Contents lists available at ScienceDirect

Science of the Total Environment

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

Improving the robustness of beach water quality modeling using an ensemble machine learning approach

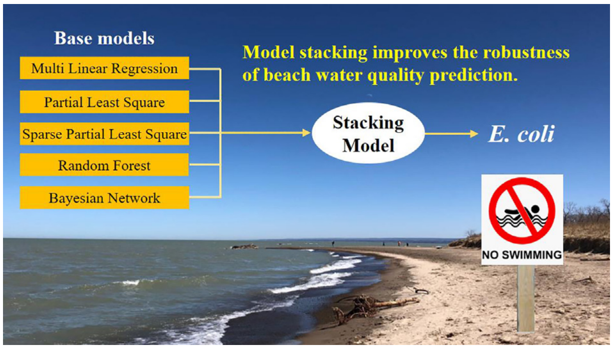
Leizhi Wang [a,b,c], Zhenduo Zhu [a,*], Lauren Sassoubre [a,1], Guan Yu [d], Chen Liao [e], Qingfang Hu [b,c], Yintang Wang [b,c]

[a] Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo 14220, NY, USA
[b] Nanjing Hydraulic Research Institute, State Key laboratory of Hydrology, Water Resources and Hydraulic Engineering & Science, Nanjing 210029, China
[c] Yangtze Institute for Conservation and Development, Nanjing, 210098, China
[d] Department of Biostatistics, University at Buffalo, The State University of New York, Buffalo 14220, NY, USA
[e] Program for Computational and Systems Biology, Memorial Sloan-Kettering Cancer Center, NY 10065, New York, USA

HIGHLIGHTS

- Predictive models often suffer from high variability in performance from one year or beach to another.
- A two-layered model stacking approach was proposed for beach water quality modeling.
- The model stacking approach improved the robustness of model prediction power.
- Random forest contributed the most by weight for the stacking model.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 July 2020
Received in revised form 28 September 2020
Accepted 28 September 2020
Available online 17 October 2020

Editor: Fernando A.L. Pacheco

Keywords:
Water quality
Fecal indicator bacteria
E. coli
Model stacking
Machine learning model

ABSTRACT

Microbial pollution of beach water can expose swimmers to harmful pathogens. Predictive modeling provides an alternative method for beach management that addresses several limitations associated with traditional culture-based methods of assessing water quality. Widely-used machine learning methods often suffer from high variability in performance from one year or beach to another. Therefore, the best machine learning method varies between beaches and years, making method selection difficult. This study proposes an ensemble machine learning approach referred to as model stacking that has a two-layered learning structure, where the outputs of five widely-used individual machine learning models (multiple linear regression, partial least square, sparse partial least square, random forest, and Bayesian network) are taken as input features for another model that produces the final prediction. Applying this approach to three beaches along eastern Lake Erie, New York, USA, we show that generally the model stacking approach was able to generate reliably good predictions compared to all of the five base models. The accuracy rankings of the stacking model consistently stayed 1^{st} or 2^{nd} every year, with yearly-average accuracy of 78%, 81%, and 82.3% at the three studied beaches, respectively. This study highlights the value of the model stacking approach in predicting beach water quality and solving other pressing environmental problems.

1. Introduction

Swimming in contaminated waters poses a health risk for beachgoers (Fewtrell and Kay, 2015). Fecal Indicator Bacteria (FIB),

[*] Corresponding author.
E-mail address: zhenduoz@buffalo.edu (Z. Zhu).
[1] Present address: University of San Francisco, San Francisco, 94117, CA, USA.

https://doi.org/10.1016/j.scitotenv.2020.142760
0048-9697/© 2020 Elsevier B.V. All rights reserved.

specifically *Escherichia coli* (*E. coli*) and *Enterococcus* spp., are widely used to assess beach microbial water quality in order to protect swimmers' health. The assessment is based on the premise that FIB, which are mostly harmless commensal bacteria, indicate the presence of pathogens. Current methods for assessing microbial water quality rely on culture-based enumeration of FIB. One of the major limitations of these culture-based methods is that results take 18–24 h, preventing beach managers from knowing the current water quality condition and issuing same-day swimming advisories for beachgoers. This limitation can be addressed with predictive models.

Predictive models trained on real-time surrogate data (e.g., rainfall, water temperature and turbidity) can serve as effective tools for predicting FIB concentrations and assessing risks at recreational beaches in real time. Researchers have evaluated the performances of a variety of statistical and machine learning models at different beaches with varying hydrodynamic, climatic and environmental conditions as model input variables (Zhang et al., 2012; de Brauwere et al., 2014; Palazón et al., 2017; Rossi et al., 2020).

Multiple linear regression (MLR), which minimizes the sum of the squares of residuals by either ordinary least square (OLS) or partial least square (PLS) methods, remains the most widely used statistical model (Olyphant and Whitman, 2004; de Brauwere et al., 2014; Herrig et al., 2015; Thoe et al., 2012). Compared to OLS which assumes little or no correlations between variables, PLS and sparse partial least square (SPLS) (Chun and Keles, 2010) are more robust against multicollinearity and have been shown to provide better predictions in some studies (Hou et al., 2006; Brooks et al., 2016; Brooks et al., 2013; Thoe et al., 2014a; Thoe et al., 2014b; Park et al., 2018). However, the spatiotemporal complexity of FIB concentrations and its inherent nonlinearity with surrogate variables hinder the predictive power of statistical models that require the linearity of regression coefficients (Panidhapu et al., 2020).

In contrast, many other machine learning methods, which do not assume any data structure (e.g., linearity), aim to build an empirical model that captures nonlinear, high-order dependences of response variables on explanatory variables in order to maximize predictability. So far, two popular algorithms, random forest (RF) (Parkhurst et al., 2005; Jones et al., 2013; Friedman et al., 2001) and Bayesian network (BN) (Avila et al., 2018), have shown potential for improving the accuracy of FIB predictions. For example, Brooks et al. (2016) studied seven beaches in Wisconsin, USA, finding that RF was the most accurate model (mean cross-validation error rate (MCVER): 24%) among 14 candidate models including MLR (MCVER: 27%). Avila et al. (2018) suggested that BN (MCVER: 21%) was superior to other models including RF (MCVER: 23%) for predicting *E. coli* levels at a recreational site in Wallacetown, New Zealand.

Previous research shows that the best model selected by cross validation varies spatially and temporally. For instance, RF was found to be the most accurate model for beaches in Wisconsin, USA (Brooks et al., 2016), but it did not perform the best in a study in New Zealand (Avila et al., 2018). Similarly, MLR was the second best model and much more accurate than PLS for Wisconsin beaches (Brooks et al., 2016), but it performed worse than PLS for other beaches along the Great Lakes in a different study (Brooks et al., 2013). In addition to variability in model performance between beaches, model performance also varies year to year at a given beach. A carefully selected model that performed satisfactorily at a given beach in previous years may not perform as well the next year. Year to year variability at a given beach has been largely overlooked in previous studies using mean cross-validation error as the metric, which may result in models with low prediction error on average but high variability across different test datasets from different years. Notably, such a prediction with low prediction error on average is often insufficient for beach management because it makes it difficult for managers to make decisions that are consistent, justifiable and maximize the trade-off between beach access and protecting human health.

