some of these cyanobacteria are toxin-producing, these events are called HABs (Beaver et al., 2018). According to the United States Environmental Protection Agency (EPA), there were at least 281 notices for freshwater HABs from June 2nd to August 1st, 2017 in the United States (U.S. Environmental Protection Agency, 2019), and Schaeffer et al. (2022) found an increase in HAB extent in the U.S. from 2017 to 2020 of 6.9 %. Furthermore, climate change is expected to increase the frequency, severity, and duration of HABs by increasing water temperature and thermal stratification (Paerl and Huisman, 2009). Lake Mead, a drinking water source for approximately 40 million people in Nevada, Arizona, and California, is one of the bodies of water in which these climate-affected HABs could be expected to occur with increased frequency (Beaver et al., 2018; Milly and Dunne, 2020). Long-term monitoring has revealed the phytoplankton communities in Lake Mead have remained relatively stable (Beaver et al., 2018), as have the phosphorous concentrations, the nutrient that could drive an algal bloom, and the chlorophyll *a* concentrations (Hannoun and Tietjen, 2023).

Online instruments have become available for indirect detection of cyanobacteria. While oxygen production rate has been proposed as a proxy of microalgae concentration (Sarrafzadeh et al., 2015), elucidating the impact of the algae in a complex water matrix would be too difficult. Chlorophyll *a*, a pigment used in photosynthesis, can be detected via absorbance or fluorescence (Choo et al., 2018; Myers et al., 2013). However, chlorophyll *a* is found in both cyanobacteria and most other algae (Choo et al., 2018). Fluorometers can also target the characteristic wavelengths of phycocyanin (590–610 nm excitation and 660–685 nm emission) (Choo et al., 2018), a pigment found in cyanobacteria but not green algae (Chang et al., 2012). Online instrumentation is also commercially available for adenosine triphosphate (ATP), an energy-carrying molecule in all known lifeforms. ATP has been proposed for cyanobacteria monitoring due to its sensitive detection threshold and strong correlations with cyanobacteria cell counts and optical measures in controlled conditions (Greenstein and Wert, 2019). However, ATP is not specific to cyanobacteria because of its presence in other biologically active cells.

Another water quality event, DFR, occurs when there is wastewater effluent at the drinking water intake and is widespread globally. In the USA, approximately 25 % of drinking water intakes for drinking water treatment plants serving >10,000 people were estimated to have >1 % DFR under average streamflow conditions (Rice et al., 2015). The percentage of DFR is highly dependent on streamflow conditions. For example, DFR in the Llobregat River in Spain was estimated to vary from 8 to 82 % (Drewes et al., 2017) and the Neosho River in Oklahoma is estimated to vary from around 2 % under median conditions to 100 % at the 5th percentile flow (Rice and Westerhoff, 2015). In lakes and reservoirs, the percentage of recycled water is less variable due to the larger dilution volume. DFR at the Southern Nevada Water Authority's drinking water intake in Lake Mead is currently approximately 1.4 % DFR (Hannoun et al., 2021) but is expected to increase over time if lake levels decline in response to climate change and drought (Milly and Dunne, 2020). DFR can increase concentrations of anthropogenic chemicals, pathogens, and fecal indicators. Burnet et al. (2019) found a statistically significant correlation ($p < 0.05$) between flow at an upstream wastewater treatment plant (WWTP) and ß-D-glucuronidase activity, which in turn correlates with *E. coli*. DFR correlated with disinfection byproducts (DBPs) in drinking water systems in a watershed in Virginia (Weisman et al., 2019). Modeled DFR and the DFR indicator sucralose strongly correlated with per- and polyfluoroalkyl substances (PFAS) in the Trinity River, Texas (Islam et al., 2023).

Alert systems are needed to notify drinking water utilities when water quality events such as HABs or sudden increases in DFR are beginning. Supervised classification machine learning applied to online water quality data could potentially provide such a system. Most previous studies applying machine learning for event detection in the drinking water field have focused on treated water in real or simulated distribution systems (Arad et al., 2013; Asheri-Arnon et al., 2018; Dogo et al., 2019). However, an alert system monitoring raw water at the drinking water intake could detect events sooner, thus allowing more time for corrective action. For example, mitigation strategies, such as source control for nutrients promoting algae growth or increased ozone dosing for algal toxins, could be implemented. Kibuye et al. (2021) conducted a survey of 35 drinking water utilities and found that 68 % of the utilities currently or had in the past implemented control strategies to mitigate cyanobacterial blooms in their source water. The most frequently used control strategies were aeration and algaecides, such as copper sulfate or hydrogen peroxide.

In Lake Mead, increases in DFR would be much slower due to the dilution, but knowing the trends would be useful for planning purposes. Utilities with intakes at rivers or smaller volume reservoirs are more likely to see spikes in DFR, so an alert system could be more beneficial. Many published classification methods are binary (i.e., classify observations as either "Normal" or "Anomaly") (Liu et al., 2020). However, multiclass models that can predict among three or more categories [i.e., weighted k-nearest neighbors (*kknn*)] could provide more information about the best course of action to utility operators. Alert systems in the drinking water context must also be highly specific (e.g., less than one false positive per week) to avoid complacency or poor allocation of resources. However, achieving high specificity and sensitivity with raw surface water quality data can be challenging due to instrument drift or limited sensitivity, or differentiating natural diurnal and seasonal patterns from events.

This study used supervised machine learning (SML) to rapidly detect low levels of cyanobacteria in the raw water, enabling prompt detection of the early onset of HAB events at the intake, allowing time for mitigation. While SML has been applied for HABs but focused on forecasting major HABs weeks in the future (Fleming et al., 2019; Jeong et al., 2022; Kim et al., 2021), this study focused on real time detection. Furthermore, this study used multi-class SML classification to simultaneously monitor for and detect low levels of DFR. While a previous study applied SML to detect increases in DFR at levels of 2 % or more in surface waters (Thompson and Dickenson, 2021), this study tested 0.5 % to 2 % DFR levels. The models in the previous study could not reliably differentiate between DFR and other events, such as stormwater. This study improved upon that limitation through greater true positive sample size and additional instrumentation. Overall, this study demonstrated the potential of SML for water quality monitoring, demonstrating that low concentrations of DFR and HABs at a drinking water intake can be detected and differentiated. This would allow for more proactive water management which could be necessary to ensure safe drinking water and protect human health if water quality declines.

2. Methods

In this study, online water quality instruments measured water quality in a pipe conveying raw water from Lake Mead. The data from these instruments was recorded at 15-min intervals for one month.

During this month, blending tests were conducted in which cyanobacteria or wastewater effluent were injected into the instrument sampling line at known blending ratios. Observations were labeled “Normal,” “Algae,” or “Wastewater.” The water quality data was then divided into training and testing sets. Fifty-six classification models were screened on the raw data with their default hyperparameters in the *caret* package (version 6.0.94) (Kuhn, 2020, 2008) in R (version 4.3.1) (R Core Team, 2023). Four high performing models were then further evaluated with (1) a wider range of hyperparameters, (2) the least important water quality feature(s) omitted, and (3) various preprocessing methods to correct for instrument drift, natural patterns, or random instrument error.

2.1. *Study area*

Lake Mead is located on the Colorado River, on the border between Nevada and Arizona (Fig. 1) and is the largest reservoir by volume in the United States (Holdren and Turner, 2010). It provides water for drinking and irrigation for over 25 million people (Holdren and Turner, 2010), so maintaining the water quality and quantity is essential. However, drought conditions have persisted in reservoirs in the southwestern U.S. since 2000 (Beaver et al., 2018), and between 2000 and 2022, Lake Mead has experienced a 71 % decrease in volume (Hannoun and Tietjen, 2023).

2.2. *Online instruments*

Six online instruments measured fourteen water quality features and flow in a pipe conveying raw water from Lake Mead (Table S1). A RealTech PL3500 (Ontario, Canada) measured absorbance at multiple wavelengths along the UV–visible spectrum (200–800 nm). It recorded three features: UVA_{254}; an estimate of total organic carbon (TOC) (expressed in mg/L) based on absorbances at 256, 280, and 322 nm; and an estimate of chlorophyll *a* (expressed in μg/L) based on absorbances at 687, 689, 692, 719, and 722 nm. A BlueI 702 TurbiPlus Hydroguard (Rosh Ha'ayin, Israel), which is no longer commercially available, monitored basic water quality features, including turbidity, conductivity, pH, temperature, and redox potential (ORP). An Atrato Series 700 Ultrasonic Inline Liquid Flowmeter (United Kingdom) measured total flow to the water quality instruments. A Ketos Shield (United States) measured nitrate with a specified method reporting limit (MRL) of 0.5 mg/L, as well as dissolved oxygen (DO). A Turner Designs Enviro-T2 In-Line Fluorometer (California, United States) with Red Excitation LED monitored a phycocyanin-like fluorescence. Relative fluorescence units were converted to an estimated concentration in μg/L based on manufacturer instructions. A Hach EZ7300 Series Online Microbiology Load Analyzer for ATP (United States) measured total, free, and intracellular ATP.

