A Bayesian Approach to Recreational Water Quality Model Validation and Comparison in the Presence of Measurement Error

Eric Potash[a†] Scott Steinschneider[b]

April 30, 2020

Abstract

Methods for measuring recreational water quality vary in analysis time, precision, availability, and cost. Decision-makers often use predictions from statistical models to compensate for the shortcomings of available measurements. However, these models and analyses of their performance have largely omitted an important source of uncertainty: measurement error. This has led to inefficient models and misestimation of their performance. In this study we develop two methods to account for measurement error in cross-validation, a non-parametric method applicable to certain models and performance metrics, and a more general Bayesian simulation method. To that end, we develop a new Bayesian multivariate, multilevel, measurement error model of fecal indicator bacteria at 19 recreational beaches in Chicago. We find that estimates of prediction performance change substantially when accounting for measurement error. We also find that when used to make predictions the Bayesian model is expected to outperform past models with the greatest improvement under limited sampling resources.

Keywords: measurement error; Bayesian; multilevel; cross validation; recreational water quality; fecal indicator bacteria

Abbreviations: Fecal indicator bacteria (FIB), quantitative polymerase chain reaction (qPCR), cell equivalents (CE), mean squared error (MSE), mean absolute error (MAE), area under receiver operating characteristic curve (AUC), credible interval (CI).

[a]University of Chicago, 1307 E 60th St, Chicago, IL 60637, USA
[†]Corresponding author epotash@uchicago.edu
[b]Cornell University, 111 Wing Dr, Ithaca, NY 14853, USA

1 Introduction

Recreational waterways are subject to contamination by bacteria from various sources including stormwater, sewage, and wildlife (Whitman and Nevers 2008). To mitigate the public's exposure to contaminated water and associated gastrointestinal illness (Prüss 1998), managers of recreational beaches monitor the presence of fecal indicator bacteria (FIB) as a proxy measure of contamination. Managers issue warnings or close sites based on this information. There is a trade off between the protective public health benefits of these actions and the recreational benefits of access to waterways (Rabinovici et al. 2004). A major challenge in this decision process is how to appropriately account for measurement error in FIB data, which can be substantial (Whitman and Nevers 2004; Whitman, Ge, et al. 2010).

Measurement error has long been recognized as a major issue in water resources management, and the literature is rich with methods to incorporate measurement uncertainty in modeling and decision analysis. For example, in hydrology, Bayesian rainfall-runoff models have been developed to account for significant measurement error in catchment-scale precipitation to support improved parameter inference, predictive uncertainty bounds, and structural error diagnostics (Kuczera et al. 2006; Vrugt et al. 2008; Renard et al. 2011). Similar methods have also been extended to urban stormwater models to propagate bias and variance in both input (e.g. rainfall) and calibration (e.g., stormwater quality) data through the model fitting process (Dotto et al. 2014). Accommodations for measurement error have also been incorporated into decision-making processes, for instance with respect to groundwater remediation. For example, Liu et al. 2012 used a value-of-information approach to estimate remediation cost reductions afforded by reduced model, parameter, and measurement uncertainty. Likewise, Leube, Geiges, and Nowak 2012 used Bayesian methods to consider the effect of integrated groundwater modeling uncertainties (including measurement error) on optimal sampling design.

Measurement error has also played a prominent role in recreational water quality analysis. Modeling in this literature is often oriented towards decision support, where model-based predictions of FIB concentrations (including estimated moments or percentiles of measured data) are compared to water quality standards to guide management actions. A significant body of work has considered the impacts of measurement error on these decisions. For instance, several studies have used Bayesian analyses to explore the potential of concentration-based FIB standards that account for measurement error in indirect FIB concentration proxy measures (Gronewold, Borsuk, et al. 2008; Gronewold and Borsuk 2010; Gronewold, Sobsey, and McMahan 2017). A similar approach was used to show that a significant fraction of space-time variability in FIB proxy measures is driven by measurement error and not underlying variability of in-situ FIB concentrations

(Gronewold, Stow, et al. 2013).