The reasons that lead to spatiotemporal variability of predictions using statistical models can be complicated and has been investigated previously (Hellweger and Bucci, 2009; Racine et al., 2012). One possible reason is that the dependence of FIB concentration on environmental variables follows a complex spatiotemporal pattern that exhibits multiple coexisting characteristics that are difficult for any single model to capture. Each model has strengths and weaknesses, and one model may be superior in predicting the effects of one environmental variable (e.g., rainfall) but not the effects of other variables (e.g., temperature). When environmental conditions change annually, the predictability of a single model can be uncertain because it is unclear if the chosen single model appropriately predicts the effect of the environmental variable that most influences FIB concentrations during the given year.

The objective of this study is to address the variability in model performance with respect to different beaches and across consecutive years, by introducing a model stacking method to assemble the predictions of multiple individual models for the final prediction. Model stacking belongs to the family of ensemble modeling whose predictive power relies on the “wisdom of crowds”. Integration of different base models' results/outputs can reduce variance and increase the stability of the final model. Over the last decade, model stacking techniques have been broadly employed to various fields, such as marketing, managing, financial, and water resources problems (Alizamir et al., 2020; Rice and Emanuel, 2017; Zhai and Chen, 2018; Cham et al., 2020), but rarely applied to water quality assessment. The model stacking algorithm uses a two-layered learning framework where the outputs generated by individual base models are inputs in another model to generate final predictions. Specifically, five base models (MLR, PLS, SPLS, RF, and BN) were chosen due to their high popularity and performance based on previous studies. We formulated the second learning layer as a linear regression and estimated the linear weight coefficients for each base model by minimizing the least square error. The model stacking method was applied to three beaches along the eastern Lake Erie, New York, USA, and evaluated against the five base models.

2. Materials and methods

2.1. Data sources

Three beaches along eastern Lake Erie in the New York State, USA, are studied: Woodlawn, Hamburg, and Bennett beaches (Fig. 1). There are two main reasons: i) the relatively long historical records (from 2011 to 2017 for Woodlawn; from 2013 to 2017 for Hamburg and Bennett) of predictive environmental variables and *E. coli* concentrations, and ii) concerns about water quality during the swimming season. Woodlawn and Hamburg beaches are approximately 6 km away from each other, while Bennett Beach is around 26 km away from the other two. All three are sandy beaches and popular for recreational uses.

E. coli data and environmental data were provided by the Erie County Department of Health. Data are generally collected daily during the summer months between June and September when the beach is open to recreation. There is variation in sample size between years mainly due to missing data. In New York State, beaches are closed when *E. coli* concentration exceeds the recreational water quality criteria of 235 CFU/100 mL recommended by US EPA. Exceedance rates range from 10%–40% (2011–2017), 17%–29% (2013–2017), and 10%–31% (2013–2017) for Woodlawn, Hamburg, and Bennett, respectively (Fig. 2).

Modeling efforts are intended to predict daily *E. coli* concentrations and inform daily decisions about beach closures. At the end of the swimming season, the *E. coli* results and model predictions are compared to evaluate model performance. One beach in particular, Woodlawn Beach, often has high and variable *E. coli* concentrations and is challenging to model correctly. There are no clear point sources of pollution at these beaches. While sewage is often the source of *E. coli* at recreational

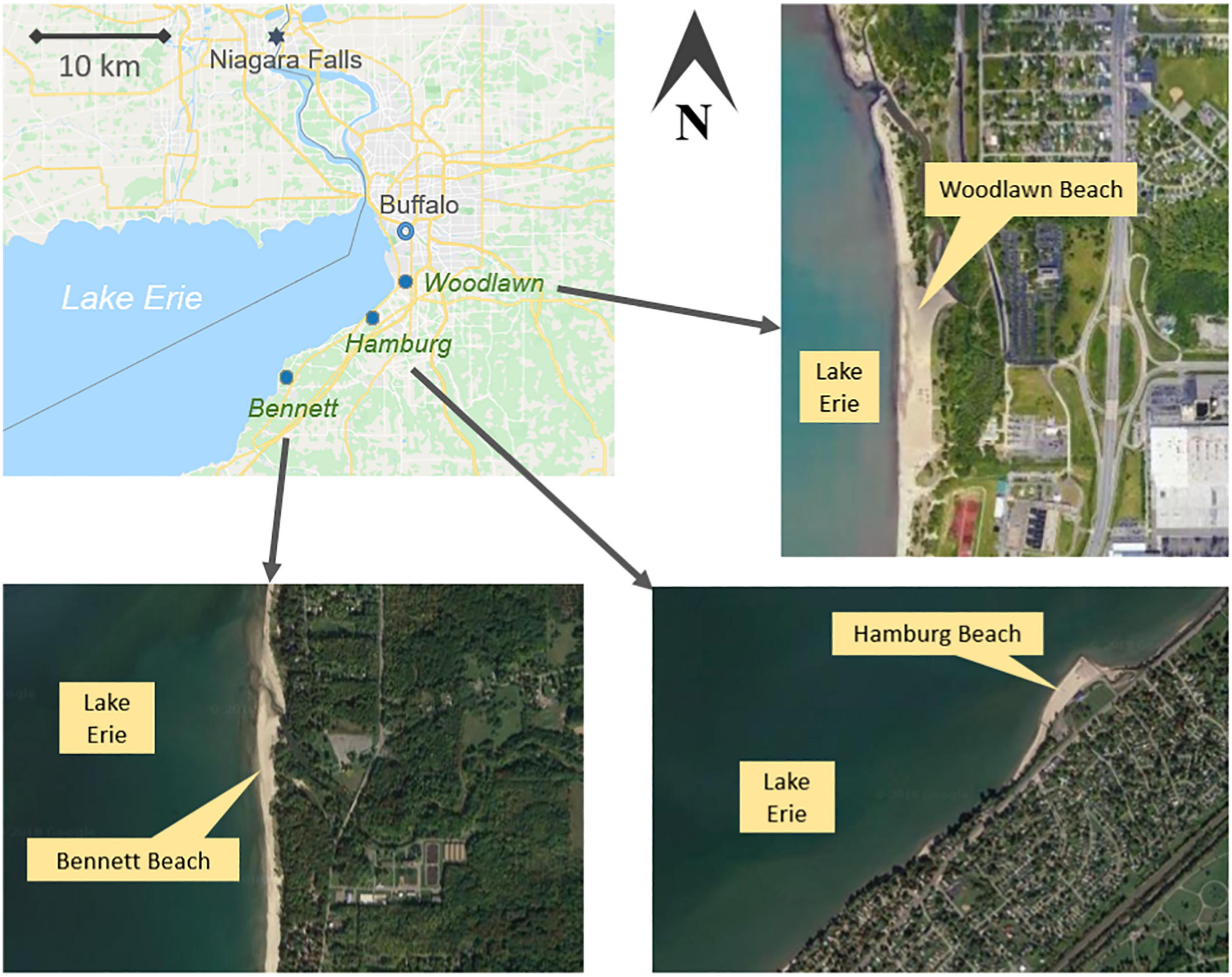


Fig. 1. Studied beaches along the eastern Lake Erie, New York, USA: Woodlawn Beach, Hamburg Beach and Bennett Beach. Picture sources: *Google Maps*.