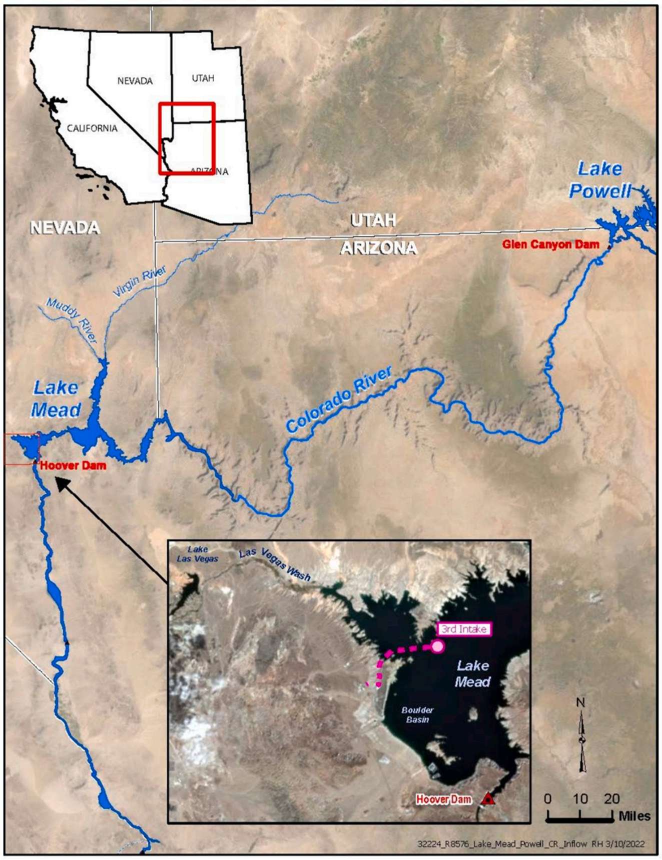


Fig. 1. Map of Lake Mead.

All instruments were maintained and calibrated, zeroed, or validated on August 29th, 2019, two days before the start of the training set. The RealTech PL3500 cleaned itself automatically with a calcium, lime, and rust removal solution every 48 h and it contained a dehumidifier that was recharged weekly on Thursdays. The ATP Analyzer automatically cleaned itself daily with NaOH and HCl and was recalibrated on September 19th, 2019. The BlueI Hydroguard was manually cleaned every Thursday. See Table S2 for a timeline of maintenance and blending events.

2.3. *Benchtop water quality methods*

The water quality of the raw Lake Mead water and the stocks to be used for blending events were measured offline to (1) assess what level of blending or dilution might be detectable and (2) check the stability of the stocks between the training set blending event and testing set blending event (Section 2.7). pH, conductivity, turbidity, and UVA_{254} were measured as described in Thompson and Dickenson (2021). UVA_{254} was measured with 0.45 μm filtration as the standard method and without prefiltration to simulate UVA_{254} as measured by the Realtech PL3500 (Potter and Wimsatt, 2009). ATP was measured using the same instrument as the online ATP data using the instrument's grab sample line. Nitrate and nitrite were determined by ion chromatography with EPA Method 300.0 (Pfaff, 1993) and optical density benchtop measurements were taken with the Spectronic 20D+ (Thermo Scientific).

2.4. *Data frequency, acquisition, and interpolation*

Data was collected for one month, September 2019. The Hydroguard, PL3500, flowmeter, and phycocyanin-like fluorometer had a measurement frequency of 15 min or less and were connected to the drinking water utility's SCADA system via 4–20 mA analog outputs. Data from these instruments were downloaded via PI Datalink at 15-min intervals. Data from the Ketos Shield was downloaded from the instrument's web portal. Nitrate data was collected hourly. Dissolved oxygen (DO) was recorded at irregular intervals ranging from hourly to more than once per minute. Data from the Ketos Shield were interpolated onto the same time vector as the SCADA-compatible instruments using the *approx* function in R with a step function giving full weight to the left value. That is, at each 15-min interval, the Ketos datapoint was the most recently recorded datapoint before that 15-min interval. The ATP Analyzer took measurements at approximately 15-min intervals, though not necessarily in alignment with the time vector of the SCADA-compatible instruments (i.e., not at precisely 12:00, 12:15, etc.). Thus, data from this instrument was also interpolated with the same function applied to the Ketos Shield data. Calibrations and automated cleanings were excluded from the ATP data.

2.5. Time of week feature

Real water quality events—as well as certain potential sources of error—would be more likely to occur at certain times of day or week. Some cyanobacteria have diurnal cycles with photosynthesis peaking in late afternoon and nitrogen fixation occurring after dark (Sherman et al., 1998). Wastewater flow—and thus DFR—has a distinctive diurnal pattern and can differ between weekdays and weekends (Atinkpahoun et al., 2018; Tchobanoglous et al., 2002). Routine sensor maintenance is typically scheduled for certain times of the week (i.e., in the case of this study, Thursday mid-day). Therefore, a feature was calculated to inform models of the time of week. Specifically, values were calculated as the number of seconds since midnight on Sunday divided by the total number of seconds in one week (604,800 s). Thus, a value of 0.001 would represent Monday shortly after midnight, 0.999 would represent Sunday shortly before midnight, and 0.5 would represent Thursday at noon.

2.6. Blending stock sampling and preparation

Tertiary wastewater effluent was collected from a large (>400,000 m3/day) wastewater treatment plant (WWTP) in Nevada and stored at 4 °C for less than one week. The WWTP employs a modified Johannesburg process for biological phosphorus removal followed by alum coagulation and dual media filtration. The final UV-disinfected effluent from this WWTP flows into Lake Mead via the Las Vegas Wash (Blunt et al., 2018). The effluent sample was collected after dual media filtration but before UV disinfection, as was done in Thompson and Dickenson (2021). The water quality of this wastewater effluent sample is in Table S3.

M. aeruginosa is a common species of cyanobacteria in freshwater HABs, many strains of which produce microcystin-LR (Ouahid et al., 2005). However, the strain UTEX 2061 does not produce microcystin-LR and lacks the gene associated with the production of this toxin. Thus, for laboratory safety and to avoid the accidental introduction of a toxic cyanobacterial strain into the local environment, UTEX 2061 was selected as a surrogate for wild *M. aeruginosa* or HABs. UTEX 2061 was purchased from the Culture Collection of Algae at the University of Texas at Austin. In duplicate, 15 mL of stock strain was added to 150 mL of Bold 3 N Medium and incubated in foil-capped 250-mL Erlenmeyer flasks on a shaker table at 150 rpm in an incubator at 20 °C with a 12/12 h light/dark cycle. Based on experiments with a different strain of *M. aeruginosa* incubated under the same conditions, approximately two weeks of lag phase followed by approximately two weeks of rapid growth were anticipated (Greenstein and Wert, 2019). So, the stocks used in the blending events during the training and testing sets were incubated for 20 and 22 days, respectively, to be relatively stable in the middle of the lag phase. The water quality for a 1:100 dilution of the UTEX 2061 with Lake Mead water is in Table S4. Cell counts for the UTEX 2061 stock were calculated from optical density measurements, as described by Greenstein and Wert (2019). UTEX 2061 stock water quality was measured as a dilution in Lake Mead water (1) to simulate *M. aeruginosa* ATP generation in Lake Mead water, (2) to fall within benchtop method calibration ranges, (3) to avoid cell lysis with deionized water, and (4) due to water volume constraints for analysis.

2.7. Blending events

Blending events were conducted as in Thompson and Dickenson (2021). A 0.5-in. PVC pipe conveyed raw water from Lake Mead to the online instruments. The effluent from the instruments went to a septic tank. Wastewater effluent or UTEX 2061 stock were diluted with Lake Mead water at 50:50 or 1:100 ratios, respectively, in a 17-L carboy on a stir plate. The diluted contaminant stocks were then injected into the pipe with a peristaltic pump upstream of an inline static mixer to ensure the waters were well-blended before reaching the instruments. To verify the intended flow rate and blending ratio, the volume in the carboy was manually recorded at 15-min intervals. For each blending event, the diluted wastewater effluent or UTEX 2061 stock was blended into the pipe at three blending ratios for two hours each, consecutively from lowest to highest to simulate the gradual, low-level onset of a real event. True HABs or periods of elevated DFR would likely last longer than the total six hours, but nevertheless this approach sufficed to evaluate the ability of SML to detect the onset of events. The overall blending ratios were 0.5 %, 1 %, and 2 % for wastewater effluent and 0.005 %, 0.015 %, and 0.045 % for the UTEX 2061 stock (corresponding to approximately 950 cells/mL, 2900 cells/mL, and 8600 cells/mL). One blending event of each type was conducted during the training set and then replicated during the testing set.

2.8. Supervised machine learning methods

Data were labeled as “Normal,” “Wastewater”, or “Algae,” based solely on which if any blending events were occurring or if the influent was raw surface water. In terms of classification performance metrics, Wastewater and Algae were considered positives while Normal was considered negative. In practice, multiple approaches could be considered for automated alert systems during times of known maintenance (i.e., a “Maintenance” alert provided, the alert system disarmed prior to maintenance, or alerts ignored during maintenance). In this study, the maintenance events were labeled as Normal.