When trying to improve water quality management decisions in the presence of model structural uncertainty, it is also common to compare the predictive performance of multiple FIB concentration models. In this facet of recreational water quality modeling, however, measurement error has been given less attention. When comparing predictive models using cross-validation, we found that researchers often ignored measurement error, simply assuming that a measurement (or mean of multiple measurements) represented the true bacteria level at the time the sample was taken (Nevers and Whitman 2011; Shively et al. 2016; Lucius et al. 2019). This is true even in studies that consider measurement error in the model estimation process (e.g., see figure 5 and associated discussion in Gronewold, Myers, et al. 2011). The omission of measurement error thus distorts a comparison of prediction performance across models, although the magnitude of this effect is unknown.

Given the methodological gap above, this study contributes two ways to account for measurement error when evaluating and comparing the performance of prediction models. The first is a non-parametric method that makes minimal assumptions but is limited in its applicability to models which are fit using fewer measurements than are recorded in the data and to a single metric of model performance, namely mean squared error (MSE). The second method is a simulation method that uses posterior samples from a Bayesian measurement error model. This method has the advantage of being applicable to any model regardless of its measurement inputs and any metric of model performance including those assessing the utility of predictions for decision-making around management-relevant FIB thresholds.

These methods are generally applicable to any inter-model comparison, and are thus relevant across a range of modeling exercises in water quantity and quality analysis. However, they are particularly relevant to recreational water quality modeling, given the common task of comparing multiple FIB concentration models for decision support and the high degree of measurement error in these data. We thus demonstrate the approach in a case study of 19 recreational beaches in Chicago, which has been used extensively to compare statistical models that aid in estimation of bacteria levels (Nevers and Whitman 2011; Shively et al. 2016; Lucius et al. 2019). In the process, we develop a new multi-level, multivariate Bayesian model for FIB concentration prediction that outperforms state-of-the-art models.

The remainder of the paper proceeds as follows. Section 2 describes the data and case study used to assess the proposed methods for inter-model comparison. Methodological details are presented in section 3, including the models (estimators) being compared and proposed approaches to cross-validation in the presence of measurement error. Results are presented in section 4. Finally we discuss limitations, potential avenues for future work, and implications in section 5.

2 Study Site and Data

The city of Chicago has 23 beaches along approximately 42 km of the Southwest shoreline of Lake Michigan. Of these, 19 beaches (figure 1) are currently subject to FIB monitoring during the swimming season from late May to early September. The beaches receive about 20 million visits during this period each year (Nevers and Whitman 2011).

Traditionally, administrators have relied on two culture measurements of *E. coli* per site to make management decisions. This method takes at least 12-24 hours. Because water quality can change rapidly, decisions based on measurements that are subject to such delays are likely to result in unnecessary closures as well as exposure (Kinzelman et al. 2003). Predictive models using these delayed measurements together with covariates such as rainfall, temperature, and sunlight have been used in Chicago to improve predictions of current bacteria levels (Shively et al. 2016).

Starting in 2015 and initially limited to five of the most contaminated beaches, quantitative polymerase chain reaction (qPCR) measurements of *Enterococci* have been employed. This method can quantify indicator bacteria in less than 2 hours (Noble et al. 2010). Adoption of qPCR methods has been limited by their increased cost and limited availability (Whitman, Ge, et al. 2010).

In 2017, however, administrators in Chicago switched completely to qPCR, making two such measurements at each of the 19 beaches. Thus the data for this study consists of two years (2015-2016) of qPCR measurements at 5 beaches and 3 years (2017-2019) of qPCR measurements at all 19 beaches. These data can be retrieved from the Chicago Data Portal. In addition, this study employs daily meteorological and hydrological covariates collected between 2015-2019 for the months of May-September. Site-specific rainfall, cloud cover, and wind speed data are collected from Dark Sky. Lake Michigan water levels at Calumet Harbor are taken from NOAA.

Figure 1: Map of the 19 recreational beaches on Lake Michigan in Chicago whose water quality is sampled. Diamonds indicate ten beaches to be sampled under the proposed targeted sampling design of Lucius et al. 2019 described in section 3.2.3.