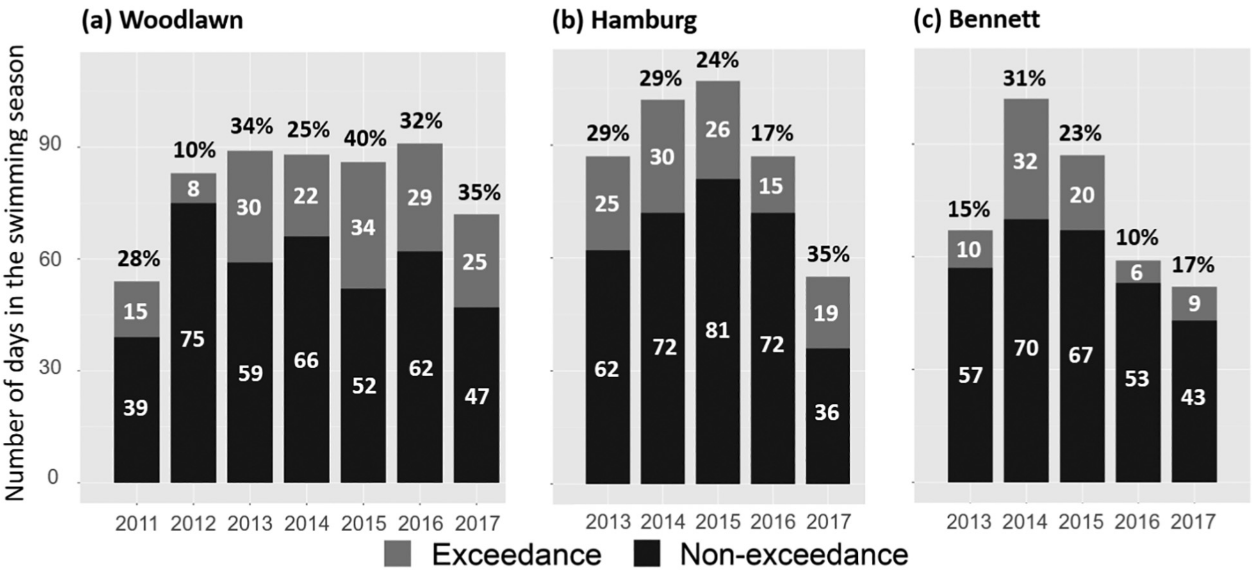


Fig. 2. Exceedances and non-exceedances of *E. coli* concentration data at the three beach sites. Numbers in grey bars refer to exceedances, while the ones in black bars are non-exceedances. Exceedance rate (number of exceedances/total) is shown on top of the bar.

beaches, previous studies have shown that *E. coli* can grow in beach sands, Cladophora and other algae in the Great Lakes (Francy et al., 2013). Therefore, we cannot assume that the pollution at these beaches is from sewage because it is likely that there are multiple sources (human and non-human) of *E. coli* at these beaches which complicates modeling efforts.

For developing predictive models, data of 43 environmental variables were collected for Woodlawn Beach, while 29 were collected for Hamburg and Bennett beaches, respectively. More details about the environmental variables are listed in Supplementary Materials (see Table S1). The wastewater treatment facility for the city of Buffalo is downstream of these beaches. It is located after Lake Erie flows into the fast-moving Niagara River so it could not influence the beaches analyzed in this study. The Southtowns Advanced Wastewater Treatment Facility that is closer to these beaches employs tertiary treatment and releases effluent far offshore that has lower concentrations of *E. coli* than Lake Erie itself so we do not believe this facility is a source of *E. coli*. The influence of tides is important in many coastal marine systems, but there is no tidal influence in Lake Erie. Lake Erie is affected by seiches, but these are much rarer and not on the same timescales as the variable *E. coli* concentrations observed at these beaches. While there is not a tidal influence, we do believe that variables such as offshore winds, lake height (influenced by seiches), and flow direction and speed could influence *E. coli* concentrations so they are included in the models. *E. coli* data were log10-transformed to reduce skewness and variance for analysis, and to meet statistical assumptions for normality (Brooks et al., 2016; Thoe et al., 2014a; Thoe et al., 2014b).

2.2. Model stacking

The model stacking method makes a composite prediction based on multiple base models' results/outputs. It is a process that uses a set of base models, each of them obtained by applying a certain learning process to the given problem (in this study predicting the *E. coli* concentrations). All base models are integrated to yield a final prediction. In the last decade, model stacking techniques have been widely applied in a broad range of fields (Zhai and Chen, 2018; Rice and Emanuel, 2017). Roli et al. (2001) categorize its learning process into three steps: stacking generation, stacking pruning, and stacking integration. The phase of stacking generation mainly refers to the generation of base models, and the last two steps generally mean optimally combining the base model predictions to form a final set of predictions using a second-level algorithm. How to combine the base models is very important for the success of the model stacking approach.

Different methods can be used to combine the base models, among which linear combination is the most widely used. Let y represents the stacking target (i.e. the final prediction of *E. coli* concentrations), and $f_1, f_2, \cdots, f_M$ denote the base model predictions from M individual algorithms ($M = 5$ in this study). A linear stacking model has a prediction function as Eq. (1).

$$y = w_1 f_1 + w_2 f_2 + \cdots + w_M f_M \tag{1}$$

where w_m ($m = 1, \cdots, M$) is the weight assigned for each base model. The key problem here lies in how to obtain the optimal set of weights. A "quadratic programming" (Frank and Wolfe, 1956) based algorithm was adopted to estimate the set of weights.

