Predicting recreational water quality advisories: a comparison of statistical methods

Wesley Brooks[a], Rebecca Carvin[a], Steven Corsi[a], Michael Fienen[a]

[a] *Wisconsin Water Science Center, United States Geological Survey, 8505 Research Way, Middleton, WI 53562*

Abstract

Epidemiological studies have indicated that the concentration of fecal indicator bacteria (FIB) in beach water is associated with illnesses among people who have contact with the water. In order to mitigate public health impacts, many beaches are managed so that an advisory is posted when the concentration of FIB exeeds a beach action value. The most commonly used method of measuring FIB concentration takes $18-24$ hours before returning a result. It has become common to base beach management decisions on the output from nowcast models that use environmental and meteorological conditions to predict the current concentration of FIB, avoiding the 24h lag. Most commonly, nowcast models are estimated using ordinary least squares regression, but other regression methods from the statistical and machine learning literature are appearing in growing numbers. The choice of regression method is quite important to the accuracy of the nowcast model, and the literature comparing the performance of different methods has often made those comparisons at a single site, which may or may not be representative. We compare several regression methods to identify which produces the most accurate predictions. The comparison is made at several sites in Wisconsin, including beaches on Lake Superior and Lake Michigan.

Email addresses: wrbrooks@usgs.gov (Wesley Brooks), rbcarvin@usgs.gov (Rebecca Carvin), srcorsi@usgs.gov (Steven Corsi), mnfienen@usgs.gov (Michael Fienen)

A random forest model is identified as the most accurate. That is followed by the adaptive Lasso, which also includes a variable selection step that reduces the number of covariates that need to be measured in order to make predictions.

Keywords: Beach water quality, Statistical model, Performance evaluation, Real-time prediction

---

1. Introduction

Fecal indicator bacteria (FIB) in beach water are often used to indicate contamination by harmful pathogens (Cabelli et al., 1979; Wade et al., 2006, 2008; Fleisher et al., 2010). The United States Environmental Protection Agency (USEPA) has established, through epidemiological studies, that FIB concentration is associated with human health outcomes (Cabelli, 1983; Dufour, 1984; USEPA, 1986). Accordingly, many states have established regulatory standards for water quality; Wisconsin states that a beach should be posted with a swimmer's advisory when the concentration of the FIB *Escherichia coli* exceeds the beach action value (BAV) of 235 colony forming units (CFU) / 100 mL (USEPA, 2012; Wisconsin Department of Natural Resources, 2012). The BAV of 235 CFU / 100 mL was recommended by the USEPA as the "do not exceed" threshold in order to limit gastrointestinal illnesses among those coming into contact with beach water to 36 cases per 1000 people (USEPA, 2012). Traditional analysis methods for FIB concentration requires $18-24$ hours for culturing a sample, so the decision to post an advisory is often made based on the previous day's FIB concentration, which is the so-called "persistence model" for beach management (USEPA, 2007). Previous research has shown that the concentration of FIB in beach water can vary substantially during the $18-24$ h analysis period, with the result

that the persistence model often provides incorrect information for posting warnings (Whitman et al., 2004; Whitman and Nevers, 2008). Thus, at beaches managed using the persistence model, the public is sometimes exposed to health risks or unnecessarily deprived of recreation opportunities.

In order to have more immediate knowledge of the FIB concentration, it is now common to use regression models that "nowcast" the FIB concentration based on easily observed surrogate covariates, e.g. turbidity and running 24 h rainfall total (Brandt et al., 2006; Olyphant and Whitman, 2004). Numerous regression techniques have been used to generate nowcast models of FIB concentration. The techniques include ordinary least squares (OLS) (Nevers and Whitman, 2005; Francy and Darner, 2007), partial least squares (PLS) (Hou et al., 2006; Brooks et al., 2013), logistic regression (Waschbusch et al., 2004; Jin and Englande Jr., 2006), decision trees (Stidson et al., 2012), random forests (Parkhurst et al., 2005; Jones et al., 2012), and artificial neural networks (Kashefipour et al., 2005; He and He, 2008). A thorough review of the regression techniques being used in nowcast models for FIB concentration is provided by (de Brauwere et al., 2014). An assessment of seven methods of regression for FIB concentration in beach water at Santa Monica Beach in California identified classification trees, artificial neural networks, and logistic regression as the three best methods (Thoe et al., 2014).

Ordinary least squares (OLS) regression is the most commonly used regression technique in the nowcast models (de Brauwere et al., 2014). However, OLS is well-known for drawbacks like overfitting, difficulty of covariate selection, and the inflexibility of its linear modeling structure (Ge and Frick, 2007). The literature suggests that many

regression techniques have been successfully used for nowcast modeling, but due to differences in such factors as local conditions, data handling, and performance validation, it is not possible to identify the best regression technique for nowcast modeling by comparing different models at different sites. In this study, fourteen regression techniques are evaluated in nowcast models at seven Wisconsin Great Lakes beaches with four years of data. The results are compared to identify the techniques that most accurately predict instances when a swimmer's advisory should be posted. This comparison is designed to provide insights that may be lost when comparing individual methods at single sites.

The remainder of the paper is organized as follows: in the next section we discuss data collection and handling, describe the regression techniques, and explain how the comparisons were made. Next, we present the results of comparing the methods by several metrics including: area under the receiver operating characteristic (ROC) curve; predictive error sum of squares; and raw number of correct/incorrect predictions. Finally, we discuss what the comparison suggests about which are the best choices for a regression technique in a nowcast model.