3 Methods

Management decisions for each beach in Chicago are currently made by estimating FIB levels using the (geometric) mean of the samples at that site. We dub this the *empirical model*. Due to the cost of these measurements, the city has also proposed reducing sampling to ten of the sites and using a random forest model to predict levels at the remaining sites (Lucius et al. 2019). In this study we develop a third, Bayesian model and then (re-)evaluate the estimates and management consequences of all three models using our proposed cross-validation methods that account for measurement error.

In the sections below, we first introduce our novel, hierarchical Bayesian model for estimating the true FIB level across beaches that facilitates partial pooling of information across sites to improve the efficiency of estimation with limited and error-prone observations. We then introduce a set of FIB estimators based on the proposed Bayesian model and the existing empirical and random forest models that need to be compared. Finally, we present the proposed methods for inter-model comparison and cross-validation in the presence of measurement error.

3.1 Bayesian model

We denote the true level of *Enterococci* cell equivalents (CE) per mL on the (natural) log scale by θ_{jt} with $j = 1 \dots J$ a site index and $t = 1 \dots T$ a day index.

Next we propose a linear regression (on the log scale) model for these states with coefficients varying by site:

$$\theta_{jt} = X_{jt}\beta_j + \epsilon_{jt} \tag{1}$$

where X_{jt} is a vector of $K = 15$ covariates (including an intercept) and β_j is a vector of K coefficient parameters.

The covariates X_{jt} are listed in table 1 and mirror those of (Lucius et al. 2019) with minor changes. First, we excluded forecasts of future meteorological conditions based on a prior belief that, conditional on past conditions, current bacteria levels are independent of future conditions. Second, since our model is linear as opposed to their non-linear random forest, we parameterized day of year as a B-spline with 4 degrees of freedom and added a separate wind speed for each cardinal direction. Finally, we reparameterized aggregated covariates to reduce their correlation to speed up model fitting and improve interpretability. For example we included 2-3 day total rainfall instead of their 1-3 day total rainfall as the former is less correlated with 1 day total rainfall.

Category	Covariate
Precipitation	1 day total rainfall
	2-3 day total rainfall
	1-2 day change in water level
Sunlight	1 day average cloud cover
	2-3 day average cloud cover
Wind	1 day average North wind speed
	1 day average South wind speed
	1 day average East wind speed
	1 day average West wind speed
Temporal	Day of year B-spline
	Weekday indicator

Table 1: Bayesian model covariates

While our model is similar to previous regression models of bacteria levels, we add three innovations. First we use a multivariate normal error distribution with covariance matrix Σ to model correlation in the errors across beaches on a given day t:

$$\epsilon_t \sim \text{Normal}(0, \Sigma) \tag{2}$$

This enables us to combine the measurements at other beaches with those at a given beach in estimating the bacteria level at that beach.

Second we add a multilevel structure on the coefficients, that is we have the second-level model:

$$\beta_{jk} \sim \text{Normal}(\mu_{\beta_k}, \sigma^2_{\beta_k}) \tag{3}$$

This allows us to partially pool information across beaches to more efficiently estimate the coefficients at a given beach (Stow et al. 2009; Cha et al. 2010).

Our final innovation is to include an additive and normally distributed measurement error with standard deviation τ (Gronewold, Qian, et al. 2009). That is, if y_{ijt} is an observation at beach j on day t then it is normally distributed with mean θ_{jt} and variance τ^2:

$$y_{ijt} \sim \text{Normal}(\theta_{jt}, \tau^2) \tag{4}$$

We put the following uninformative priors on these parameters (Gelman et al. 2013). Decomposing Σ into a

correlation matrix Ω and a vector of coefficient scales σ

$$\Sigma = \text{diag}(\sigma) \cdot \Omega \cdot \text{diag}(\sigma) \tag{5}$$

we put a uniform prior over Ω and a $\text{Cauchy}_+(0, 1)$ prior on the components of σ. The mean and variance hyperparameters μ_{β_k} and $\sigma^2_{\beta_k}$ are given uninformative $\text{Cauchy}(0, 1)$ and $\text{Cauchy}_+(0, 1)$ priors, respectively. All priors are defined after standardizing all predictors and the outcome.