Assume the dataset based on which we intend to estimate the weights has N observations (see training process in Fig. 3). First, a base model m is trained using the dataset with the i^{th} observation removed. $\widehat{f}_m^{-i}(x_i)$ represents the prediction of the model m for the i^{th} observation (note this observation is not used during model training). The estimation of the weights is obtained from the least square linear regression of y_i (observed value of the i^{th} observation) on the linear combination of $\widehat{f}_m^{-i}(x_i)$, $m = 1, \cdots, M$. The optimal set of stacking weights are estimated by minimizing the following objective function under two constraints (Eq. (2)).

$$\begin{aligned} &\widehat{\omega}^{st} = argmin \sum_{i=1}^{N} \left[y_i - \sum_{m=1}^{M} \omega_m \widehat{f}_m^{-i}(x_i) \right]^2 \\ &\omega_m \geq 0 \qquad m = 1, 2, 3, \cdots, M \\ &\sum_{m=1}^{M} \omega_m = 1 \qquad m = 1, 2, 3, \cdots, M \end{aligned} \tag{2}$$

where $\widehat{\omega}^{st}$ is the objective function, x_i refers to the i^{th} observation composed of all environmental variables. There are two constraints: i) weight should be large or equal to zero, and ii) the sum of weights equals to one. The two constraints are reasonable if we interpret the weights as posterior model probabilities. It is noteworthy that the i^{th} observation is removed from the training data when training model m, in order to avoid assigning unfairly high weights to models with higher complexity (Friedman et al., 2001). Finally, it leads to a tractable quadratic programming problem (Roli et al., 2001). R package **quadprog** (Koenker and Mizera, 2014) was used to assign the weights.

2.3. Base models

There are numerous machine learning models that can be included into an ensemble model. Three linear regression models (MLR, PLS, and SPLS) and two nonlinear models (RF and BN) were chosen based on previous literature that compared a number of models. MLR is estimated based on OLS, which is often criticized due to the inflexibility of its linear structure and overfitting (Nevers and Whitman, 2005), but still widely accepted in water quality modeling given its ease of implementation and accessibility (Mas and Ahlfeld, 2007). PLS uses different linear combinations of the original exploratory variables to construct a certain number of orthogonal variables, and then regress the response variable on those orthogonal variables. This process has an advantage over principle component regression since it guarantees that each variable is related to the response. SPLS combines the idea of PLS with a variable selection routine based on the Least Angle Regression algorithm, which can drop some environmental variables entirely (Chun and Keles, 2010). It is noteworthy that RF already uses an ensemble learning approach named bagging and decision tree as the individual model (Parkhurst et al., 2005). BN is a graphical model that encompasses probabilistic relationships among a set of variables (Fenton and Neil, 2018). Since these models have been widely used and extensively discussed in literature, readers are referred to the previous studies for more details on each model (de Brauwere et al., 2014; Herrig et al., 2015; Brooks et al., 2013; Brooks et al., 2016; Avila et al., 2018). This paper is focused on the model stacking method described in the next section.

The predictive models using each individual model were developed via R (R-Core-Team, 2006). The base algorithm in R was used for MLR, the **pls** package (Mevik et al., 2011) was used for PLS, **spls** (Chung et al., 2012) for SPLS, **randomForest** (Liaw and Wiener, 2002) for RF, and **bnlearn** (Scutari and Denis, 2014) for BN.

2.4. Model evaluation

Cross validation (CV) is widely used to estimate the accuracy of a regression model with a completely new dataset (Kohavi, 1995). CV measures predictive performance by dividing a dataset into folds, then cycling through the folds to make predictions. In one cycle of CV, one fold is reserved as a test set and the remaining data is the training set. A model is first estimated using the training set and then evaluated by predicting the responses in the test set. Given the nature of beach management, a leave-one-year-out (LOYO) CV is typically adopted (Brooks et al., 2016), which means a model is trained using all data except for the year left out, then tested using data from the year that was left out (see data split in Fig. 3). For instance, *E. coli* data and environmental