2. Data

The seven beach sites analyzed in this study are located within two distinct regions of Wisconsin (Figure 1). Three of the sites are on Chequamegon Bay in Lake Superior and the remaining five are in Manitowoc County in Lake Michigan. For each site in the study, the data used to estimate the predictive models for FIB concentration were measured by a combination of automatic sensing, hydrodynamic and atmospheric

modeling, and manual sampling. A listing of the covariates included for modeling the FIB concentration at each beach site is in the Appendix.

[[Figure 1 about here]]

2.1. Site descriptions

2.1.1. Chequamegon Bay sites

Chequamegon Bay is approximately 19 km long and ranges from 3 to 10 km in width, with a maximum depth of 11 m. Water quality at the three Chequamegon Bay/Lake Superior beaches is influenced by nearby streams, as well as by urban runoff from Ashland and Washburn, Wisconsin. Thompson beach is within the small town of Washburn, on the north side of the bay. Next to the beach are a playground, RV campsites, piers and boat launch. There are two flowing artesian wells that drain to the beach. Thompson Creek, about 300 m from the beach, is the nearest stream. Maslowski beach is on the west side of Ashland, on the south side of the Chequamegon Bay. A playground and parking area are near the beach. Two flowing artesian wells are near the swim area, and Fish Creek (1.5 km west of the beach) and Whittlesley Creek (3 km northwest) are the nearest streams. Kreher beach is in Ashland, 4 km northeast of Maslowski beach. Kreher Park has an RV campground, playground and boat launch, and is nearest to Bay City Creek, which is 1 km east of the beach. All of the listed streams are influenced by areas of agricultural and forested land use, with Bay City Creek also influenced by urban land use (Francy et al., 2013). Contributions from Fish Creek are dynamic due to a wetland at the creek's outlet that is influenced by the lake level.

2.1.2. Manitowoc County sites

Red Arrow beach is within the city of Manitowoc. It has numerous potential influences on water quality, including the mouth of the Manitowoc River one mile north and urban runoff draining to the beach through storm sewer outlets. The Manitowoc River is dominated by agricultural land use, but there is some urban influence from the city of Manitowoc. The Manitowoc sewage treatment plant sits at the mouth of the Manitowoc River. Neshotah beach is in the small community of Two Rivers. Small storm sewers drain to the north and to the south directly adjacent to the beach boundaries, and the mouth of the Twin River is 1 km south of the beach. The Twin River drains an agricultural watershed. Point Beach State Park is approximately 18 km north of Manitowoc, about 4 km north of the mouth of Molash Creek whose watershed encompasses a mix of agricultural land use and wetland area. The mouth of Twin River is 10 km south and the mouth of the Kewaunee River is 26 km north of Point Beach State Park. The Kewaunee River is also dominated by agricultural land use. FIB concentration was measured at three beaches within Point Beach State Park, with the samples being considered in the models as three independent observations. Hika beach is south of the city of Manitowoc near the small community of Cleveland. Large floating mats of *Cladophora* algae are common. Centerville Creek, a small stream dominated by agricultural land use, drains to the lake adjacent to the beach.

2.2. Data sources

Field data collection and sample analysis followed methods described in Francy et al. (2013). Concentration of *E. coli* was measured at each beach $2-4$ times per week

for $12 - 14$ weeks between Memorial Day and Labor Day from 2010 through 2013. Samples were collected from the center of the length of the beach, 30 cm below the water surface where total water depth was 60 cm. All samples were quantified by use of the Colilertő QuantiTray/2000 method, which were reported as the most probable number (MPN) of *E. coli* colony forming units (CFU) and were read after 24 hours of incubation (National environmental methods index, 2013).

Additional covariates were compiled from a variety of sources including online data and manual measurements. Online data were accessed using Environmental Data Discovery and Transformation (EnDDaT), a web service that accesses data from a variety of sources, compiles and processes the data, and performs common transformations (USGS, 2014a). Three sources of data were accessed: The U.S. Geological survey National Water Information System (NWIS) (USGS, 2014b), the National Weather Service North Central River Forecasting Center (National Oceanic and Atmospheric Administration, 2012), and the Great Lakes Costal Forecasting System (Schwab and Bedford, 1999). Covariates acquired through these sources included: river discharge, precipitation, lake current vectors, wave height, wave direction, lake level, water temperature, air temperature, wind vector, and percent cloud cover.

Most covariates from online sources were available in hourly increments with the exception of NWIS data which were available in 15 minute increments. In order to make best use of this high-frequency data for daily predictions, several summary statistics were calculated over several time windows for use as potential covariates. The use of 1, 2, 6, 12, 24, 48, 72, and 120 hour time windows for calculating the summary statistics followed recent research showing that selecting from windowed and lagged versions

of raw high-frequency covariates can improve the predictive accuracy of regression models (Cyterski et al., 2012). The choice of summary statistics to include as potential covariates was guided by scientific judgement regarding phenomena that could affect the FIB concentration. For example, standard deviation of water temperature measurements over the window period reflected the variability in water temperature, which may affect the survival and growth of FIB; the sum of rainfall measurements over the window period indicated the magnitude of recent rain events, which may be associated with FIB washed into the lake from sources on land; and the mean of cloud cover measurements over the window period may measure the degree to which UV light was inhibited from breaking down FIB colonies in the water. The summary statistics computed by EnDDaT were the mean, minimum, maximum, difference, sum, and standard deviation.

Manually observed data were instantaneous observations that had the benefit of being measured when and where the FIB samples were collected. However, these covariates were measured only once per day and at greater expense than the online data because the data had to be collected by field personnel. Manual data collection techniques were guided by the USEPA's Great Lakes Beach Sanitary Survey (USEPA, 2008). Among the manually measured data were turbidity, wave height, number of birds present, number of people present, amount of algae floating in the swim area and on the beach, specific conductance, water and air temperature, wind direction, and wind speed. Every beach dataset included turbidity, but other field covariates occasionally had to be dropped from some of the datasets because of missing values or questionable reliability.