We denote by Ψ the collection of model parameters (Σ, β, τ). We fit the model using the Markov Chain Monte Carlo software Stan (Carpenter et al. 2017), which uses No-U-Turn sampling (Hoffman and Gelman 2014), an extension Hamiltonian Monte Carlo (Duane et al. 1987). We generated 4 chains with 1000 iterations each, saving the last 500 to produce $N = 2000$ draws from the joint posterior parameter distribution. We assessed mixing using the criteria $\hat{R} < 1.05$ and $n_{\text{eff}}/N > .001$ where $\hat{R}$ is the Gelman-Rubin convergence statistic and n_{eff} is the effective sample size (Gelman et al. 2013).

3.2 Estimators

Various estimates of the *Enterococci* CE θ_{jt} at site j on day t have been proposed. Here we describe our proposed Bayesian estimator based on the model in 3.1, as well as two other estimators based on a baseline, empirical approach and a recently published random forest model.

In all cases the input to an estimator on day t is a vector Y_t of I FIB observations, where I is the total number of samples per day, and possibly a vector of K covariates X_t. Here, I can vary depending on the sampling design, i.e., the number of samples taken at each site each day. In considering various estimators, we include the sampling design as part of the estimator. Thus the same type of estimator (i.e. empirical, Bayesian, random forest) under different sampling designs are considered different estimators.

Formally we write a sampling design as a matrix F of dimension $I \times J$. If the i^{th} sample is at site j then $F_{ij} = 1$ and the remaining entries are zero. Thus if Y_t is a vector of observations with design F then

$$\mathbb{E}[Y_t] = F\theta_t. \tag{6}$$

Here we consider three sampling designs: one sample per site (19 samples total), two samples per site (38 samples total), and the targeted design of Lucius et al. (2019) (described in section 3.2.3 below) which draws two samples from each of ten sites (20 samples total). These sampling designs are selectively paired with the three estimator types to develop the final estimators considered in this study (table 2). Because there are two

	19-20 samples total		38 samples total
Model type	1 sample per site	Targeted sampling	2 samples per site
Empirical	Empirical[1]		Empirical[2]
Bayesian	Bayesian[1]		Bayesian[2]
Random Forest		Random Forest[1]	

Table 2: Estimators by model type and sampling design. Estimator superscripts indicate (approximate) average number of samples per site.

samples per site in our data (see section 2), estimators which use one sample per site do not use all available data, thereby creating a hold-out set which can be exploited for cross-validation. For those estimators that use two samples per site, no hold-out set can be created, but certain methods for cross-validation are still possible (as explained in section 3.3).

3.2.1 Empirical estimator

The simplest estimator, which is currently being used in Chicago, is what we call the empirical estimator. When there is one sample y_{jt} per site, the empirical estimate is equal to the observation: $\hat{\theta}^{\text{emp}}_{jt} = y_{jt}$. When the sampling design F includes more than one sample at a site, the empirical estimate is the site-specific (geometric) mean, which can be written in matrix notation as

$$\hat{\theta}^{\text{emp}}(Y_t) = (FF')^{-1}F'Y. \tag{7}$$

3.2.2 Bayesian estimator

Our Bayesian estimator of θ_t is the posterior mean of θ given measurements Y_t and covariates X_t:

$$\hat{\theta}^{\text{bayes}}(Y_t, X_t) = \mathbb{E}[\theta_t | Y_t, X_t]. \tag{8}$$

Note that given the design matrix F we can rewrite equation 4 as

$$Y_t \sim \text{Normal}(F\theta_t, \mathbb{1}_I \tau^2) \tag{9}$$

where $\mathbb{1}_I$ is the identity matrix of dimension I. For each posterior sample of parameters Ψ, the posterior distribution $\theta_t | Y_t, X_t, \Psi$ is normal with a closed form mean and variance (e.g. Eaton, Giovagnoli, and Sebastiani 1996, lemma 3.1).