SML was conducted using the *caret* package in R. The *caret* package was chosen because it allows the screening of a large number of machine learning model types with consistent cross-validation and testing protocols. This creates the possibility of identifying effective but previously underutilized models for time series water quality SML analysis. Further research could use model-specific packages for greater flexibility with hyperparameter tuning. The dataset was split 60:40 into a training set and testing set. While 80:20 is the more common or default data split in machine learning, 60:40 has also been used, including in studies applying SML to surface water quality (Khullar and Singh, 2020; Liu et al., 2023). A smaller training set would be beneficial because it reduces computational requirements and training time, allowing for quicker experiments. The split was consecutive, not random, (1) to ensure one blending event of each type in each set, (2) to enable time series-specific preprocessing methods, and (3) to simulate a scenario in which a drinking water utility were to train models based on full-scale data from actual events to detect future similar events. The training set was from September 1st through 18th and the testing set was from September 19th through 30th, 2019. Before each model training, the seed for random number generation was set to 1 with the *set.seed* function for reproducibility. Models were trained on the training set and hyperparameters were selected based on highest average accuracy over 25 bootstraps of the training set (hereon referred to as “training set accuracy”). Multiple measures of error were calculated including accuracy (the overall percentage of observations for which the model predicted the correct category), sensitivity (how often the model predicted the correct category when a given category was occurring), and total false alerts (how often the model predicted Wastewater or Algae when the correct label was Normal).

Fifty-eight models were initially included in this study based on whether they had achieved at least 90 % testing set accuracy or 98.5 % training set accuracy with raw data from a related prior study (Thompson and Dickenson, 2021). These 58 models were screened on the raw data with default hyperparameter ranges in the *caret* package. Fuzzy Rules Using Chi’s Method (*FRBCS.CHI*) and Fuzzy Rules with Weight Factor (*FRBS.W*) were omitted as too computationally intensive, leaving a total of 56 tested models, including several neural networks (Table S5). Four models were selected for further evaluation based on the following criteria: the two with highest training set accuracy or the two with highest testing set accuracy.

For the four selected models, preprocessing methods were tested as described in Section 2.9. Next, least important features were identified using the *varImp* function and omitted to determine whether equal or greater accuracy could be achieved in equal or less computation time. Lastly, models were tested across a wider range of hyperparameter settings. Fig. 2 shows a schematic of the development of the models.

2.9. Data preprocessing

2.9.1. Smoothing with rolling median

Certain features (e.g., intracellular ATP) appeared to have occasional high outliers that could potentially cause false alerts (Fig. S1). However, these high outliers tended to be non-consecutive (i.e., rarely, if ever, occurring on consecutive 15-min intervals). Thus, a method of smoothing or data cleaning was studied in which each datapoint was replaced by the median of the past three datapoints, including itself. An example with intracellular ATP is shown in Fig. S1. This way, outliers would be omitted unless two occurred consecutively. This method was expected to reduce the number of false positives while only delaying event detection by 15 min.

2.9.2. Difference from rolling median

Many water quality features (i.e., temperature, ORP, pH, and conductivity) had long-term trends due to seasonal patterns or instrument drift. For example, mean daily temperature gradually declined from around 18 °C to around 16 °C over the month of September (Fig. S2). To correct for this drift, the median of the previous 24 h (i.e., 96 datapoints) was calculated. Then, that 24-h rolling median was subtracted from each datapoint. The output of this preprocessing method is hereon referred to as "seasonally adjusted" and is shown for temperature in Fig. S2.

2.9.3. Diurnal modelling

For water quality features with visually apparent diurnal patterns, a sinusoidal model was fit using data during Normal observations in the training set and the *nls* function (Fig. S2). Sinusoidal models were fit based on the equation:

$$x = a + b^*sin\left(\frac{t^*2\pi}{86,400\ s} + c\right) \tag{1}$$

where x is the modeled feature, a is vertical shift, b is amplitude, t is time in seconds, and c is horizontal shift. Predictions from these models were then made for all observations in the training and testing sets. The difference between the sinusoidal model predictions and the seasonally adjusted data were calculated and are hereon referred to as "diurnally adjusted". The seasonal and diurnal adjustment procedures were conducted both with and without prior smoothing as described in Section 2.9.1.

3. Results

3.1. Water quality

The median water quality data for Lake Mead (Normal condition) is shown in Table 1, while the water quality features for the tertiary wastewater effluent and a 1:100 dilution of the UTEX 2061 stock are shown in the SI (Tables S3 and S4). Except for total ATP and intracellular ATP, median water quality was within 10 % between the training set and testing set, indicating little seasonal change within the month of September or sensor drift.

Cell count data was not measured directly for the UTEX 2061 stock or its 1:100 dilution. However, based on a linear model from Greenstein and Wert (2019), the average OD730 from the UTEX 2061 stocks would indicate a cell count of ~1.8 × 10^7 cells/mL, which is the value we used in this study to estimate cell concentration. However, based on another linear model from that same study, the intracellular ATP of the UTEX

Table 1
Median water quality of Lake Mead, blending events omitted.

Feature	Units	Median	
		Training set (n = 1729)	Testing set (n = 1152)
Nitrate	mg/L as N	<0.5	<0.5
Dissolved oxygen	mg/L	7.2	7.2
Flow	L/min	1.54	1.44
Total organic carbon (TOC)	mg/L	2.75	2.75
UV absorbance at 254 nm	1/cm	0.0604	0.0602
Chlorophyll *a*	μg/L	0.0513	0.0475
Phycocyanin-like fluorescence	μg/L	0.62	0.61
Temperature	°C	17.4	15.9
Total ATP	pg/mL	7.39	9.16
Free ATP	pg/mL	3.0	2.7
Intracellular ATP	pg/mL	4.07	6.35
pH		7.89	7.87
Conductivity	μS/cm	869	890
Turbidity	NTU	0.31	0.33
Redox potential (ORP)	mV	830	844

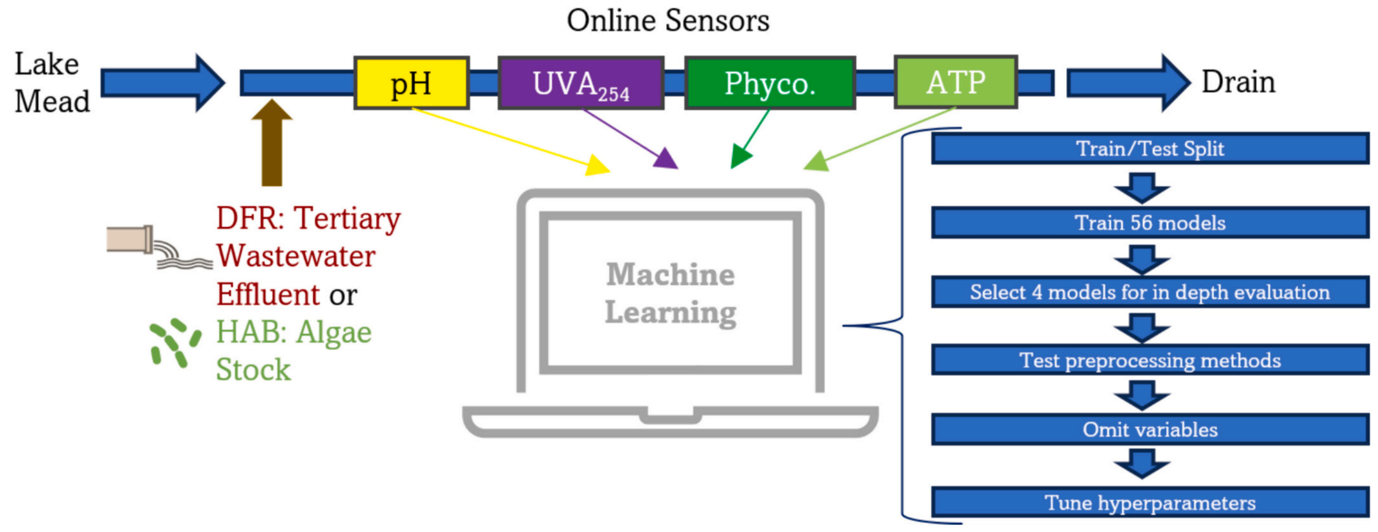


Fig. 2. Schematic of machine learning workflow.

2061 stock—estimated from the 1:100 dilution after subtracting the background intracellular ATP in the Lake Mead water—indicates the cell count would be ~1.5 × 10^6 cells/mL. Since ATP per cell could depend on factors like growth phase, light availability, or health of the culture, OD730 was used for the primary estimate of cell count.

The training set raw data is plotted in Fig. 3. The features are plotted against the standard deviations from their means and grouped by the extreme datapoints, so it is easier to see deviations in the different features. Some of the features, specifically temperature, pH, conductivity, and ORP, had visually apparent diurnal patterns and so were diurnal adjusted based on sinusoidal models as in Section 2.9.3. The sinusoidal models converged and had statistically significant amplitude in all cases (Table S6).

3.2. Screening results

Of the 56 models that were screened on the raw data (Table S5), four were selected for further evaluation: the two with highest training set accuracy and the two with highest testing set accuracy (Table 2). The models selected for further analysis based on these criteria were logistic regression boosting (*LogitBoost*), mixture discriminant analysis (*mda*), bagged flexible discriminant analysis (FDA) using generalized cross-validation (GCV) pruning (*bagFDAGCV*), and extreme gradient boosting (*xgbTree*).