2.3. Data transformations

The response for the continuous regression models was the base-10 logarithm of the FIB concentration. For the binary regression models, the response is an indicator of whether the concentration exceeds the BAV. Transformations were applied to some of the covariates during pre-processing: the beach water turbidity and the discharge of tributaries near each beach were log-transformed, and rainfall covariates were all square root transformed. These transformations were based on the performance of previous studies and were applied to all datasets (Ge and Frick, 2007; Frick et al., 2008).

3. Methods

3.1. Definitions

For each site, let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the vector of $\log_{10}$ FIB concentration measurements, let n be the number of observations, and let p be the number of explanatory covariates. The beach action value (BAV) of 235 CFU / 100 mL is represented symbolically in equations by δ. Define an exceedance as a meansured FIB concentration that exceeds the BAV. Conversely, a nonexceedance is a measured FIB concentration that does not exceed the BAV.

Predictions are the result of applying a model to data that was not used to estimate the model. The predicted $\log_{10}$ FIB concentration is denoted by a tilde (e.g., $\tilde{y}_i$). On the other hand, applying the model to the same data as was used to estimate the model produces fitted values, which are denoted by a hat (e.g., $\hat{y}_j$). We define a

predicted exceedance as when a model predicts that the FIB concentration exceeds the BAV. This is not the same as $\tilde{y}_i > \delta$ because predictions are compared to a decision threshold $\hat{\delta}$ rather than to the BAV δ. The decision threshold $\hat{\delta}$ is a parameter that can be adjusted to tune the predictive performance. For instance, increasing the decision threshold reduces the number of false positives but increases the number of false negatives. Setting the decision threshold is an important detail that is discussed in Section 5.4.

3.2. Listing of statistical techniques

Fourteen different regression modeling techniques were considered (Table 1). Each technique uses one of five modeling algorithms: the gradient boosting machine (GBM), the adaptive Lasso (AL), the genetic algorithm (GA), partial least squares (PLS), or sparse PLS (SPLS). Each technique is applied to either continuous or binary regression and to either covariate selection and model estimation, or covariate selection only.

3.2.1. Continuous vs. binary regression

The goal of predicting exceednaces of the water quality standard is approached in two ways: one is to predict the bacterial concentration and then compare the prediction to a threshold, which is referred to as continuous modeling. The other is referred to as binary modeling, in which we predict the state of the binary indicator z_i:

$$z_i = \begin{cases} 0 & \text{if } y_i < \delta \\ 1 & \text{otherwise} \end{cases}$$

where y_i is the FIB concentration and δ is the BAV. The indicator is coded as zero when the concetration is below the BAV and one when the concentration exceeds the BAV. All of the binary modeling techniques herein use logistic regression (Hosmer Jr and Lemeshow, 2004). Binary regression methods are indicated with a (b).

3.2.2. Weighting of observations in binary regression

The concentration of FIB in the water at a single beach on a single day can be subject to a large degree of spatiotemporal heterogeneity (Whitman and Nevers, 2004). Thus, when the concentration in a sample is observed to fall near the BAV, there is considerable uncertainty as to whether an independent sample from the same date and location would or would not exceed the BAV. A weighting scheme for the binary regression techniques was designed to reflect this ambiguity by giving more weight to observations far from the BAV. In the weighting scheme, observations were given weights w_i for $i = 1, \ldots, n$, where

$$
\begin{aligned}
w_i &= (y_i - \delta)/\hat{\mathrm{sd}}(y) \\
\hat{\mathrm{sd}}(y) &= \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / n} \\
\bar{y} &= \sum_{i=1}^{n} y_i / n.
\end{aligned}
$$

That is, the weights are equal to the number of standard deviations that the observed concentration lies from the BAV. All the techniques that were implemented with this weighting scheme were separately implemented without any weighting of the observations. The methods using the weighting scheme are indicated by (w).

3.2.3. Selection-only methods

Modeling methods indicated by an (s) were applied only to select covariates for a regression model. Once the covariates were selected, the regression model using the selected covariates was estimated using ordinary least squares for the continuous methods, or ordinary logistic regression for the binary methods.

3.2.4. Listing of modeling algorithms

GBM. A GBM model is a so-called random forest model - a collection of many regression trees, each fit to a randomly drawn subsample of the training data (Friedman, 2001). Prediction is done by averaging the outputs of the trees. Two GBM-based techniques were studied - we refer to them as GBM-OOB and GBM-CV. The difference between these two techniques is in how the optimal number of trees is determined - GBM-CV selects the number of trees in a model using leave-one-out cross validation (CV), while GBM-OOB uses the so-called out-of-bag error estimate, where the predictive error of each tree is estimated by its predictive error over the observations that were left out when fitting the tree. In contrast, the predictive error of CV is estimated from observations that are left out from the training data altogether, and are therefore not used in the fitting of any trees. The CV method is much slower (it has to construct as many random forests as there are observations, while the OOB method only requires computing a single random forest). However, GBM-CV should more accurately estimate the prediction error.