3.2.3 Random forest estimator

Lucius et al. (2019) proposed a "hybrid nowcast model" which, like our Bayesian estimator, takes advantage of the correlation of bacteria levels across beaches. This is done by first using a targeted sampling design that takes two measurements at each of 10 beaches. At these sampled sites empirical estimates are used. Next, to estimate FIB levels at the unsampled sites, a random forest regression model with 400 trees was used (Breiman 2001). The outcomes used to fit the model were the empirically estimated levels at these unsampled sites. The inputs to the model were the empirical estimates at the sampled sites together with covariates.

Formally the random forest estimate $\hat{\theta}^{\text{rf}}$ can be written as

$$\hat{\theta}^{\text{rf}}(Y_t, X_t)_j = \begin{cases} \hat{\theta}^{\text{emp}}(Y_t)_j, & \text{if } j \in \mathcal{J} \\ f(\hat{\theta}^{\text{emp}}(Y_t), X_t), & \text{otherwise} \end{cases} \tag{10}$$

where $\mathcal{J} \subset \{1, \ldots, J\}$ is the set of ten sampled sites, Y_t is the vector of 20 measurements at these sites, $\hat{\theta}_t^{\text{emp}}$ is the vector of corresponding empirical estimates at these sites, X_t are covariates, and f is the random forest regression function. For this study we refit the random forest using the covariates in table 1 and our training set, which is larger than that of the original publication.

The ten sampled beaches (see figure 1) were chosen by Lucius et al. (2019) as follows. First, five beaches were selected to be sampled due to their historically high FIB levels. Next, the remaining beaches were grouped into five geographic clusters and a single beach was selected to be sampled from each cluster. Since two samples are taken at each of the ten selected beaches, this estimator uses 20 samples. In our analysis, we compare this estimator against the Bayesian and empirical estimators using one sample at site, i.e. 19 samples.

Because the random forest estimator relies on empirical estimates for ten of the sites, its performance is closely linked to that of the empirical estimator.

3.2.4 Exceedance predictions

We have presented estimators of continuous FIB levels, but Environmental Protection Agency guidance suggests making management decisions based on the binary event of exceeding 1000 CE (United States Environmental Protection Agency 2012). For this an estimator's continuous predictions must be transformed into binary predictions. This is done using a threshold decision rule, i.e. $D(\hat{\theta}) = \hat{\theta} > C$, where the threshold C may depend on the estimator.

Regarding the choice of threshold, decisions currently made in Chicago using the empirical estimator with two samples per site simply use the allowance of 1000 CE as the threshold.

For their random forest, Lucius et al. 2019 calibrated the threshold to match the specificity of a reference model (Shively et al. 2016). That is, they chose C such that the resulting specificity of their predictions (equivalently the false positive rate) would match that of their reference model. We follow this approach, taking the current empirical estimator using two samples per site as our reference and calibrating thresholds for all other estimators to match its specificity.

The Bayesian model is richer than the others and so when predicting exceedance we replace the posterior expectation estimator above with the posterior probability of exceedance:

$$\hat{\theta}_t^{\text{bayes}} = \mathbb{E}[\theta_t > \log(1000) | Y_t, X_t]. \tag{11}$$

3.3 Cross-validation

The primary purpose of this study is to develop an approach to compare the performance of multiple models (i.e., the various estimators above) in the presence of measurement error. In cross-validation we evaluate the fidelity of estimated states $\hat{\theta}(Y)$ to the true state θ by a function $L(\theta, \hat{\theta}(Y))$. Here Y is an observation vector input to the estimator under evaluation and L is one of various performance metrics (e.g. MSE). We include days t in the *test period* which we chose to be the most recent beach season, 2019. For the random forest and Bayesian estimators, model parameters were fit using data from the *training period*, i.e. prior to 2019; the empirical estimator does not require fitted parameters and thus only uses data from the 2019 test period.

The challenge in cross-validation is that we never observe θ_t. In the literature, θ_t is often assumed to be exactly equal to the empirical estimate $\hat{\theta}_t^{\text{emp}}$ using all available samples (Nevers and Whitman 2011; Shively et al. 2016; Lucius et al. 2019). However, this does not account for measurement uncertainty. For example, this would imply that predictions of the empirical estimator using all available samples are perfect.[1] We term this method of cross-validation *naive*, and propose two additional methods: *non-parametric* and *simulated*. The methods are described below and summarized in figure 2.