3.3. Mixture discriminant analysis

mda is a modification of linear discriminant analysis in which the distribution of each class is assumed to consist of a mixture of superimposed Gaussian distributions rather than a single Gaussian distribution (Hastie and Tibshirani, 1996). Thus, *mda* is better suited than linear discriminant analysis for non-normally distributed data. Several features were non-normally distributed, as can be seen with Quantile-Quantile plots (Fig. S4). Mixture discriminant analysis (*mda*) had the highest testing set accuracy using the raw data, 97.74 %. This testing set accuracy was significantly above the NIR (p-value = 0.0003). The NIR was 95.83 % and represents the accuracy of a model that always assumes the most common class in the data, which was Normal in this study. *mda* had the second highest cyanobacteria sensitivity (62.5 %) and only 3 false alerts, but the lowest Wastewater sensitivity (41.67 %) among selected models. Smoothing lowered the testing set accuracy from 97.74 % to 94.01 % and seasonally adjusting made the testing set accuracy 97.66 %.

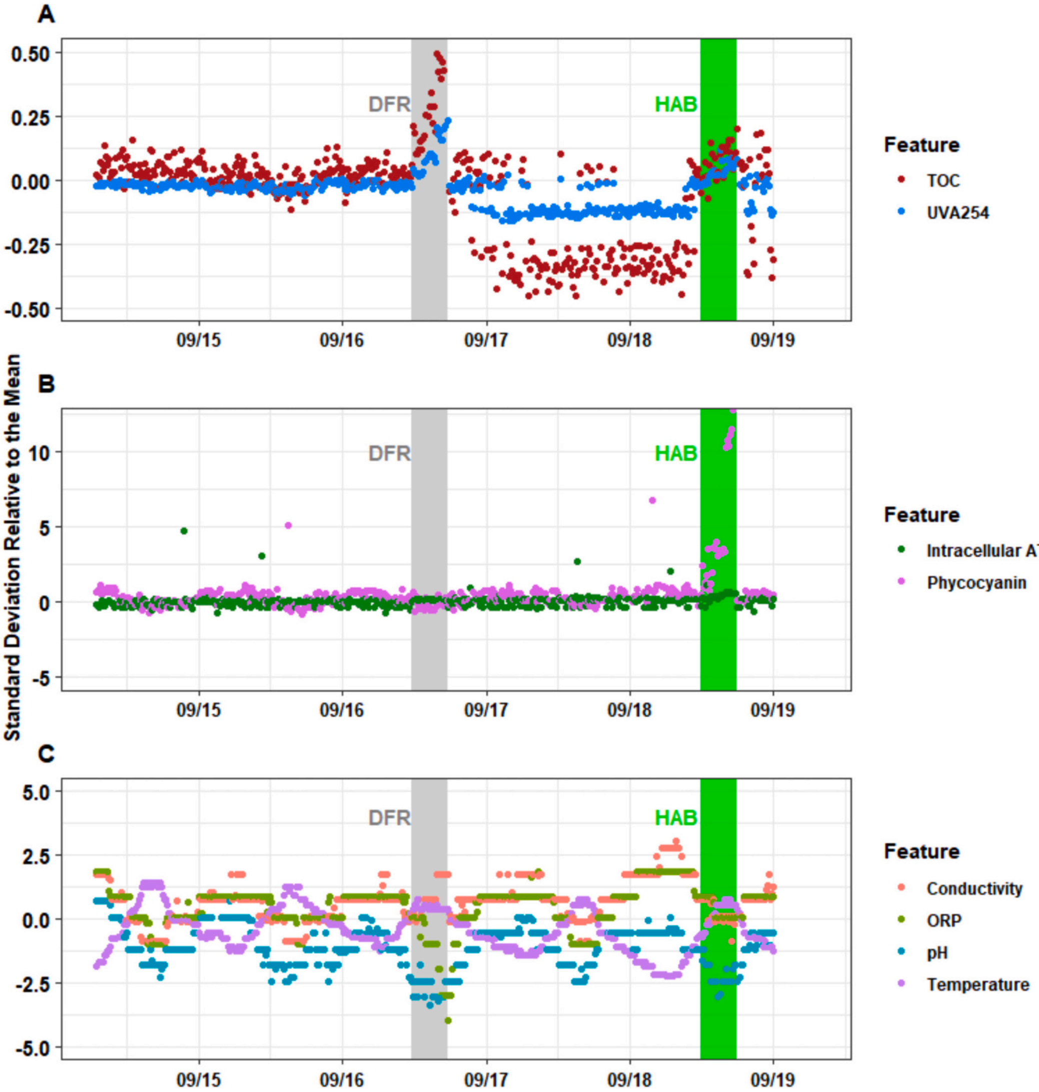


Fig. 3. Training set data for (A) UVA_{254} and TOC; (B) phycocyanin-like fluorescence (phycocyanin) and intracellular ATP; and (C) Conductivity, ORP, pH, and temperature. Shaded gray areas represent blending with wastewater effluent to simulate DPR and shaded green areas represent mixing events with higher concentrations of cyanobacteria. Some outliers are excluded from the range of the y-axis and the whole time period is not shown to highlight the impacts of blending events. Features less impacted by the blending events are also omitted for clarity. The full dataset is provided in the SI (Fig. S3).

Table 2
Summary of screening results with raw data. Full screening results are in Table S5.

Model	Abb.	Training set accuracy	Testing set accuracy	Wastewater effluent sensitivity	Cyanobacteria sensitivity	p-Value	Total false alerts	Reason selected for further analysis
Mixture discriminant analysis	*mda*	98.99 %	97.74 %	41.67 %	62.50 %	3.0×10^{-4}	3	1st testing set accuracy
Logistic regression boosting	*LogitBoost*	99.75 %	92.42 %	72.73 %	12.50 %	1	58	1st training set accuracy
Bagged FDA using GCV pruning	*bagFDAGCV*	99.09 %	97.66 %	75 %	87.50 %	5.7×10^{-4}	18	2nd testing set accuracy
Extreme gradient boosting	*xgbTree*	99.71 %	93.84 %	54.17 %	12.50 %	1	39	2nd training set accuracy

Both seasonally adjusting and smoothing the data resulted in a testing set accuracy of 97.05 %. Therefore, the data was not seasonally adjusted or smoothed. Diurnally adjusting the pH increased the testing set accuracy to 97.83 %, and also diurnally adjusting the temperature and conductivity resulted in a testing set accuracy of 98.09 % (Table 3). From the seven false positives, two were maintenance events labeled as Normal and predicted as Wastewater, four were Normal events predicted as Algae, and one was a Normal event labeled as Wastewater. So, omitting maintenance or separating it as a prediction class would not have rectified the majority of false alerts. Chlorophyll *a* was the least important feature but excluding it decreased the testing set accuracy to 87.32 %. The hyperparameter in *mda* is subclasses, the number of Gaussian distributions assumed to be within the mixture distributions. Over the default range in the *caret* package, two to four, two subclasses resulted in the highest training set accuracy. Increasing the number of subclasses decreased the testing set accuracy. Fig. 4 shows the raw testing set results for *mda*, with false positives and false negatives labeled. False positives occurred when the model predicted an Algae or Wastewater event when it was Normal, and false negatives occurred when it was Normal, and the model predicted an Algae or Wastewater event. The SI contains a zoomed in figure of UVA_{254} and phycocyanin-like fluorescence (Fig. S5) to better see the false positives and false negatives. Fig. S6 is the full dataset for the optimized version of *mda*.

If the data was seasonally adjusted and then optimized by diurnally adjusting conductivity and temperature and omitting flow and nitrate, the testing set accuracy was 98.00 % (*p*-value = 3.58×10^{-5}), with 1 false alert. The cyanobacteria sensitivity was 66.7 % and the Wastewater sensitivity was 41.7 %. While the testing set accuracy is lower, the lower number of false positives could be beneficial, depending on the priorities of the model user, so seasonal adjustment might be advantageous.

3.4. Logistic regression boosting

Logistic regression boosting (*LogitBoost*) is an implementation of gradient boosting framework by Friedman et al. (2000) and Friedman (2001), where a series of weak learners are combined to create a stronger predictive model. *LogitBoost* uses logistic regression as the weak learners for classification. It had the highest training set accuracy on the raw data, 99.75 %, but a testing set accuracy of just 92.42 %, indicating overfitting (Table 2). *LogitBoost* had 58 false alerts, the most of any of the models selected for further testing, a Wastewater sensitivity of 72.72 % and a cyanobacteria sensitivity of 12.50 %. Smoothing the data increased the testing set accuracy to 95.49 % and seasonal adjustment also increased the testing set accuracy, to 96.78 %. However, first smoothing and then applying seasonal adjustment resulted in a lower testing set accuracy of 96.16 % so the data was seasonally adjusted but not smoothed. While diurnal adjustment for each feature individually led to an increase in testing set accuracy to 96.70 % for pH, 97.13 % for conductivity, 96.95 % for ORP, and 96.88 % for temperature, diurnal adjustments on all features had a testing set accuracy of 96.96 %. The highest testing set accuracy (97.13 %) resulted from diurnally adjusting pH and conductivity. Nitrate was the least important feature and excluding it increased the testing set accuracy to 97.82 % but also excluding flow, the next least important feature, reduced the testing set accuracy to 97.04 %. The best tune was with 42 boosting iterations (testing set accuracy of 97.82 %) (Table 3), but it was only 0.006 % more accurate than the default tune, which used 31 iterations. There was only one false alert, in which the model predicted Wastewater when the event was Normal. Although the test set accuracy was less than the NIR on the raw data, after the preprocessing and optimization, the testing set accuracy was significantly over the NIR (*p*-value = 6.98×10^{-4}). Fig. S7 shows the testing set results for *LogitBoost*.