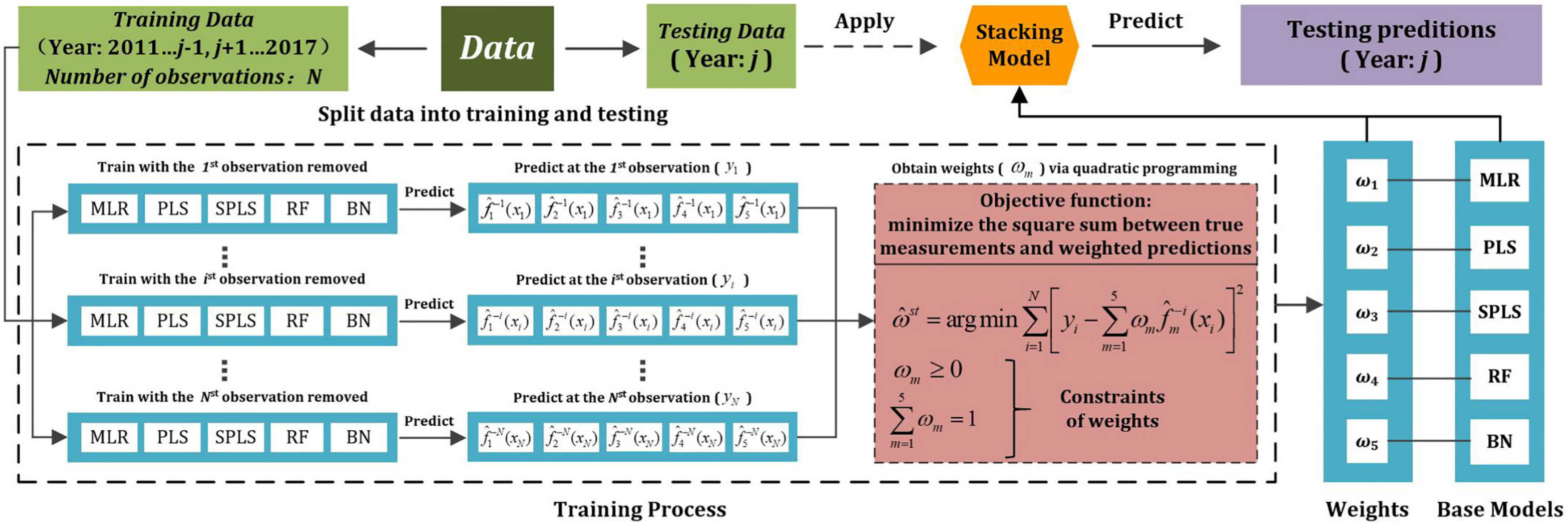

Training Data
(Year: 2011…j-1, j+1…2017)
Number of observations: N
Data
Testing Data
(Year: j)
Apply
Stacking Model
Predict
Testing preditions
(Year: j)
Split data into training and testing
Train with the 1st observation removed
MLR
PLS
SPLS
RF
BN
Predict
Predict at the 1st observation (y_1)
$\hat{f}_1^{-1}(x_1)$
$\hat{f}_2^{-1}(x_1)$
$\hat{f}_3^{-1}(x_1)$
$\hat{f}_4^{-1}(x_1)$
$\hat{f}_5^{-1}(x_1)$
Train with the ist observation removed
Predict at the ist observation (y_i)
$\hat{f}_1^{-i}(x_i)$
$\hat{f}_2^{-i}(x_i)$
$\hat{f}_3^{-i}(x_i)$
$\hat{f}_4^{-i}(x_i)$
$\hat{f}_5^{-i}(x_i)$
Train with the Nst observation removed
Predict at the Nst observation (y_N)
$\hat{f}_1^{-N}(x_N)$
$\hat{f}_2^{-N}(x_N)$
$\hat{f}_3^{-N}(x_N)$
$\hat{f}_4^{-N}(x_N)$
$\hat{f}_5^{-N}(x_N)$
Obtain weights (ω_m) via quadratic programming
Objective function:
minimize the square sum between true measurements and weighted predictions
$\hat{\omega}^{st} = \arg\min \sum_{i=1}^{N} \left[y_i - \sum_{m=1}^{5} \omega_m \hat{f}_m^{-i}(x_i) \right]^2$
$\omega_m \geq 0$
$\sum_{m=1}^{5} \omega_m = 1$
Constraints of weights
ω_1
ω_2
ω_3
ω_4
ω_5
Training Process
Weights
Base Models

Fig. 3. Details of the model stacking process when data of year *j* are reserved as the test set and the rest of data are the training set.

variables over seven years (2011–2017) were collected for Woodlawn Beach. A typical process of LOYO CV is to first set data from year 2011–2016 as the training set, and data from 2017 as the testing set. This process is executed seven times, at each time one year is taken out as the testing set while the remaining six years are used as the training set. One of the main reasons for using LOYO CV is to mimic the way that beach managers evaluate the predictive model after the beach season every year to decide whether it is necessary to develop a new model for the following year.

Generally, there are two types of metrics for evaluating model performance, goodness of fit and predictive power. The former refers to the extent to which observed data match the model predicted values, such as mean squared error (MSE). The latter includes sensitivity (the proportion of correctly predicted exceedances), specificity (the proportion of correctly predicted non-exceedances), and accuracy (the proportion of correct predictions), as an ultimate beach water quality prediction is to open or close the beach, depending on whether or not the predicted *E. coli* concentration exceeds the threshold (235 CFU/ 100 mL in this study) or not. (Brooks et al., 2013; Thoe et al., 2014b). We selected MSE and accuracy to evaluate the base models and the stacking model. Accuracy was adopted rather than sensitivity and specificity because managers for the beaches included in this study decide whether or not a new model is needed depending on whether or not the accuracy of the model used in the past year is smaller than 80%. In the following section, accuracy is mainly used for presentation and discussion of results as the same conclusions would be made with the results based on MSE. The MSE-based results are shown in the Supplementary Materials (see Table S2 and Fig. S1).

3. Results and discussion

3.1. Performance of base models

Performance of the five base models varied significantly at different beaches as well as for different years (see Table 1), demonstrating that there is not one best model for any of the beaches. For example, RF provided the highest accuracy in 2012, 2013 and 2016 at Woodlawn, while it only ranked 4th and 5th out of 5 in 2011 and 2014, respectively. At Hamburg, RF was superior to other models in 2013, 2014, 2017, while SPLS had the highest accuracy in the other years. Interestingly, BN showed poor performance in every year at Hamburg Beach. At Bennett, BN became the second best in 2013, 2015 and 2017, but turned out to be the worst in 2014. Similarly, RF performed well in 2016, but the worst in 2014. Meanwhile, model performance based on MSE was similar to accuracy, showing substantial year to year variability (see Table S2 in the Supplementary Materials). For instance, BN was superior to others in 2014 at Woodlawn, while it showed poor performance in 2011 and 2012. At Hamburg, SPLS provided the lowest MSE in 2016, but only ranked 3rd in 2013 and 2014.

Since no base model consistently performed the best, it is not practical to identify one best model for predicting beach closures. Essentially, data-driven models are devices used to capture relationships between the relevant input and output variables. Such models typically do not really represent the physics of modelled processes (Solomatine et al., 2009). Previous studies also found that no modeling technique was superior to all others, as all data-driven models have their strengths and weaknesses. With a wide range of modeling options to choose from, the challenge remains choosing a particular model that generates the best results for a given beach (Solomatine and Ostfeld, 2008).

Inter-annual variations in the probability density distribution of *E. coli* data might partly explain the unstable performance of these data-driven, individual models. One critical assumption made when building a statistical predictive model is that the unseen data (testing) comes from the same distribution as the training data (Solomatine and Ostfeld, 2008), so it is problematic for a single statistical model if the distribution of the training data is different from that of the testing. For example, Fig. 4 shows the distributions of log-transformed *E. coli* concentrations at Bennett Beach from 2013 to 2017. It shows that *E. coli* concentrations had a different pattern in 2014 compared to other years. Specifically, the frequency of log-transformed *E. coli* concentrations between 2 and 3 in 2014 was higher than that of other years, while *E. coli* concentrations smaller than 2 was lower. Meanwhile, BN was inferior to all the other models in 2014 while it performed well in other beach seasons. Moreover, it is worth pointing out that currently, beach managers assess their models after each beach season is over and decide to change or redevelop the models if the performance is unsatisfactory (e.g. accuracy <80%). This practice can be problematic because if the past year is unusual, development of new models might be unnecessary or even worse for the next year. Ideally the model stacking method combines the results of multiple base machine learning methods, so the variance of the prediction can be reduced.

3.2. Stacking model

3.2.1. Stacking weights

Stacking weights of the five base models were estimated by minimizing objective function (see Eq. (2)) via quadratic programming. With cross validation applied to each beach season, weights can differ

Table 1
Accuracy of different base models with leave-one-year-out cross validation. Note that numbers in the brackets represent the model's ranks in terms of accuracy (the stacking model is also included). The ranking sum (the last column) is derived simply by adding up ranks of all beach seasons.