Adaptive Lasso. The least absolute shrinkage and selection operator (Lasso) is a penalized regression method that simultaneously selects relevant covariates and esti-

mates their coefficients (Tibshirani, 1996). The AL is a refinement of the Lasso that possesses the so-called "oracle" properties of asymptotically selecting exactly the correct covariates and estimating them as accurately as would be possible if their identities were known in advance (Zou, 2006). To use the AL for prediction requires selecting a tuning parameter. In this study, the AL tuning parameter λ was selected to minimize the corrected Akaike Information Criterion ($\mathrm{AIC_c}$) (Akaike, 1973; Hurvich et al., 1998). The $\mathrm{AIC_c}$ for the continuous regression models is

$$\mathrm{AIC_c} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} + 2df + \frac{2df(df+1)}{n - df - 1}. \tag{1}$$

where $\hat{\sigma}^2$ is the variance estimate from the model that used all of the covariates, n is the sample size, y_i and $\hat{y}_i$ are respectively the observed and fitted value of the ith FIB measurement, and df is the number of covariates in the model. For binary regression models, the $\mathrm{AIC_c}$ is

$$\mathrm{AIC_c} = \sum_{i=1}^{n} 2\left\{ z_i \log(\frac{z_i}{\hat{z}_i}) + (1 - z_i) \log\left(\frac{1 - z_i}{1 - \hat{z}_i}\right)\right\} + 2df + \frac{2df(df+1)}{n - df - 1} \tag{2}$$

where z_i and $\hat{z}_i$ are respectively the observed and fitted value of the ith BAV exceedance indicator.

Genetic algorithm. The GA was used to select covariates for either an OLS or a logistic regression model. By analogy to natural selection, so-called chromosomes in the

GA represent regression models (Fogel, 1998). A covariate is included in the model if the corresponding element of the chromosome is one, but not otherwise. Chromosomes are produced in successive generations, where the first generation is produced randomly and subsequent generations are produced by combining chromosomes from the current generation, with additional random drift. The chance that a chromosome in the current generation produces offspring in the next generation is an increasing function of its fitness. The fitness of each chromosome is calculated by the $\mathrm{AIC_c}$ (1), (2).

PLS. Partial least squares (PLS) regression is a tool for building regression models with many covariates (Wold et al., 2001). PLS works by decomposing the covariates into mutually orthogonal components, with the components then used as the covariates in a regression model. This is similar to principal components regression (PCR), but the way PLS components are chosen ensures that they are aligned with the response, whereas PCR is sometimes criticised for decomposing the covariates into components that are unrelated to the response. To use PLS, one must decide how many components to use in the model. Following (Brooks et al., 2013), we use the PRESS statistic to select the number of components.

SPLS. Sparse PLS (SPLS) combines the orthogonal decompositions of PLS with the sparsity of Lasso-type covariate selection (Chun and Keles, 2010). To do so, SPLS uses two tuning parameters: one that controls the number of orthogonal components and one that controls the Lasso-type penalty. The optimal parameters are those that minimize the mean squared prediction error (MSEP) over a two-dimensional grid search. The MSEP is estimated by 10-fold cross-validation.

Name	Algorithm	Binary	Weighted	Selection Only
GBM-OOB	Gradient boosting			
GBM-CV	Gradient boosting			
AL	Adaptive Lasso			
AL (s)	Adaptive Lasso			X
AL (b)	Adaptive Lasso	X		
AL (b,w)	Adaptive Lasso	X	X	
AL (s,b)	Adaptive Lasso	X		X
AL (s,b,w)	Adaptive Lasso	X	X	X
GA	Genetic algorithm			
GA (b)	Genetic algorithm	X		
GA (b,w)	Genetic algorithm	X	X	
PLS	Patrial least squares			
SPLS	Sparse partial least squares			
SPLS (s)	Sparse partial least squares			X

Table 1: Comprehensive list of the modeling methods analyzed in this study. Listed for each method are the method's abbreviation, the algorithm used by the method, and indicators of whether the method uses binary regression, observation weighting, and/or covariate selection separately from estimation. The two methods GBM-CV and GBM-OOB differ in how the number of GBM trees are selected, as described in the main text.

3.3. Cross Validation

Our assessment of the modeling techniques was based on their performance in predicting exceedances of the BAV. Two types of cross validation were used to measure the performance in prediction: leave-one-out (LOO) and leave-one-year-out (LOYO). In LOO CV, one day's observation was held out for validation while the rest of the data was used to train a model. At Point Beach State Park, where FIB concentration was measured at three locations each day, all three daily observations were left out of the LOO CV models together. The model was used to predict the result of the held out observation(s), and the process - including estimating a new predictive model - was repeated for each date with available data. On the other hand, each cycle of LOYO CV held out an entire year's worth of data for validation instead of a single observation. It was intended to approximate the performance of the modeling technique under a typical use case: a new model is estimated before the start of each annual beach season and then used for predicting exceedances during the season. The LOYO models in this study were estimated using all the available data except for the held out year, even that from future years. So for instance the 2012 models were estimated using the 2010-2011 and 2013 data.

Some methods also used CV internally to select tuning parameters. In those cases the internal CV was conducted by subdividing the model data, and never looking at the held-out observation(s). This process was independent of the CV to assess predictive performance.

Row	$\log_{10}$ FIB	PLS (LOO)	PLS (LOYO)	...	SPLS (LOO)	SPLS (LOYO)
1	2.54	2.35	2.22	...	2.29	2.55
2	2.59	1.87	1.79	...	1.91	1.23
⋮	⋮	⋮	⋮	...	⋮	⋮
166	1.57	1.93	2.06	...	1.83	2.07
167	3.38	1.84	2.01	...	1.80	1.71

Table 2: An example of how the results for a site (Hika here) were compiled into a results table. The summary statistics used to compare predictive performance (area under the ROC curve and predictive error sum of squares) were calculated from the table. Confidence intervals for the summary statistics were computed via the bootstrap by resampling (with replacement) the rows of the results table.