3.5. Bagged FDA using GCV pruning

Bagged flexible discriminant analysis (FDA) using generalized cross-validation (GCV) pruning can capture nonlinear relationships between the predictors and classes with GCV pruning to avoid overfitting. Bagging, or bootstrap aggregating, creates subsets of the training data by sampling with replacement, resulting in certain values being repeated or

Table 3
Summary of optimized models.

Abb.	Preprocessing	Tuning	Features excluded	Cohen's kappa	Testing set accuracy	p-Value	Total false alerts	Most important features
mda	Diurnally adjusted (pH, Cond., Temp)	Default	None	0.745	98.09 %	1.6×10^{-5}	7	Phycocyanin-like fluorescence, day, pH, UVA_{254}, and TOC
LogitBoost	Seasonally adjusted (All) and diurnally adjusted (pH, Cond.)	nIter = 42	Nitrate	0.620	97.82 %	7.0×10^{-4}	1	Conductivity, day, phycocyanin-like fluorescence, UVA_{254}, and chlorophyll *a*.
bagFDAGCV	None	Default	None	0.739	97.66 %	5.7×10^{-4}	18	Phycocyanin-like fluorescence, TOC, pH, UVA_{254}, and nitrate
xgbTree	Seasonally adjusted (All) and diurnally adjusted (pH)	nrounds = 100, max_depth = 1, eta = 0.15, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1, subsample = 1	Turbidity	0.551	97.40 %	3.1×10^{-3}	1	UVA_{254}, phycocyanin-like fluorescence, day, pH, and nitrate

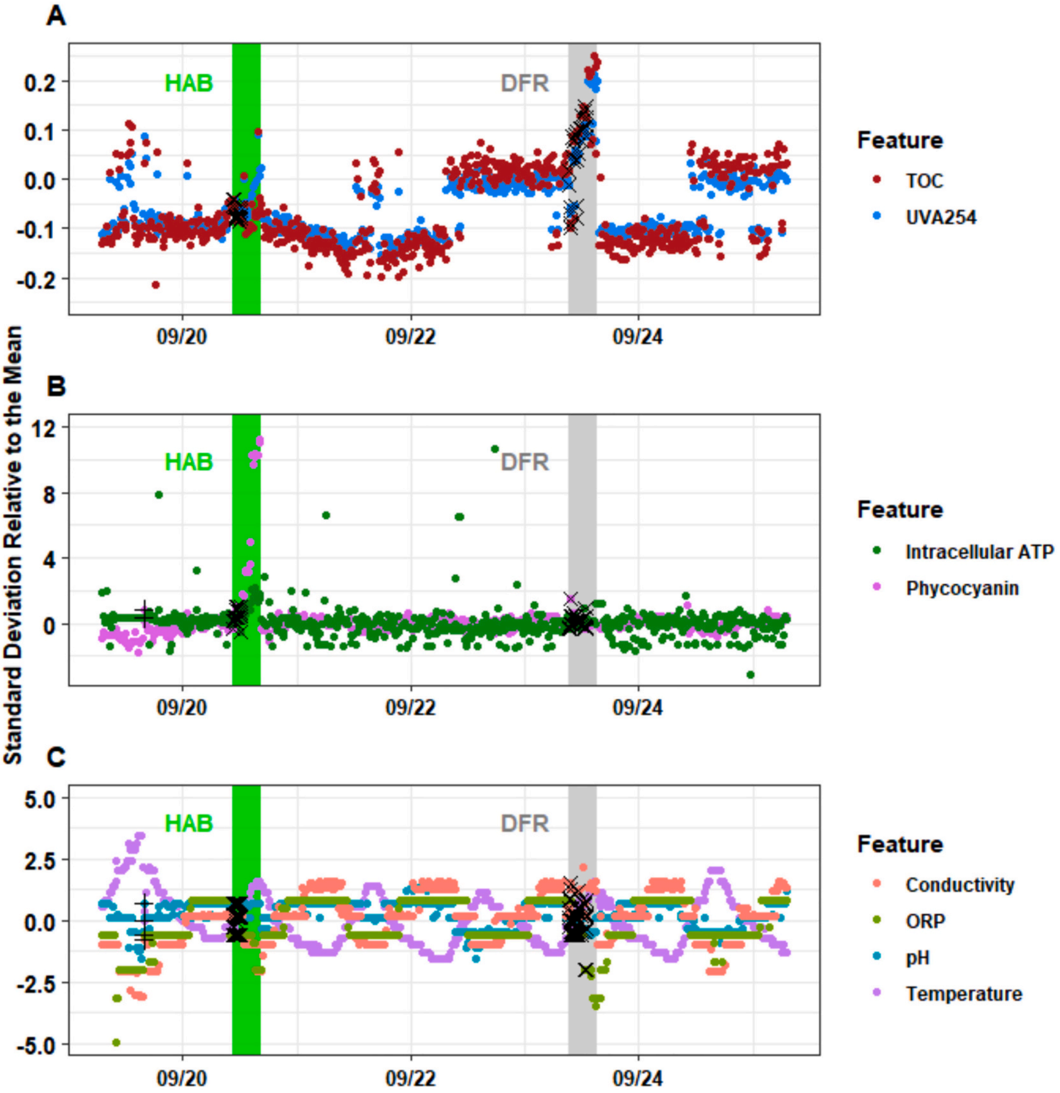


Fig. 4. Testing set results for the optimized version of *mda* for (A) UVA_{254} and TOC; (B) phycocyanin-like fluorescence (phycocyanin) and intracellular ATP; and (C) Conductivity, ORP, pH, and temperature. False positives are shown with + and false negatives are labeled with X. Shaded gray areas represent blending with wastewater effluent to simulate DPR and shaded green areas represent mixing events with higher concentrations of cyanobacteria. Some outliers are excluded from the range of the y-axis and the whole time period is not shown to highlight the impacts of blending events. Features less impacted by the blending events are also omitted for clarity. The full dataset is provided in Fig. S6.

omitted from different subsets. FDA is an extension of linear discriminant analysis that can be used for nonlinear relationships for classification which can be applied to the bootstrapped datasets (Mallet et al., 1996). GCV pruning can be applied to the FDA models to determine the optimal level of complexity and prevent overfitting. Bagged FDA using GCV Pruning (*bagFDAGCV*) had the second highest testing set accuracy using the raw data, 97.66 % (Table 2), which was significantly above the NIR (p-value = 0.00057). It had the highest Wastewater sensitivity (75 %) and the highest cyanobacteria sensitivity (87.5 %) and 18 false alerts. Smoothing substantially decreased the testing set accuracy to 52.00 %, and seasonal adjustment resulted in a testing set accuracy of 97.05 %. The raw data next underwent diurnal adjustment for pH, ORP, conductivity, and temperature, but these all also decreased the testing set accuracy compared to the raw data. Conductivity was the least important feature, but omitting it resulted in a testing set accuracy of 97.48 %. The default tuning, with the maximum *degree* of interaction for FDA being 1, was the most accurate. Therefore, *bagFDAGCV* underwent no preprocessing and was most accurate with the raw data, the default tuning, and all the features included (Table 3). Of the 18 false alerts, four were when the model predicted Wastewater when it was Normal, 13 were when the model predicted Algae when it was Normal, and one occurred when the model predicted Algae during a maintenance event labeled as Normal. Results for *bagFDAGCV* are shown in Fig. S8. If the data did undergo seasonal adjustment and was then optimized, the pH and conductivity were diurnally adjusted, and all the features were included. This resulted in a lower testing set accuracy of 97.31 %, but only had 6 false positives.