3.3.1 Naive validation

In naive validation we simply assume that the empirical estimate using all available samples is true: $\theta_t = \hat{\theta}_t^{\text{emp}}(Y_t^{\text{obs}})$.

[1] For estimators using prior-day culture measurements (Nevers and Whitman 2011; Shively et al. 2016), the issue is less severe but remains.

In order to evaluate the metric $L(\theta, \hat{\theta}(Y))$ we also need to produce Y. When the estimator $\hat{\theta}$ being evaluated uses all available observations (i.e. 2 per site), we let $Y = Y^{\text{obs}}$ and for any performance metric L we have a single number $L(\theta, \hat{\theta}(Y))$.

When the estimator being evaluated uses one sample per site (table 2), we produce Y by subsampling one of the two observations from each site on each day. We sample among possible subsamples, resulting in a distribution for $L(\theta, \hat{\theta}(Y))$. Note that when evaluating the random forest estimator which uses two samples per site at ten sites, subsampling is not necessary (or possible).

Thus in this case we have (by necessity) augmented the conventional naive validation procedure and incorporated into $L(\theta, \hat{\theta}(Y))$ uncertainty in the measurements Y. However this does not account for uncertainty in the true state θ.

As mentioned above this naive validation method is flawed as can be seen from the fact that when it is used to evaluate the empirical estimator that uses two samples per site, the predictions are exactly equal to the "truth" ($\theta = \hat{\theta}$) and so are considered perfect by any L. We expect this naive validation to over-estimate the performance of any other estimator which does not account for measurement error since these will produce estimates which are correlated with the measurement error in the empirically estimated true state. This includes the empirical estimator using one sample per site as well as the random forest estimator which implicitly uses empirical estimates for the ten sites which its design samples (equation 10).[2]

3.3.2 Non-parametric validation

Consider the special case of estimating MSE of the empirical estimator using one sample per site. Assume a measurement error model

$$y_{ijt} = \theta_{jt} + \nu_{ijt} \tag{12}$$

where ν_{ijt} are independent measurement errors whose distribution need not be known. Then expected MSE on beach-day jt is simply $\text{Var}(\nu_{ijt})$. So to estimate MSE in this case is equivalent to estimating the measurement error variance. Since we have two measurements per beach-day in the data, this can be done using the standard sample variance estimator and then averaged across beaches and days.

Can this result be extended to include other estimators? For simplicity assume that there are two observations per site in the data and the estimator $\hat{\theta}$ uses only one of them, as in the one sample per site sampling design (table 2).[3] Next subsample Y_t accordingly (as in naive validation above) and denote the remaining or *hold-out*

[2] We do not expect naive validation to be biased for estimators based on prior day *E. coli* measurements because in this case the measurement error in the prior and current day observations are independent. However, naive validation still has the disadvantage here of not reporting uncertainty.

[3] In fact these results can be generalized to whenever there are more observations available than used by the estimator.

observations by Y_t^{hold}. In appendix A we derive the following estimate of the MSE:

$$\frac{1}{T}\sum_t |\hat{\theta}(Y_t) - Y_t^{\text{hold}}|^2 - \frac{1}{2}|Y_t - Y_t^{\text{hold}}|^2 \tag{13}$$

which we can bootstrap across t to estimate the sampling distribution.

Note that this result does not make any distributional assumptions. We only assumed that the observations are equal to the true state plus independent measurement error (equation 12). However, this approach only applies to estimators using strictly fewer observations of each state θ_{jt} than are available. Thus we only use non-parametric validation to estimate the empirical and Bayesian estimators with one sample per site. It is also limited to the specific error metric of MSE.[4]

3.3.3 Simulated validation

The approaches to cross-validation presented above either: assume the true FIB level θ_t is equal to the geometric mean of available observations (naive); or estimate a specific error metric, MSE, under a specific sampling design (non-parametric). An alternative and more general approach is a kind of multiple imputation using simulations from our Bayesian model (Rubin 2004).