3.4. Comparing methods, and quantifying uncertainty in the ranks

Results were compiled into one table for each site where each observation corresponds to a row in the table. For example, a few rows from the results table at Hika are presented in Table 2. The results table has a column for the observed $\log_{10}$ FIB concentration and, for each method, columns for the predicted concentration by LOO CV and by LOYO CV. From the table, performance of the modeling methods was summarized by calculating the area under the receiver operating characteristic (ROC) curve (AUROC) and the predictive error sum of squares (PRESS).

The ROC curve is an assessment of how well predictions are separated into exceedances and nonexceedances (Hanley and McNeil, 1982). Every possible value of the decision threshold $\hat{\delta}$ corresponds to a point on the ROC curve, with coordinates $(1-\text{specificity}(\hat{\delta}), \text{sensitivity}(\hat{\delta}))$. Specificity is the fraction of decision threshold nonexceedances that have been correctly predicted.Sensitivity is the fraction of decision threshold exceedances that have been correctly predicted. Specificity and sensitivity are mathematically defined as

$$\text{specificity}(\hat{\delta}) = \sum_{i=1}^{n} I(\tilde{y}_i \leq \hat{\delta}) I(y_i \leq \delta) / \sum_{j=1}^{n} I(y_j \leq \delta)$$
$$\text{sensitivity}(\hat{\delta}) = \sum_{i=1}^{n} I(\tilde{y}_i > \hat{\delta}) I(y_i > \delta) / \sum_{j=1}^{n} I(y_j > \delta).$$

where $I(A)$ is the indicator function that takes value one if A is true and zero is A is false.

The AUROC averages the model's performance over the range of possible thresholds. A model which perfectly separates exceedances from non-exceedances in prediction would have an AUROC of one, while a model that predicts exceedances no better than a coin flip would have an expected AUROC of 0.5.

While AUROC quantifies how well a model sorts exceedances and non-exceedances, PRESS measures how accurately a model's predictions match the observed FIB concentration. The PRESS can only be computed for continuous regression methods. Recalling that the ith observed $\log_{10}$ FIB concentration is denoted y_i and that the corresponding prediction is denoted $\tilde{y}_i$ for $i = 1, \ldots, n$ where n is the total number of predictions, the PRESS is given by

$$\text{PRESS} = \sum_{i=1}^{n} \left(\tilde{y}_i - y_i\right)^2.$$

To identify which modeling methods had the best performance across all sites, the methods at each site were ranked from worst to best according to AUROC and PRESS

(the ranks were taken worst to best so that larger numbers represent better performance). The mean rank of each method was then taken across the sites as a measurement of how each of our modeling methods performed relative to the others. Uncertainty in the rankings was quantified by the bootstrap: since PRESS and AUROC are functions of the results tables, the bootstrap procedure was carried out by resampling the rows of each results table and recalculating the ranks for each bootstrap sample. We used $1,000$ bootstrap samples of each results table in the analysis that follows.

4. Results

4.1. AUROC

The mean LOO and LOYO ranks were computed for all of the methods as determined by AUROC (Figure 2). The three top-ranked methods were GBM-CV, GBM-OOB, and AL. In order to facilitate a pairwise comparison between modeling methods, the frequency that the mean AUROC rank of GBM-OOB, GBM-CV, or AL exceeded each of the other modeling methods for the leave-one-year-out and the leave-one-out analyses were also computed (Table 3).

[[Figure 2 about here]]

4.2. PRESS

The PRESS statistic is of interest because a good model should accurately predict the bacterial concentration, but for assessing regression models for FIB concentration,

Leave-one-year-out cross-validation:

	GBM-OOB	AL	AL (s)	AL (b,w)	SPLS	PLS	AL (b)	SPLS (s)	AL (b,s)	AL (b,w,s)	GA	GA (b,w)	GA (b)
GBM-CV	0.86	0.87	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		0.69	0.92	0.95	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
AL			0.96	0.87	1.00	0.97	0.98	1.00	1.00	1.00	1.00	1.00	1.00

Leave-one-out cross-validation:

	GBM-OOB	AL	AL (b,w)	AL (s)	AL (b,w,s)	GA	SPLS	SPLS (s)	PLS	GA (b)	AL (b)	AL (b,s)	GA (b,w)
GBM-CV	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GBM-OOB		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AL			0.72	0.99	0.93	0.96	0.98	1.00	0.99	1.00	1.00	1.00	1.00

Table 3: Under leave-one-year-out (top) and leave-one-out (bottom) cross validation, frequency of the mean AUROC rank of GBM-OOB, GBM-CV, or AL (in the rows) exceeding that of the other methods (in the columns).

AUROC is more important than PRESS because it directly measures the models' abilities to distinguish exceedances from nonexceedances. That said, we expect the two statistics to usually agree on which modeling methods perform best.

The top three techniques under both LOO and LOYO analysis as determined by PRESS were GBM-CV, GBM-OOB, and AL (Figure 3). Again, to facilitate a pairwise comparison between modeling methods, the frequency that the mean PRESS rank of GBM-OOB, GBM-CV, or AL exceeded each of the other modeling methods for the leave-one-year-out and the leave-one-out analyses were computed (Table 4).