Beach	Model	Beach season							Ranking sum
		2011	2012	2013	2014	2015	2016	2017	
Woodlawn	MLR	79.6% (4)	86.7% (5)	71.9% (3)	73.9% (3)	72.1% (4)	69.4% (5)	73.6% (3)	27
	PLS	81.5% (3)	88.0% (3)	71.9% (3)	79.5% (1)	73.3% (2)	70.8% (3)	73.6% (3)	18
	SPLS	83.3% (1)	88.0% (3)	71.9% (3)	73.9% (3)	70.9% (5)	69.4% (4)	73.6% (3)	22
	RF	79.6% (4)	90.4% (1)	74.2% (1)	72.7% (5)	72.1% (3)	73.6% (1)	75.0% (2)	17
	BN	72.2% (6)	86.7% (5)	70.8% (6)	72.7% (5)	74.4% (6)	75.0% (6)	72.2% (6)	40
	Stacking	82.4% (2)	89.2% (2)	74.2% (1)	76.1% (2)	74.4% (1)	73.6% (2)	76.4% (1)	11
Hamburg	MLR	–	–	74.8% (3)	82.4% (3)	80.4% (3)	86.2% (2)	70.9% (5)	16
	PLS	–	–	74.8% (3)	82.4% (3)	80.4% (3)	85.1% (5)	71.8% (2)	16
	SPLS	–	–	74.8% (3)	82.4% (3)	81.3% (2)	86.2% (2)	71.8% (2)	12
	RF	–	–	75.9% (2)	85.4% (1)	80.4% (3)	86.2% (2)	71.8% (2)	9
	BN	–	–	71.3% (6)	78.4% (6)	74.8% (6)	85.1% (5)	70.9% (5)	28
	Stacking	–	–	77.0% (1)	84.3% (2)	81.3% (1)	89.7% (1)	72.7% (1)	6
Bennett	MLR	–	–	73.1% (5)	75.5% (3)	75.9% (5)	89.8% (1)	80.8% (4)	18
	PLS	–	–	71.6% (6)	75.5% (3)	75.9% (5)	89.8% (1)	76.9% (6)	21
	SPLS	–	–	80.6% (4)	77.5% (1)	78.2% (2)	89.8% (1)	80.8% (4)	12
	RF	–	–	83.6% (2)	74.5% (5)	77.0% (3)	89.8% (1)	82.7% (3)	14
	BN	–	–	88.1% (1)	71.6% (6)	80.5% (1)	89.8% (1)	84.6% (1)	11
	Stacking	–	–	83.6% (2)	76.5% (2)	77.0% (3)	89.8% (1)	84.6% (1)	9

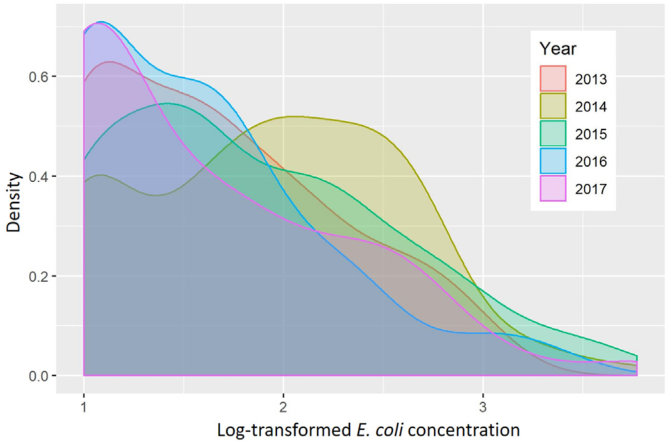


Fig. 4. Distributions of log-transformed *E. coli* concentration at Bennett from 2013 to 2017.

over the years as they are trained using different datasets. RF (black bars in Fig. 5) was always picked for the stacking model and contributed the most by weight (weights larger than 0.6 for all years). Contributions from other base models varied. Generally, MLR served as the second most important contributor to the stacking model, it was picked 5, 4, 3 times out of 7, 5, 5 years for each location respectively. SPLS became the third most frequently chosen model, contributing to the stacking model three times for each location. PLS was only used in the stacking model once at Bennett, and BN was the only base model that was never selected for the stacking model. Notably, even though there was no limitation on the number of base models included in the stacking model, the stacking method usually picked two base models (12 times out of 17).

3.2.2. Performance of stacking model

To compare models, radar plots showing the ranks of models according to their performance are presented (see Fig. 6). As discussed above, RF provided an unstable performance over the beach seasons at Woodlawn, especially for 2014 when it performed the worst out of all 5 base models. In contrast, the stacking model consistently provided an accurate prediction, as its rankings stayed 1st or 2nd every year at Woodlawn Beach (see Fig. 6a). The ranking sum (see Table 1) of the stacking model was 11, much smaller than that of RF (Hou et al., 2006) and other base models.

Similarly, at Hamburg (Fig. 6b) the stacking model outperformed all base models for almost every beach season, except for 2014 when it was second to RF. At Bennett (Fig. 6c), the stacking model ranked 1st or 2nd in 4 out of 5 years. In 2014, it ranked 2nd while the best base model, RF, fell to the 5th. These results show how the stacking model continues to perform well over time. Overall, the stacking model showed promise with its robustness and stability in performance, thus it can be a powerful tool for water quality predictions. Assessment using MSE values support these conclusions. The MSE values and the corresponding radar plots can be found in the Supplementary Materials (Table S2 and Fig. S1).

It is worth noting that the five base models were selected due to their high performances at other sites (e.g., beaches along Lake Michigan in Wisconsin). The stacking model should inherit the strengths of its members and likely is more accurate than its members when the base models are diverse (Dietterich, 2000). Therefore, for applying this

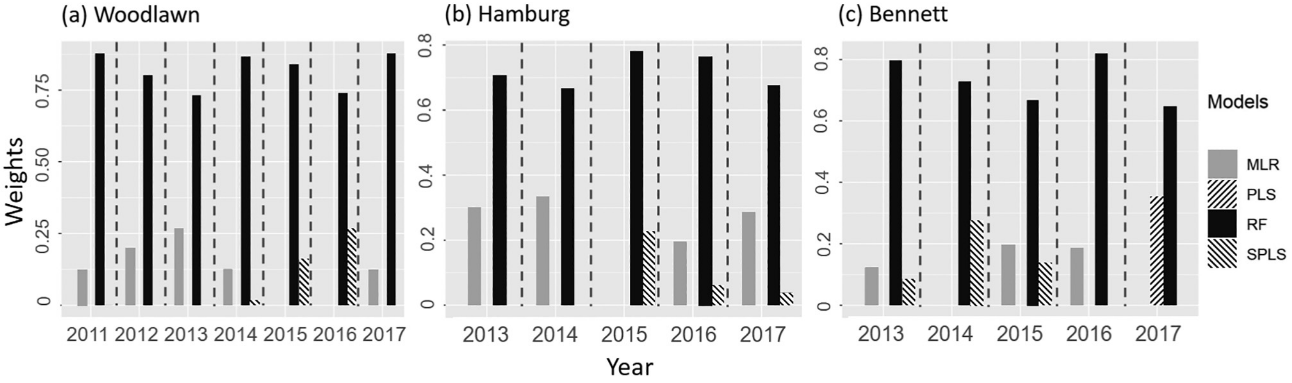


Fig. 5. Stacking weights of each base model at Woodlawn, Hamburg and Bennett, trained and constrained via quadratic programming by Eq. (2). Note that the grey dash lines help distinguish which year each constituent is included in.