There are two simulation steps. First, we simulate the true states θ from the posterior distribution $\theta|Y^{\text{obs}}, X$. Next we simulate measurements $Y|\theta$ according to the design F of whatever estimator $\hat{\theta}$ we are evaluating. For each draw (θ, Y) we can evaluate $L(\theta, \hat{\theta}(Y))$. By repeatedly drawing these quantities we thus estimate a posterior distribution for $L(\theta, \hat{\theta}(Y))$. Note that whereas parameters for the Bayesian estimator (section 3.2.2) are fit using only training data, parameters used for simulation of the true states and measurements are fit using all of the data, including the test set.

The simulated cross-validation method assumes that the Bayesian model is correct. Thus one may suspect that using this method to validate predictions from the same model will lead to (optimistically) biased performance estimates. However note that the simulations incorporate uncertainty in the Bayesian model parameters, and the simulations are made with the Bayesian model fit to all data while predictions are made by a model fit only to past data. Thus the simulation and prediction models are not strictly the same.

The simulated validation method has two advantages. First, unlike the non-parametric validation above which can only estimate MSE, simulation can be used to estimate any prediction performance metric. We

[4] One might try using the bootstrap to estimate the distribution of θ_{jt} which would provide a non-parametric approach to cross-validation for any metric or estimator. But with only two samples of each θ_{jt} this approach is not viable here.

consider several (table 3), including MSE and mean absolute error (MAE) to evaluate predictions of the continuous FIB level and other metrics to evaluate predictions of binary exceedance (section 3.2.4). The area under the receiver operating curve (AUC) evaluates these exceedance predictions as continuous scores. The remaining metrics use binary classifications which are obtained from scores using a threshold as described in section 3.2.4: precision measures the proportion of exceedance predictions which are correct; sensitivity measures the proportion of exceedances which are correctly predicted; accuracy measures overall performance as the proportion of all predictions which are correct.

Second, the simulation approach can be used in scenarios where there are not enough remaining observations to form the hold-out vector needed in the non-parametric approach, such as the random forest and the empirical and Bayesian estimators using two samples per site. Moreover, simulated validation could be used to evaluate estimators using *more* observations than are available in the data by simply simulating them. This could be used for example to estimate the prediction performance benefits of increasing the number of samples per site in Chicago to three.

	Cross-validation method		
Estimator	Naive	Non-parametric	Simulated
Empirical[1]	✓	✓	✓
Random Forest[1]	✓		✓
Bayesian[1]	✓	✓	✓
Empirical[2]	✓		✓
Bayesian[2]	✓		✓

(a)

	Cross-validation method		
Metric	Naive	Non-parametric	Simulated
Mean squared error	✓	✓	✓
Mean absolute error	✓		✓
AUC	✓		✓
Accuracy	✓		✓
Precision	✓		✓
Sensitivity	✓		✓

(b)

Table 3: Applicability of cross-validation methods to estimators (a) and prediction performance metrics (b). Estimator superscripts indicate average number of samples per site. Abbreviations: area under receiver operating characteristic curve (AUC).

Input: Test set observations Y^{obs}, covariates X
Estimator $\hat{\theta}$ with sampling design F
Performance metric L
Result: Estimate of $L(\hat{\theta}, \theta)$

Estimate θ using the empirical estimator $\theta = \hat{\theta}^{\text{emp}}(Y^{\text{obs}})$
Subsample Y from Y^{obs} according to F
Estimate $\hat{\theta} := \hat{\theta}(Y, X)$
Calculate $L(\hat{\theta}, \theta)$

(a) Naive cross-validation

Input: Test set observations Y^{obs}, covariates X
Estimator $\hat{\theta}$ with sampling design F
Result: Estimate of $\text{MSE}(\hat{\theta}, \theta)$

Bootstrap resample $\tilde{Y}^{\text{obs}}$ across t of Y^{obs}
Subsample $\tilde{Y}^{\text{obs}}$ to Y and Y^{hold} according to F
Estimate $\hat{\theta} := \hat{\theta}(Y, X)$
Calculate $\frac{1}{T}\sum_t |\hat{\theta}(Y_t) - Y_t^{\text{hold}}|^2 - \frac{1}{2}|Y_t - Y_t^{\text{hold}}|^2$