[[Figure 3 about here]]

4.3. Narrowing the focus

By AUROC and PRESS, and for LOO and LOYO analyses, the three highest-ranked modeling methods were GBM-CV, GBM-OOB, and AL. The fourth-ranked method was not consistent across the different analyses. By the LOO CV analysis, AL was ranked better than the fourth-ranked method by AUROC, AL (b,w), on 72% of bootstraps and better than the fourth-ranked method by PRESS, SPLS (s), on 73%

Leave-one-year-out cross-validation:

	GBM-OOB	AL	SPLS	PLS	SPLS (s)	AL (s)	GA
GBM-CV	0.58	0.96	0.97	1.00	1.00	1.00	1.00
GBM-OOB		0.94	0.96	1.00	1.00	1.00	1.00
AL			0.52	0.88	0.94	0.99	1.00

Leave-one-out cross-validation:

	GBM-OOB	AL	SPLS (s)	SPLS	PLS	AL (s)	GA
GBM-CV	0.66	0.99	1.00	1.00	1.00	1.00	1.00
GBM-OOB		0.97	1.00	1.00	1.00	1.00	1.00
AL			0.73	0.80	0.85	0.93	1.00

Table 4: Under leave-one-year-out (top) or leave-one-out (bottom) cross validation, frequency of the mean PRESS rank of GBM-CV, GBM-OOB, or AL (in the rows) exceeding that of the other methods (in the columns).

of bootstraps. And by the LOYO CV analysis, AL was ranked better than the fourth-ranked method by AUROC, AL (s), on 96% of bootstraps and better than the fourth-ranked method by PRESS, SPLS, on 52% of bootstraps. Therefore, we consider only the GBM methods and AL for the following analyses because they consistently outperform the other methods.

4.4. Classification of responses, and the decision threshold

In operational use, a model's performance will be judged by how well it distinguishes between exceedances and nonexceedances. While AUROC measures how well exceedances and nonexcedances are sorted among the predictions, AUROC is an average accuracy over all possible thresholds. In order to provide the assessment most relevant for operational use, the LOYO CV results were used to simulate how many correct and incorrect predictions would be seen from an AL, GBM-OOB, or GBM-CV model

with a specific choice of decision threshold. Using the LOYO CV results simulates the common scenario that a model is estimated at the beginning of each beach season and used to make predictions during that season, with a new model incorporating the new season of data estimated the following year into the new model's training data.

Intuitively, the decision threshold should adapt to the conditions that are observed in the beach's training data. If, for instance, exceedances were rare in the training data, then we expect few exceedances in the future, and should set the decision threshold high to reflect this expectation. On the other hand, if the bacterial concentration often exceeds the BAV, then the decision threshold should be set lower in order to properly flag more of those exceedances. This intuition was encoded into how the decision threshold was set for the LOYO models. Specifically, the decision threshold $\hat{\delta}$ was set to the q^{th} quantile of the fitted values of non-exceedances in the training set, where q is the proportion of training set observations that are non-exceedances.

In Figure 4, we examine the counts on a per-beach basis of four categories of decisions: true negatives (correct predictions of nonexceedances), false positives (incorrect predictions of exceedances) true positives (correct predictions of exceedances), and false negatives (incorrect predictions of nonexceedances). In most cases, the counts were similar between the three techniques, with GBM-OOB and GBM-CV commonly resulting in a few more correct decisions than AL. There are, however, exceptions where AL results had more correct decisions (e.g., Hika and Red Arrow).

[[Figure 4 about here]]

4.5. Covariate selection

It was noted in Section 4.3 that GBM-OOB and AL are two of the three best-ranked methods. One difference between the two is that AL does covariate selection while GBM-OOB uses all of the available covariates. We explore here how many covariates were used in AL models compared to the GBM-OOB models (Figure 5).

[[Figure 5 about here]]

At most of the sites, AL uses only a small fraction of the available covariates, but at Point beach, AL uses almost all of the available covariates. This is due to the covariate selection criterion we used (AIC_c) which is intended to minimize predictive error. As the amount of data increases, we accumulate enough information to discern an effect even of covariates that are only slightly correlated with the response. As our dataset grows, then, we should expect more covariates to be selected for an AL model, and Point has far more observations than the other sites.

5. Conclusions

The GBM-CV, GBM-OOB, and AL methods showed the best results by both PRESS and AUROC, under LOO and LOYO cross validation. Though GBM-CV was a bit more accurate than GBM-OOB in all the settings, the small improvement in accuracy may not outweigh the large additional cost in computational time to fit the model. However, the additional computational cost is incurred only once when the model is estimated - given a new observation of beach data, both the GBM-CV and GBM-OOB models produce predictions nigh-instantaneously. Where predictive accuracy

is the most important consideration and no difficulty is anticipated in acquiring the data, it is hard to argue against using a GBM-type model.

The predictive performance of the AL models was somewhat worse than that of the GBM models, but by including a covariate selection step, the AL models reduce the number of covariates that must be measured in order to make daily predictions. A model that requires fewer covariates is less expensive and more robust (since each additional covariate increases the chance of missing data). This is particularly important for manually-collected covariates because collecting data by hand takes more time and is more costly than accessing publically available data from a web service. Across all of the sites, the ratio of manually-collected to automatically collected covariates in the AL models seems to mirror the ratio among all available covariates, indicating that neither the manually- nor automatically-collected covariates are systematically more important to predicting the bacterial concentration. Some covariates tended to appear at every site in the AL models (and other models that include a covariate selection step). The manually-collected covariates that were consistently selected for the models were the (log) turbidity in the beach water, and wave height at the beach.

Another advantage of the AL over GBM-type models is interpretability. As a linear regression technique, fitting an AL model means generating a set of coefficients, which can be interpreted as the marginal effect of a change in the corresponding covariate. On the other hand, GBM produces black-box models that typically make more accurate predictions but are difficult to intrepret. One common way to interpret a random forest model (such as from the GBM algorithm) is to observe the proportion of splits in the underlying trees that involve a particular covariate. The split proportion is a

measurement of that covariate's importance to the model but gives no indication of how the covariate affects the bacterial concentraion.