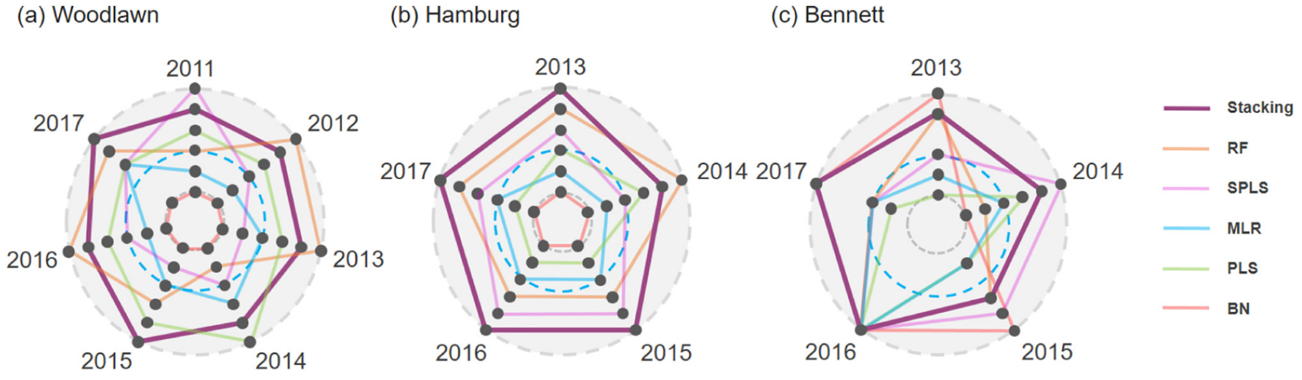


Fig. 6. Radar plots of model performance ranks based on accuracy at three beaches. Models achieving the best performance (highest accuracy) appear on the periphery of radar plots and vice versa. Note that RF and BN are visually missing in 2016 and 2017 in the accuracy radar plot at Bennett, because the accuracy can be identical to each other in certain years. The lines can be covered by others when they are drawn in sequence.

approach in other beach sites, other complementary models of different kinds, if shown to be predictive for the sites, could also be included in the base model collection. However, it does not necessarily mean that any base models with fine predictive power should be chosen, because one of the key premises of using model staking lies in the different structures of base models. Errors generated by these models may not be overlapped, so the stacking model can take advantage of different base models to make a better prediction.

4. Conclusions

Literature reviews on predictive models for FIB concentrations in beach water reveal that while a single model can be accurate for some years it can have poor predictive power in others. We found that this was a general issue for all statistical and machine learning models tested in previous studies as well as ours. Our study demonstrates the utility of using a model stacking approach for predictive modeling of beach water quality. Since model stacking averages out noise from its base models, it is theoretically more promising than individual models in generating predictions with greater accuracy and robustness. Minimizing overfitting is a special concern for stacking a library of models, thus techniques such as cross validation, regularization and bagging should be used to assess the generalization performance of a stacking model.

We applied this model stacking method to three beaches along the eastern Lake Erie, New York, USA, and showed that the model stacking approach outperformed all five base models in generating robust, cross-validated predictions. The accuracy rankings of the stacking model consistently stayed 1^{st} or 2^{nd} every year, with yearly-average accuracy of 78%, 81%, and 82.3% at the three studied beaches, respectively. The results from this study suggest that the model stacking algorithm has promise for improving the reliability of predictive modeling for beach microbial water quality of other sites with similar hydrogeological and environmental conditions such as other beaches along the Great Lakes. Although model stacking has been applied to a few environmental problems, its full potential as a tool for recreational water quality monitoring has not been fully tested. A comprehensive test needs to be done for more beach sites, especially those with different environmental variables from Lake Erie such as saltwater sites, to understand the strength and weakness of individual base models and the stacking approach. This study indicated that model stacking approach may improve the robustness of beach water quality modeling.

CRediT authorship contribution statement

Leizhi Wang: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Zhenduo Zhu:** Conceptualization, Methodology, Resources, Data curation, Writing - original draft. **Lauren Sassoubre:** Resources, Writing - review & editing. **Guan Yu:** Methodology, Writing - review & editing. **Chen Liao:** Writing - review & editing. **Qingfang Hu:** Writing - review & editing. **Yintang Wang:** 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.

Acknowledgment

The research was partly funded through a grant from the Great Lakes Research Consortium. The authors thank Rebecca R. Wightman and Erie County Department of Health for the support of research data and Berwin A. Turlach for the development of R package **quadprog**. L. Wang thanks financial support from China Scholarship Council. G. Yu thanks financial support from the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001412. The content is solely the responsibility of the authors and does not necessarily represent the official views of any agency mentioned above.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.142760.

References

Alizamir, M., Kisi, O., Muhammad Adnan, R., Kuriqi, A., 2020. Modelling reference evapotranspiration by combining neuro-fuzzy and evolutionary strategies. Act. Geoph. 68, 1–14. https://doi.org/10.1007/s11600-020-00446-9.

Avila, R., Horn, B., Moriarty, E., Hodson, R., Moltchanova, E., 2018. Evaluating statistical model performance in water quality prediction. J. Environ. Manag. 206, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049.

Brooks, W.R., Fienen, M.N., Corsi, S.R., 2013. Partial least squares for efficient models of fecal indicator bacteria on great lakes beaches. J. Environ. Manag. 114, 470–475. https://doi.org/10.1016/j.jenvman.2012.09.033.

Brooks, W., Corsi, S., Fienen, M., Carvin, R., 2016. Predicting recreational water quality advisories: a comparison of statistical methods. Environ. Mode. Softw. 76, 81–94. https://doi.org/10.1016/j.envsoft.2015.10.012.

Cham, D.D., Son, N.T., Minh, N.Q., Thanh, N.T., Dung, T.T., 2020. An analysis of shoreline changes using combined multitemporal remote sensing and digital evaluation model. Civ. Engi. J. 6, 1–10. https://doi.org/10.3390/s19102401.

Chun, H., Keles, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. Roy. Stati. Soci. 72, 3–25. https://doi.org/10.1111/j.1467-9868.2009.00723.x.

Chung, D., Chun, H., Keles, S., 2012. Spls: Sparse partial least squares (spls) regression and classification. R Package, Version 2.

de Brauwere, A., Ouattara, N.K., Servais, P., 2014. Modeling fecal indicator bacteria concentrations in natural surface waters: a review. Crit. Rev. Environ. Sci. Techno. 44, 2380–2453. https://doi.org/10.1080/10643389.2013.829978.

Dietterich, T.G., 2000. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems. Springer, pp. 1–15.

Fenton, N., Neil, M., 2018. Risk Assessment and Decision Analysis with Bayesian Networks. Crc Press https://doi.org/10.1201/b21982.

Fewtrell, L., Kay, D., 2015. Recreational water and infection: a review of recent findings. Curr. Environ. Health Rep. 2, 85–94. https://doi.org/10.1007/s40572-014-0036-6.