3.6. Extreme gradient boosting

Extreme gradient boosting (*xgbTree*) is an extension of the gradient boosting framework by Friedman et al. (2000) and Friedman (2001) using decision trees (Chen and He, 2023). Extreme gradient boosting uses regularization, parallel processing, and tree pruning, and has more flexibility than traditional gradient boosting. Extreme gradient boosting (*xgbTree*) had the second highest training set accuracy on the raw data, 99.71 %, but its testing set accuracy was below the NIR at 93.84 %, indicating overfitting (Table 2). The Wastewater sensitivity was 54.17 %, the cyanobacteria sensitivity was 12.50 %, and there were 39 false alerts. Smoothing improved the testing set accuracy to 95.31 % and seasonal adjustment improved the testing set accuracy to 96.44 %. However, first smoothing and then applying seasonal adjustment resulted in a lower testing set accuracy, 95.83 %. Diurnally adjusting pH after seasonal adjustment improved the testing set accuracy to 96.88 % but also diurnally adjusting temperature, ORP, or conductivity did not improve accuracy. The least important feature for *xgbTree* was turbidity and omitting turbidity increased the testing set accuracy to 96.96 % but omitting chlorophyll *a*, the next least important feature, reduced testing set accuracy to 96.79 %. The only false positive occurred when the model predicted Wastewater for a Normal event. The number of boosting iterations (*nrounds*), the maximum tree depth (*max_depth*), the shrinkage factor (*eta*), the minimum loss reduction (*gamma*), subsample ratio of columns (*colsample_bytree*), minimum sum of instance weight (*min_child_weight*), and the subsample percentage all underwent tuning. The best tune (*nrounds* = 100, *max_depth* = 1, *eta* = 0.15, *gamma* = 0, colsample_bytree = 0.6, *min_child_weight* = 1, *subsample* = 1) resulted in a testing set accuracy of 97.40 % with a *p*-value of 3.12×10^{-3} (Table 3). Results for the optimized *xgbTree* model are shown in Fig. S9.

4. Discussion

4.1. Detection sensitivity

While a previous SML study was able to detect 2, 5, and 10 % of DFR (Thompson and Dickenson, 2021), this study attempted lower blending ratios of 0.5. 1, and 2 % wastewater effluent and certain SML algorithms were still able to detect the lowest levels evaluated. The blending ratios for the algal stocks were also low (0.005 %, 0.015 %, and 0.045 % of the UTEX 2061 stock, corresponding to approximately 950 cells/mL, 2900 cells/mL, and 8600 cells/mL) and yet the lowest levels were also detected. For example, the *bagFDAGCV* model had a Wastewater sensitivity of 75 % and a cyanobacteria sensitivity of 87.5 %, demonstrating that it could detect the blending ratios down to 0.5 % and 0.005 %, for DFR and the UTEX 2061 stock, respectively (Table S7). *bagFDAGCV* was able to detect 0.5 % DFR at the fourth datapoint at 15-min intervals (after 45 min) and detected 0.005 % UTEX stock solution (~950 cells/mL) on the third datapoint (after 30 min). However, *mda* (as seen in Fig. 4) had a Wastewater sensitivity of only 50 %, though the cyanobacteria sensitivity was also 87.5 %.

By detecting such low blending ratios, the SML would be able to warn of early-onset events before they reach a hazardous level. A 0.5 % increase in DFR could be considered minor relative to the current estimated 1.4 % DFR at SNWA's intake in Lake Mead (Hannoun et al., 2021). Algal blooms have been defined as extremely high phytoplankton cell densities (typically above 20,000 to 100,000 cells/mL) (Graham et al., 2008), which is a similar range that drinking water utilities have used as alert levels for their source water monitoring (Kibuye et al., 2021). A previous study investigated the *Microcystis* cell counts in urban pond algal blooms in Northern Kentucky and found concentrations up to 86,000 cells/mL (de la Cruz et al., 2017). In fact, the 0.005 % blending of the UTEX 2061 stock (~950 cells/mL) was not detected with the ATP and OD730 benchtop methods in this study (i.e., approximately 0.9 mg/L intracellular ATP and 0.000017 1/cm OD730). The phycocyanin sensor had clear spikes in the UTEX 2061 stock blending events (Fig. S10), indicating the sensitivity of this sensor relative to benchtop methods and its usefulness for monitoring algal blooms. Similarly, there was a spike for UVA_{254} during the Wastewater events, indicating it would be a useful sensor to monitor for increased DFR.

4.2. False alerts

A false alert could lead to poor allocation of resources or a tendency to ignore the alarms, especially in this case considering the blending levels would be sub-hazardous. Therefore, false positives could be considered a more important error type than false negatives for the levels in this study, though with higher blending ratios the false negatives could be more consequential. Thus, it is important to note that two of the optimized models (*LogitBoost*, and *xgbTree*) had only 1 false alert, though *mda* had 7 and *bagFDAGCV* had 18. Considering the testing set covered 12 days of data, *LogitBoost* and *xgbTree* would meet a criterion of no more than one false alert per week, while *mda* and *bagFDAGCV* would not. Under this requirement, *LogitBoost* would be the best of the optimized models.

4.3. Unbalanced data performance metrics

This study also had many more datapoints classified as Normal than as Algae or Wastewater, leading to an unbalanced dataset, which can result in poorer predictive performance (Zeinolabedini Rezaabad et al., 2023). If a model only ever predicted Normal, it would have an accuracy of 95.83 % (the NIR). A better assessment technique could be looking at the balanced accuracy, which is the accuracy if there were equal amounts of data from each class. Comparing optimized models, *bagFDAGCV* had the highest testing set balanced accuracy, 86.96 %. *mda* had had the highest testing set accuracy but its balanced accuracy was lower at 78.95 %. Among models with fewer than one false alert per week, *LogitBoost* had a higher balanced accuracy of 64.46 % than *xgbTree* (59.69 %).

Cohen's Kappa is another way to assess classification models that is consistent across studies despite unbalanced data. Cohen's Kappa compares the agreement from correct classifications and from classifications that could be due to chance (Cohen, 1960). A value of 1 indicates there is perfect agreement, a value of 0 indicates the agreement is equal to what would be expected from chance. *Mda* had a kappa of 0.745, *LogitBoost* had a kappa of 0.620, *bagFDAGCV* had a kappa of 0.739, and *xgbTree* had a kappa of 0.551. The higher overall accuracy led to higher kappa values, with *mda* having the highest kappa and testing set accuracy. If *mda* and *bagFDAGCV* were excluded as having too many false alerts, then *LogitBoost* would be the best remaining model in terms of balanced accuracy and Cohen's Kappa.

4.4. Preprocessing methods and hyperparameter tuning

Smoothing the data by taking the median of three data points decreased accuracy in all cases. With this approach, the detection of an event was automatically delayed an additional 15 min and once the event ended, there would be an additional false positive after the event. So, despite smoothing removing any non-consecutive random outliers, it was not retained in any of the optimized models.

Seasonal adjusting by subtracting the 24-h rolling median only improved two out of the four models, *xgbTree* and *LogitBoost*. If the other two models (*mda* and *bagFDAGCV*) did have their features seasonally adjusted and were then optimized, they had fewer false positives (7 vs. 1 for *mda* and 18 vs. 6 for *bagFDAGCV*) without much loss in accuracy (98.09 % vs. 98.00 % for *mda* and 97.66 % vs. 97.31 % for *bagFDAGCV*). Therefore, if minimizing false positives were the priority, seasonal adjustment might be useful for all models. If there needed to be less than one false alert per week, seasonally adjusted *mda* would be the most accurate model (Cohen's kappa was 0.686). When *bagFDAGCV* and *mda* were not seasonally adjusted, most of the false alerts were the model predicting Algae when the actual condition was Normal. Without the seasonal adjustment, gradual upward sensor drift of the phycocyanin-like fluorescence may have caused these false alerts (Fig. S10).

Four features had clear diurnal trends: pH, temperature, conductivity, and ORP (Fig. 3C). Three of the models had increased accuracy with diurnal adjustments of certain features, making diurnal adjustment the most widely effective of the preprocessing methods evaluated. However, the beneficial diurnally adjusted features were different between models. For example, for *mda*, diurnal pH, conductivity and temperature adjustments resulted in the highest accuracy, while only pH was diurnally adjusted for *xgbTree*. These results indicate (1) that SML models are generally able to handle raw data for surface water quality, (2) the best preprocessing methods are model-specific.

Each of these four models' hyperparameters underwent tuning for increased accuracy but this only improved two of the four models, and the improvements in testing set accuracy were small. For *LogitBoost*, it was <0.006 % and increased the number for false positives from 0 to 1. For *xgbTree*, adjusting *eta* to 0.15 improved the accuracy from 96.96 % to 97.40 %. The low impact of hyperparameter tuning indicates the default options within *caret* were suitable for this dataset or the selected models were not highly sensitive to hyperparameter settings.

4.5. Feature importance

UVA_{254}, TOC, and turbidity had previously been found to be useful features for wastewater effluent detection with SML (Thompson and Dickenson, 2021). It was hypothesized that ATP and phycocyanin-like fluorescence would aid algal bloom even detection based on other literature (Choo et al., 2018; Greenstein and Wert, 2019). Even features that appeared relatively noisy or did not have clear visual peaks during the blending events were useful for assessing whether the overall water quality was typical. This was determined by ranking features by importance, eliminating the least important feature, and determining if the accuracy remained the same or improved. Among the optimized models, two (*mda* and *bagFDAGCV*) were most accurate when using all the features, and the other two had only one feature excluded. Nitrate was excluded in *LogitBoost* and turbidity was excluded in *xgbTree*. Turbidity did not differ greatly between the Lake Mead water (around 0.3 NTU) and the tertiary wastewater effluent (around 0.4 NTU) (Tables 1, S3). Nitrate as measured by the Ketos Shield spiked at times that did not correspond to the blending events (Figs. S4 and S7).