6. Discussion

Where minimizing the number of covariates is important, the selection criterion used here (AIC_c) may not be ideal. In that case it may be advantageous to use a criterion that is more parsimonious about including covariates in the model, such as the Bayesian information criterion (BIC), for which the penalty term $2df + 2df(df + 1)/(n - df - 1)$ in (1) or (2) would be replaced by $n \times df$, which grows with the sample size n. The BIC does not exhibit the property of the AIC (or AIC_c) where more covariates are included in the model as the number of observations increases. However, the BIC is derived from the standpoint of identifying the most probable model, rather than minimizing the predictive error. It is therefore likely that an AL model using the BIC for covariate selection will have slightly worse predictive performance than one using the AIC_c.

All statistical methods and the comparison for this study were carried out in the R statistical software environment (R Core Team, 2014). Scripts and details of the how the modeling methods were implemented are in the online supplement. Often times, beach management practitioners are not very familiar with statistical analysis and rely on more accessible software to help guide them through development of models for recreational water quality predictions. For this purpose, the Virtual Beach software was developed (Cyterski et al., 2013). The GA, PLS, and GBM algorithms used in this study are based on simplified versions of the comparable algorithms in Virtual

Beach version 3.0. An implementation of AL is also an anticipated addition to Virtual Beach.

7. Acknowledgments

The predictive models for this study were generated on facilities and software (HT-Condor) provided by the University of Wisconsin-Madison's Center for High Throughput Computing.

8. Figure Captions

8.1. Figure 1

Map showing the location of the seven Wisconsin beaches for which models were analyzed in this work.

8.2. Figure 2

Mean ranking of the methods across all seven sites by area under the receiver operating characteristic curve (AUROC). The error bars are 90% confidence intervals computed by the bootstrap. At left are the AUROC rankings from the leave-one-year-out cross validation (a), at right are the AUROC rankings from the leave-one-out cross validation (b).

8.3. Figure 3

Mean ranking of the methods by predictive error sum of squares (PRESS) across all sites (higher is better). The error bars are 90% confidence intervals computed by the bootstrap. At left are the PRESS rankings from the leave-one-year-out cross validation (a), at right are the PRESS rankings from the leave-one-out cross validation (b).

8.4. Figure 4

At each site, the number of predictions from AL, GBM-OOB, and GBM-CV that fell into four categories, from left: true negatives, false positives, true positive, and false negatives.

8.5. Figure 5

At each site, the mean number of covariates that were selected for the AL model, and the total number of covariates, all of which were used in the gradient boosting machine with an out-of-bag estimate of the optimal tree count (GBM-OOB) models. For both AL and GBM-OOB, the covariate counts are broken down by whether the covariate values were collected automatically from web services or manually at the beach.

9. References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pp. 267–281. Akademinai Kiado.

Brandt, S., D. Schwab, T. Croley, D. Belestky, and R. Whitman (2006). *Ecosystem Forecasting: Integrating Science to Reduce the Risks to Human Health*. American Geophysical Union.

Brooks, W. R., M. N. Fienen, and S. R. Corsi (2013). Partial least squares for efficient models of fecal indicator bacteria on great lakes beaches. *Journal of Environmental Management 114*, 470–475.

Cabelli, V. J. (1983). Health effects criteria for marine recreational waters. Tech report EPA-600/1-80-031, United States Environmental Protection Agency Office of Research and Development.

Cabelli, V. J., A. P. Dufour, M. A. Levin, L. J. McCabe, P. W. Haberman, and L. D. Jensen (1979). Relationship of microbial indicators to health effects at marine bathing beaches. *American Journal of Public Health 69*(7), 690–696.

Chun, H. and S. Keles (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(1), 3–25.

Cyterski, M., W. Brooks, M. Galvin, K. Wolfe, R. Carvin, T. Roddick, M. Fienen, and S. Corsi (2013). *Virtual Beach 3: User's Guide*. United States Environmental Protection Agency.

Cyterski, M., S. Zhang, E. White, M. Molina, K. Wolfe, R. Parmar, and R. Zepp (2012). Temporal synchronization analysis for improving regression modeling of fecal indicator bacteria levels. *Water, Air & Soil Pollution 223*, 4841–4851.

de Brauwere, A., N. K. Ouattara, and P. Servais (2014). Modeling fecal indicator bacteria concentrations in natural surface waters: a review. *Critical Reviews in Environmental Science and Technology 44*(21), 2380–2453.

Dufour, A. P. (1984). Health effects criteria for fresh recreational waters. Tech report EPA-600/1-84-004, United States Environmental Protection Agency Office of Research and Development.

Fleisher, J. M., L. E. Fleming, H. M. Solo-Gabriele, J. K. Kish, C. D. Sinigalliano, L. Plano, S. M. Elmir, J. D. Wang, K. Withum, T. Shibata, M. L. Gidley, A. Abdelzaher, G. He, C. Ortega, X. Zhu, M. Wright, J. Hollenbeck, and L. C. Backer (2010). The BEACHES study: health effects and exposures from non-point source microbial contaminants in subtropical recreational marine waters. *International Journal of Epidemiology 39*(5), 1291–1298.

Fogel, D. B. (1998). *Evolutionary computation: the fossil record.* Wiley-IEEE Press.

Francy, D. S., A. M. G. Brady, R. B. Carvin, S. R. Corsi, L. M. Fuller, J. H. Harrison, B. A. Hayhurst, J. Lant, M. B. Nevers, P. J. Terrio, and T. M. Zimmerman (2013). Developing and implementing the use of predictive models for estimating water quality at Great Lakes beaches. Scientific Investigations Report 2013-5166, United States Geological Survey.

539 Francy, D. S. and R. A. Darner (2007). Nowcasting beach advisories at Ohio Lake
540 Erie Beaches. Open File Report 2007-1427, United States Geological Survey.