Francy, D.S., Stelzer, E.A., Duris, J.W., Brady, A.M.G., Harrison, J.H., Johnson, H.E., Ware, M.W., 2013. Predictive models for Escherichia coli concentrations at inland lake beaches and relationship of model variables to pathogen detection. Appl. Environ. Microbiol. 79, 1676–1688. https://doi.org/10.1128/AEM.02995-12.

Frank, M., Wolfe, P., 1956. An algorithm for quadratic programming. Naval Research Logistics Quarterly 3, 95–110. https://doi.org/10.1002/nav.3800030109.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. vol. 1. Springer series in statistics, New York, NY, USA. https://doi.org/10.1007/978-0-387-84858-7.

Hellweger, F.L., Bucci, V., 2009. A bunch of tiny individuals—individual-based modeling for microbes. Ecol. Model. 220, 8–22. https://doi.org/10.1016/j.ecolmodel.2008.09.004.

Herrig, I.M., Böer, S.I., Brennholt, N., Manz, W., 2015. Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower lahn river, Germany. Water Res. 85, 148–157. https://doi.org/10.1016/j.watres.2015.08.006.

Hou, D., Rabinovici, S.J.M., Boehm, A.B., 2006. Enterococci predictions from partial least squares regression models in conjunction with a single-sample standard improve the efficacy of beach management advisories. Environmental Science & Technology 40, 1737–1743. https://doi.org/10.1021/es0515250.

Jones, R.M., Liu, L., Dorevitch, S., 2013. Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. Environ. Monit. Assess. 185, 2355–2366. https://doi.org/10.1007/s10661-012-2716-8.

Koenker, R., Mizera, I., 2014. Convex optimization in R. J. Statis. Softw. 60, 1–23 doi: 10.1.1.704.3184.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Ijcai, Montreal, Canada, pp. 1137–1145.

Liaw, A., Wiener, M., 2002. Classification and Regression by Randomforest. R News 2. , pp. 18–22. https://CRAN.R-project.org/doc/Rnews/.

Mas, D.M.L., Ahlfeld, D.P., 2007. Comparing artificial neural networks and regression models for predicting faecal coliform concentrations. Hydrol. Sci. J. 52, 713–731. https://doi.org/10.1623/hysj.52.4.713.

Mevik, B.H., Wehrens, R., Liland, K.H., 2011. Pls: Partial Least Squares and Principal Component Regression. R Package Version 2.

Nevers, M.B., Whitman, R.L., 2005. Nowcast modeling of escherichia coli concentrations at multiple urban beaches of southern Lake Michigan. Water Res. 39, 5250–5260. https://doi.org/10.1016/j.watres.2005.10.012.

Olyphant, G.A., Whitman, R.L., 2004. Elements of a predictive model for determining beach closures on a real time basis: the case of 63rd street beach Chicago. Environ. Monit. Assess. 98, 175–190. https://doi.org/10.1023/b:emas.0000038185.79137.b9.

Palazón, A., López, I., Aragonés, L., Villacampa, Y., Navarro-González, F., 2017. Modelling of Escherichia coli concentrations in bathing water at microtidal coasts. Sci. Total Environ. 593-594, 173–181. https://doi.org/10.1016/j.scitotenv.2017.03.161.

Panidhapu, A., Li, Z., Aliashrafi, A., Peleato, N.M., 2020. Integration of weather conditions for predicting microbial water quality using Bayesian belief networks. Water Res. 170, 115349. https://doi.org/10.1016/j.watres.2019.115349.

Park, Y., Kim, M., Pachepsky, Y., Choi, S.H., Cho, J.G., Jeon, J., Cho, K.H., 2018. Development of a nowcasting system using machine learning approaches to predict fecal contamination levels at recreational beaches in Korea. J. Environ. Qual., 1094–1102 https://doi.org/10.2134/jeq2017.11.0425.

Parkhurst, D.F., Brenner, K.P., Dufour, A.P., Wymer, L.J., 2005. Indicator bacteria at five swimming beaches–analysis using random forests. Water Res. 39, 1354–1360. https://doi.org/10.1016/j.watres.2005.01.001.

Racine, J.S., Parmeter, C.F., et al., 2012. Data-Driven Model Evaluation: A Test for Revealed Performance. Department of Economics, McMaster Univ https://doi.org/10.1093/oxfordhb/9780199857944.013.010.

R-Core-Team, 2006. R: A Language and Environment for Statistical Computing. https://doi.org/10.1890/0012-9658(2002)083[3097:CFHIWS]2.0.CO;2.

Rice, J.S., Emanuel, R.E., 2017. How are streamflow responses to the el nino southern oscillation affected by watershed characteristics? Water Resour. Res. 53, 4393–4406. https://doi.org/10.1002/2016WR020097.

Roli, F., Giacinto, G., Vernazza, G., 2001. Methods for designing multiple classifier systems. Intern. Works. Multip. Classi. Syst. Springer, pp. 78–87 https://doi.org/10.1007/3-540-48219-9_8.

Rossi, A., Wolde, B.T., Lee, L.H., Wu, M., 2020. Prediction of recreational water safety using Escherichia coli as an indicator: case study of the Passaic and Pompton rivers, New Jersey. Sci. Total Environ. 714, 136814 URL. https://doi.org/10.1016/j.scitotenv.2020.136814 doi:10.1016/j.scitotenv.2020.136814.

Scutari, M., Denis, J.B., 2014. Bayesian Networks: With Examples in R. Chapman and Hall/CRC. https://doi.org/10.1111/biom.12369.

Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling: some past experiences and new approaches. J. Hydroinf. 10, 3–22. https://doi.org/10.2166/hydro.2008.015.

Solomatine, D., See, L.M., Abrahart, R., 2009. Data-driven modelling: Concepts, approaches and experiences. Practi. Hydroinfo. Springer, pp. 17–30 https://doi.org/10.1002/hyp. 6592.

Thoe, W., Wong, S.H., Choi, K., Lee, J.H., 2012. Daily prediction of marine beach water quality in Hong Kong. J. Hydro-environ. Res. 6, 164–180. https://doi.org/10.1016/j.jher.2012.05.003.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M., Boehm, A., 2014a. Predicting water quality at Santa Monica beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. 67, 105–117. https://doi.org/10.1016/j.watres.2014.09.001.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M., Boehm, A., 2014b. Sunny with a chance of gastroenteritis: predicting swimmer risk at California beaches. Environ. Sci. Technol. 49, 423–431. https://doi.org/10.1021/es504701j.

Zhai, B., Chen, J., 2018. Development of a stacked ensemble model for forecasting and analyzing daily average pm 2.5 concentrations in Beijing, China. Sci. Total Environ. 635, 644–658. https://doi.org/10.1016/j.scitotenv.2018.04.040.

Zhang, Z., Deng, Z., Rusch, K.A., 2012. Development of predictive models for determining enterococci levels at Gulf Coast beaches. Water Res. 46, 465–474. https://doi.org/10.1016/j.watres.2011.11.027.