The most important features varied among the chosen models. For example, nitrate was one of the most important features for *bagFDAGCV* and *xgbTree*, but *LogitBoost* was more accurate without including it. Wang et al. (2024) used machine learning to predict the chlorophyll *a* concentrations in a lake recharged by recycled water using the difference in nutrients between the lake and recycled water. They found nitrate from the recycled water was the key factor for algal bloom control, especially when there were higher temperatures. Lake Mead is phosphorous, not nitrogen limited (Hannoun and Tietjen, 2023), so the most important features for the machine learning models will also depend on the water matrices. Conductivity was the most important feature for *LogitBoost* but not among the top five variables for the other optimized models. Nonetheless, phycocyanin-like fluorescence and UVA_{254} were among the top five important features for all four optimized models, helping identify Algae and Wastewater events, respectively. This outcome is reasonable considering their clear peaks during events (Figs. 3 and 4). TOC also had visible peaks during Wastewater events but may have been deemphasized by certain models due to its correlation with $UVA_{254.}$ Other studies have used machine learning to predict chlorophyll *a* in lakes as a proxy for algal blooms (Chen et al., 2024; Wang et al., 2024), though phycocyanin-like fluorescence might provide more information for identifying cyanobacteria proliferation.

4.6. Study limitations

The experiments performed for this study occurred within one month, which did not capture the full range of seasonal variations in temperature and nutrient concentrations that occur throughout the year in Lake Mead. However, seasonal adjustment was included as a preprocessing technique to adjust for these variations or possible instrument drift. While the limited timeframe did not cover all possible conditions, this experiment showed that SML was successful in detecting algae and DFR. Since water utilities would need to train their models on data collected at their specific intake, demonstrating that this technique was effective was the priority of this study, rather than the collection of data with all water quality conditions.

Classification SML, as used in this study, would be useful in creating alarms when there were higher levels of DFR or algae. However, regression SML could be used to determine the percentages of DFR or concentrations of algae, which would be valuable to decide whether the increases were large enough to necessitate any additional treatment. Future work could include collecting more data to build regression models for DFR and algal concentrations. Additional data collection, capturing seasonal variations, could allow for more robust and reliable models. While the current models had accuracies up to 98.09 %, more training and testing data would validate the models' effectiveness under all environmental conditions.

5. Conclusions

Overall, this study showed SML could be used to detect and differentiate very low levels of algae and DFR in the intake water for a drinking water treatment plant. The highest accuracies were only possible providing at least 15 features (including 13 water quality variables) as model inputs. Among the optimized models, *mda* had the greatest testing set accuracy, 98.09 %. If having less than one false alert per week is a requirement, *mda* would still be the most accurate (98.00 %), but with different preprocessing methods. Diurnal adjustment was generally the most effective data preprocessing method tested, particularly for pH. Nonetheless, the models differed on which features were beneficial for preprocessing. One model, *bagFDAGCV*, was most accurate on the raw data and therefore did not undergo preprocessing for its optimized version. Some of the most important features for the different models were phycocyanin-like fluorescence and UVA_{254}. Nitrate and turbidity were the only features omitted from any optimized models. Phycocyanin-like fluorescence helped identify Algae events, while UVA_{254} helped identify Wastewater events.

Future studies can improve on this research by collecting a larger sample size, particularly of the Wastewater and Algae events, to further refine and assess the classification. A longer study period could allow for more seasonal variations to be captured, which could be used to refine the model further. While this study used classification models, future work could use SML regression to estimate the percentage of DFR or concentration of algae. This would indicate if further treatment would be necessary and, if so, how much. Regression models could also be used to determine if the percentage of DFR or algae concentrations were increasing over long time periods. While this data and the models built from it are specific to the water quality of Lake Mead, similar techniques could be applied at other intakes, prioritizing having measurements for the most important features, though they would need to train the models with their own water quality.

CRediT authorship contribution statement

Emily Clements: Writing – original draft, Visualization, Software, Formal analysis. **Kyle A. Thompson:** Writing – original draft, Visualization, Software, Methodology, Data curation. **Deena Hannoun:** Writing – review & editing, Supervision, Conceptualization. **Eric R.V. Dickenson:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This project was funded in part by the US Bureau of Reclamation, Funding Opportunity No. R22AP00236. The authors would like to thank the instrument vendors for their customer service. We would also like to acknowledge Ron Terifaj from SNWA Facilities Maintenance for instrument installation and Rich Velasco from SNWA Maintenance Engineering for SCADA support. We would like to thank Bright Huang and Natalie Piñon for instrument maintenance and Dr. Mahmut Ersan and Dr. Faith Kibuye for assistance in cultivating UTEX 2061. We would also like to thank Dr. Todd Tietjen, Dr. Ariel Atkinson, Dr. Eric Wert, and Daniel Chan for reviewing the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.174690.

References

Arad, J., Housh, M., Perelman, L., Ostfeld, A., 2013. A dynamic thresholds scheme for contaminant event detection in water distribution systems. Water Res. 47, 1899–1908. https://doi.org/10.1016/j.watres.2013.01.017.

Asheri-Arnon, T., Ezra, S., Fishbain, B., 2018. Contamination detection of water with varying routine backgrounds by UV-spectrophotometry. J. Water Resour. Plan. Manag. 144, 04018056 https://doi.org/10.1061/(ASCE)WR.1943-5452.0000965.

Atinkpahoun, C.N.H., Le, N.D., Pontvianne, S., Poirot, H., Leclerc, J.-P., Pons, M.-N., Soclo, H.H., 2018. Population mobility and urban wastewater dynamics. Sci. Total Environ. 622–623, 1431–1437. https://doi.org/10.1016/j.scitotenv.2017.12.087.

Beaver, J.R., Kirsch, J.E., Tausz, C.E., Samples, E.E., Renicker, T.R., Scotese, K.C., McMaster, H.A., Blasius-Wert, B.J., Zimba, P.V., Casamatta, D.A., 2018. Long-term trends in seasonal plankton dynamics in Lake Mead (Nevada-Arizona, USA) and implications for climate change. Hydrobiologia 822, 85–109. https://doi.org/10.1007/s10750-018-3638-4.

Blunt, S.M., Sackett, J.D., Rosen, M.R., Benotti, M.J., Trenholm, R.A., Vanderford, B.J., Hedlund, B.P., Moser, D.P., 2018. Association between degradation of pharmaceuticals and endocrine-disrupting compounds and microbial communities along a treated wastewater effluent gradient in Lake Mead. Sci. Total Environ. 622–623, 1640–1648. https://doi.org/10.1016/j.scitotenv.2017.10.052.

Burnet, J.-B., Sylvestre, É., Jalbert, J., Imbeault, S., Servais, P., Prévost, M., Dorner, S., 2019. Tracking the contribution of multiple raw and treated wastewater discharges at an urban drinking water supply using near real-time monitoring of β-d-glucuronidase activity. Water Res. 164, 114869 https://doi.org/10.1016/j.watres.2019.114869.

Centers for Disease Control and Prevention, 2016. One Health Harmful Algal Bloom System (OHHABS) Algae, Algal Toxins, and Other Pathogens List.

Chang, D.-W., Hobson, P., Burch, M., Lin, T.-F., 2012. Measurement of cyanobacteria using *in-vivo* fluoroscopy – effect of cyanobacterial species, pigments, and colonies. Water Res. 46, 5037–5048. https://doi.org/10.1016/j.watres.2012.06.050.

Chen, T., He, T., 2023. xgboost: eXtreme Gradient Boosting. Package Version: 1.7.6.1.

Chen, C., Chen, Q., Yao, S., He, M., Zhang, J., Li, G., Lin, Y., 2024. Combining physical-based model and machine learning to forecast chlorophyll-a concentration in freshwater lakes. Sci. Total Environ. 907, 168097 https://doi.org/10.1016/j.scitotenv.2023.168097.

Choo, F., Zamyadi, A., Newton, K., Newcombe, G., Bowling, L., Stuetz, R., Henderson, R. K., 2018. Performance evaluation of in situ fluorometers for real-time cyanobacterial monitoring. H2Open Journal 1, 26–46. https://doi.org/10.2166/h2oj.2018.009.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104.

de la Cruz, A., Logsdon, R., Lye, D., Guglielmi, S., Rice, A., Kannan, M.S., 2017. Harmful algae bloom occurrence in urban ponds: relationship of toxin levels with cell density and species composition. J Earth Environ Sci 25, 704–726. https://doi.org/10.29011/JEES-148.100048.

Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J. 16, 235–248. https://doi.org/10.1080/1573062X.2019.1637002.

Drewes, J., Hübner, U., Zhiteneva, V., Karakurt, S., 2017. Characterization of Unplanned Water Reuse in the EU. Final Report. European Union, Garching, Germany.

Fleming, S.W., Titus, M., Watson, J.R., Doring, D., 2019. Technology Demonstration for One-week-ahead Forecasting of Toxic Algal Blooms in the US Army Corps of Engineers Reservoir at Detroit Lake Using Machine Learning 2019, GH44A-12.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407. https://doi.org/10.1214/aos/1016218223.

Graham, J., Loftin, K., Ziegler, A., Meyer, M., 2008. Guidelines for Design and Sampling for Cyanobacterial Toxin and Taste-and-odor Studies in Lakes and Reservoirs (Scientific Investigations Report). Scientific Investigations Report. U.S. Department of the Interior and U.S, Geological Survey.