541 Frick, W. E., Z. Ge, and R. G. Zepp (2008). Nowcasting and forecasting concentrations
542 of biological contaminants at beaches: A feasibility and case study. *Environmental*
543 *Science & Technology 42*(13), 4818–4824.

544 Friedman, J. (2001). Greedy function approximation: a gradient boosting machine.
545 *Annals of Statistics*, 1189–1232.

546 Ge, Z. and W. E. Frick (2007). Some statistical issues related to multiple linear regres-
547 sion modeling of beach bacteria concentrations. *Environmental Research 103*(3),
548 358–364.

549 Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a
550 receiver operating characteristic (ROC) curve. *Radiology 148*(1), 29–36.

551 He, L.-M. L. and Z.-L. He (2008). Water quality prediction of marine recreational
552 beaches receiving watershed baseflow and stormwater runoff in southern California,
553 USA. *Water research 42*(10), 2563–2573.

554 Hosmer Jr, D. W. and S. Lemeshow (2004). *Applied logistic regression*. John Wiley
555 & Sons.

556 Hou, D., S. J. M. Rabinovici, and A. B. Boehm (2006). Enterococci predictions
557 from partial least squares regression models in conjunction with a single-sample
558 standard improve the efficacy of beach management advisories. *Environmental*
559 *Science & Technology 40*(6), 1737–1743.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Methodology) 60*(2), 271–293.

Jin, G. and A. J. Englande Jr. (2006). Prediction of swimmability in a brackish water body. *Management of Environmental Quality 17*(2), 197–208.

Jones, R. M., L. Liu, and S. Dorovitch (2012). Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. *Environmental Monitoring and Assessment 185*(3), 2355–2366.

Kashefipour, S. M., B. Lin, and R. A. Falconer (2005). Neural networks for predicting seawater bacterial levels. *Proceedings of the Institution of Civil Engineers-Water Management 158*(3), 111–118.

National environmental methods index (2013). *Colilert Test Kit Procedure*. National environmental methods index.

National Oceanic and Atmospheric Administration (2012). North Central River Forecasting Center.

Nevers, M. B. and R. L. Whitman (2005). Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan. *Water research 39*(20), 5250–5260.

Olyphant, G. A. and R. L. Whitman (2004). Elements of a predictive model for determining beach closures on a real time basis: the case of 63rd Street Beach Chicago. *Environmental Monitoring and Assessment 98*, 175–190.

Parkhurst, D. F., K. P. Brenner, A. P. Dufour, and L. J. Wymer (2005). Indicator bacteria at five swimming beaches - analysis using random forests. *Water Research 39*(7), 1354–1360.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schwab, D. J. and K. W. Bedford (1999). The Great Lakes forecasting system. *Coastal and Estuarine Studies*, 157–174.

Stidson, R. T., C. A. Gray, and C. D. McPhail (2012). Development and use of modelling techniques for real-time bathing water quality predictions. *Water and Environment Journal 26*(1), 7–18.

Thoe, W., M. Gold, A. Griesbach, M. Grimmer, M. L. Taggart, and A. B. Boehm (2014). Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming confditions. *Water Research 67*, 105–117.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 267–288.

United States Environmental Protection Agency (1986). Ambient water quality criteria for bacteria. Technical Report EPA-440-5-84-00.

United States Environmental Protection Agency (2007). Critical path science plan for the development of new or revised recreational water quality criteria. Technical Report EPA-823-R-08-002.

United States Environmental Protection Agency (2008). Great Lakes beach sanitary survey user's manual. Technical Report EPA-823-B-06-001.

United States Environmental Protection Agency (2012). Recreational water quality criteria. Technical Report EPA-820-F-12-058.

United States Geological Survey (2014a). Environmental Data Discovery and Transformation.

United States Geological Survey (2014b). The National Water Information System.

Wade, T. J., R. L. Calderon, K. P. Brenner, E. Sams, M. Beach, R. Haugland, L. Wymer, and A. P. Dufour (2008). High sensitivity of children to swimming-associated gastrointestinal illness: results using a rapid assay of recreational water quality. *Epidemiology 19*(3), 375–383.

Wade, T. J., R. L. Calderon, E. Sams, M. Beach, K. P. Brenner, A. H. Williams, and A. P. Dufour (2006). Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness. *Environmental Health Perspectives 114*(1), 24–28.

Waschbusch, R., S. Corsi, K. Sorsa, J. Walker, J. Standridge, and T. Schnieder (2004). Final report for the EMPACT project: data collection and modeling of enteric pathogens, fecal indicators and real-time environmental data at Madison, Wisconsin recreational beaches for timely public access to water quality information. Technical Report R-82933901-0, Madison Beach EMPACT Team.

Whitman, R. L. and M. B. Nevers (2004). Escherichia coli sampling reliability at a

frequently closed Chicago beach: Monitoring and management implications. *Environmental Science & Technology 38*(16), 4241–4246.

Whitman, R. L. and M. B. Nevers (2008). Summer *E. coli* patterns and responses along 23 Chicago beaches. *Environmental Science & Technology 42*(24), 9217–9224.

Whitman, R. L., M. B. Nevers, G. C. Korinek, and M. N. Byappanahalli (2004). Solar and temporal effects on Escherichia coli concentration at a Lake Michigan swimming beach. *Applied and Environmental Microbiology 70*(7), 4276–4285.

Wisconsin Department of Natural Resources (2012). Wisconsin's Great Lakes beach monitoring and notification program. Technical report.

Wold, S., M. Sjostrom, and L. Eriksson (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems 58*(2), 109–130.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.