Greenstein, K.E., Wert, E.C., 2019. Using rapid quantification of adenosine triphosphate (ATP) as an indicator for early detection and treatment of cyanobacterial blooms. Water Res. 154, 171–179. https://doi.org/10.1016/j.watres.2019.02.005.

Hannoun, D., Tietjen, T., 2023. Lake management under severe drought: Lake Mead, Nevada/Arizona. JAWRA Journal of the American Water Resources Association 59, 416–428. https://doi.org/10.1111/1752-1688.13090.

Hannoun, D., Tietjen, T., Brooks, K., 2021. The potential effects of climate change and drawdown on a newly constructed drinking water intake: study case in Las Vegas, NV, USA. Water Utility Journal 1–13.

Hastie, T., Tibshirani, R., 1996. Discriminant analysis by Gaussian mixtures. J. R. Stat. Soc. B. Methodol. 58, 155–176. https://doi.org/10.1111/j.2517-6161.1996.tb02073.x.

Holdren, G.C., Turner, K., 2010. Characteristics of Lake Mead, Arizona–Nevada. Lake and Reservoir Management 26, 230–239. https://doi.org/10.1080/07438141.2010.540699.

Islam, M., Thompson, K., Dickenson, E., Quiñones, O., Steinle-Darling, E., Westerhoff, P., 2023. Sucralose and predicted de facto wastewater reuse levels correlate with PFAS levels in surface waters. Environ. Sci. Technol. Lett. 10, 431–438. https://doi.org/10.1021/acs.estlett.3c00185.

Jeong, B., Chapeta, M.R., Kim, M., Kim, J., Shin, J., Cha, Y., 2022. Machine learning-based prediction of harmful algal blooms in water supply reservoirs. Water Quality Research Journal 57, 304–318. https://doi.org/10.2166/wqrj.2022.019.

Khullar, S., Singh, N., 2020. Machine learning techniques in river water quality modelling: a research travelogue. Water Supply 21, 1–13. https://doi.org/10.2166/ws.2020.277.

Kibuye, F.A., Almuhtaram, H., Zamyadi, A., Gaget, V., Owen, C., Hofmann, R., Wert, E. C., 2021. Utility practices and perspectives on monitoring and source control of cyanobacterial blooms. AWWA Water Science 3, e1264. https://doi.org/10.1002/aws2.1264.

Kim, J.H., Shin, J.-K., Lee, H., Lee, D.H., Kang, J.-H., Cho, K.H., Lee, Y.-G., Chon, K., Baek, S.-S., Park, Y., 2021. Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. Water Res. 207, 117821 https://doi.org/10.1016/j.watres.2021.117821.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05.

Kuhn, M., 2020. Classification and Regression Training [R Package Caret Version 6.0-86].

Liu, J., Wang, P., Jiang, D., Nan, J., Zhu, W., 2020. An integrated data-driven framework for surface water quality anomaly detection and early warning. J. Clean. Prod. 251, 119145 https://doi.org/10.1016/j.jclepro.2019.119145.

Liu, G., Savic, D., Fu, G., 2023. Short-term water demand forecasting using data-centric machine learning approaches. J. Hydroinf. 25, 895–911. https://doi.org/10.2166/hydro.2023.163.

Liu, W., Bao, Y., Li, K., Yang, N., He, P., He, C., Liu, J., 2024. The diversity of planktonic bacteria driven by environmental factors in different mariculture areas in the East China Sea. Mar. Pollut. Bull. 201, 116136 https://doi.org/10.1016/j.marpolbul.2024.116136.

Mallet, Y., Coomans, D., De Vel, O., 1996. Recent developments in discriminant analysis on high dimensional spectral data. Chemom. Intell. Lab. Syst. 35, 157–173. https://doi.org/10.1016/S0169-7439(96)00050-0.

Milly, P.C.D., Dunne, K.A., 2020. Colorado River flow dwindles as warming-driven loss of reflective snow energizes evaporation. Science 367, 1252–1255. https://doi.org/10.1126/science.aay9187.

Myers, J.A., Curtis, B.S., Curtis, W.R., 2013. Improving accuracy of cell and chromophore concentration measurements using optical density. BMC Biophys. 6, 4. https://doi.org/10.1186/2046-1682-6-4.

Ouahid, Y., Pérez-Silva, G., del Campo, F.F., 2005. Identification of potentially toxic environmental Microcystis by individual and multiple PCR amplification of specific microcystin synthetase gene regions. Environ. Toxicol. 20, 235–242. https://doi.org/10.1002/tox.20103.

Paerl, H.W., Huisman, J., 2009. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. Environ. Microbiol. Rep. 1, 27–37. https://doi.org/10.1111/j.1758-2229.2008.00004.x.

Pan, C., Zhang, L., Meng, X., Qin, H., Xiang, Z., Gong, W., Luo, W., Li, D., Han, X., 2021. Chronic exposure to microcystin-LR increases the risk of prostate cancer and induces malignant transformation of human prostate epithelial cells. Chemosphere 263, 128295. https://doi.org/10.1016/j.chemosphere.2020.128295.

Pfaff, J.D., 1993. Method 300.0, Rev. 2.1: Determination of Inorganic Anions by Ion Chromatography. US Environmental Protection Agency, Washington, DC, USA.

Potter, B.B., Wimsatt, J.C., 2009. Method 415.3, Rev. 1.2: Determination of Total Organic Carbon and Specific UV Absorbance at 254 nm in Source Water and Drinking Water. US Environmental Protection Agency, Washington, DC, USA.

R Core Team, 2023. R: A Language and Environment for Statistical Computing.

Rice, J., Westerhoff, P., 2015. Spatial and temporal variation in de facto wastewater reuse in drinking water systems across the U.S.A. Environ. Sci. Technol. 49, 982–989. https://doi.org/10.1021/es5048057.

Rice, J., Via, S.H., Westerhoff, P., 2015. Extent and impacts of unplanned wastewater reuse in US rivers. J. AWWA 107, E571–E581. https://doi.org/10.5942/jawwa.2015.107.0178.

Sarrafzadeh, M.H., La, H.-J., Seo, S.-H., Asgharnejad, H., Oh, H.-M., 2015. Evaluation of various techniques for microalgal biomass quantification. J. Biotechnol. 216, 90–97. https://doi.org/10.1016/j.jbiotec.2015.10.010.

Schaeffer, B.A., Urquhart, E., Coffer, M., Salls, W., Stumpf, R.P., Loftin, K.A., Jeremy Werdell, P., 2022. Satellites quantify the spatial extent of cyanobacterial blooms across the United States at multiple scales. Ecol. Indic. 140, 108990 https://doi.org/10.1016/j.ecolind.2022.108990.

Shen, X., Jiang, R., Liu, J., Zhao, D., Wang, L., Liu, Y., Yin, Y., Zhang, J., Shao, L., He, W., He, P., 2023. Algae extermination by a novel algicide (DMPAI) with low-dose and field experiment. Algal Res. 75, 103264 https://doi.org/10.1016/j.algal.2023.103264.

Sherman, L.A., Meunier, P., Colón-López, M.S., 1998. Diurnal rhythms in metabolism: a day in the life of a unicellular, diazotrophic cyanobacterium. Photosynth. Res. 58, 25–42. https://doi.org/10.1023/A:1006137605802.

Tchobanoglous, G., Burton, F.L., Stensel, H.D., 2002. Wastewater Engineering: Treatment and Reuse, 4th, edition. ed. McGraw-Hill Science/Engineering/Math, Boston.

Thompson, K.A., Dickenson, E.R.V., 2021. Using machine learning classification to detect simulated increases of de facto reuse and urban stormwater surges in surface water. Water Res. 204, 117556 https://doi.org/10.1016/j.watres.2021.117556.

U.S. Environmental Protection Agency, 2019. Recommended Human Health Recreational Ambient Water Quality Criteria or Swimming Advisories for Microcystins and Cylindrospermopsin (No. 822-R-19-00).

Wang, C., Liu, J., Qiu, C., Su, X., Ma, N., Li, J., Wang, S., Qu, S., 2024. Identifying the drivers of chlorophyll-a dynamics in a landscape lake recharged by reclaimed water using interpretable machine learning. Sci. Total Environ. 906, 167483 https://doi.org/10.1016/j.scitotenv.2023.167483.

Weisman, R.J., Barber, L.B., Rapp, J.L., Ferreira, C.M., 2019. De facto reuse and disinfection by-products in drinking water systems in the Shenandoah River watershed. Environ. Sci.: Water Res. Technol. 5, 1699–1708. https://doi.org/10.1039/C9EW00326F.

Weng, D., Lu, Y., Wei, Y., Liu, Y., Shen, P., 2007. The role of ROS in microcystin-LR-induced hepatocyte apoptosis and liver injury in mice. Toxicology 232, 15–23. https://doi.org/10.1016/j.tox.2006.12.010.

Zeinolabedini Rezaabad, M., Lacey, H., Marshall, L., Johnson, F., 2023. Influence of resampling techniques on Bayesian network performance in predicting increased algal activity. Water Res. 244, 120558 https://doi.org/10.1016/j.watres.2023.120558.