(b) Non-parametric cross-validation

Input: Test set observations Y^{obs}, covariates X
Estimator $\hat{\theta}$ with sampling design F
Performance metric L
Result: Estimate of $L(\hat{\theta}, \theta)$

Sample Bayesian model parameters Ψ from posterior
Sample $\theta | Y^{\text{obs}}, X, \Psi$
Sample Observations $Y|\theta$ according to F
Estimate $\hat{\theta} := \hat{\theta}(Y, X)$
Calculate $L(\hat{\theta}, \theta)$

(c) Simulated cross-validation

Figure 2: Cross-validation methods

4 Results

We first present details of the fitted Bayesian model. We then compare cross-validation procedures for specific sampling designs, and then compare estimators across all sampling designs using a subset of cross-validation procedures.

4.1 Bayesian model Fit

The Bayesian model was fit using qPCR measurements from the 2015 to 2018 seasons. There were 9329 such observations made at the 19 beaches on 328 days. A subset of these measurements at a cluster of beaches are shown in figure 3.

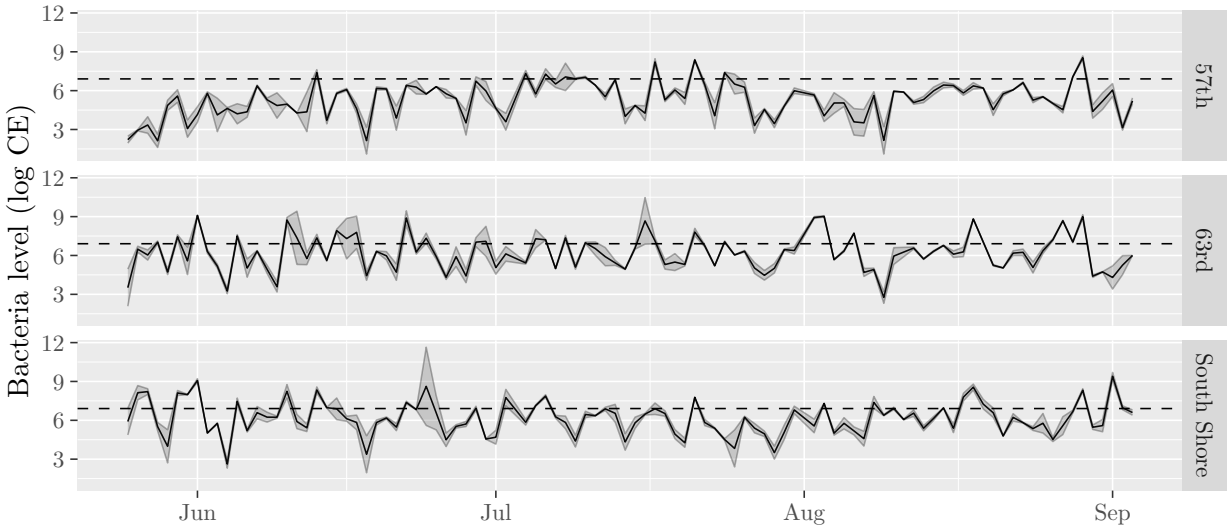

Figure 3: Fecal indicator bacteria measurements at a cluster of 3 of the 19 Chicago beaches during the 2018 beach season. Gray region spans minimum and maximum measurements, solid line connects daily means, i.e. empirical estimates currently used for decision making. Dashed lines indicate action threshold of 1000 cell equivalents (CE).

Our MCMC diagnostic criteria were satisfied and there were no divergent transitions. The posterior estimates of the multilevel regression coefficient means μ_{β_k} that relate covariates to FIB concentrations are summarized in figure 4. Rainfall and lake level coefficients are positive, consistent with stormwater causing combined sewage overflows that discharge into the lake (Olyphant and Whitman 2004). However, many of the covariates (precipitation, cloud cover, wind speed, etc.) are correlated so our interpretation of their coefficients is limited. For instance, posterior coefficient estimates on all wind speeds are positive, suggesting higher winds cause higher FIB concentrations. Yet this effect may just be an artifact of the correlation between rainfall and wind speeds during storm events. The day-of-year trend, which predicts (albeit imprecisely) increasing bacteria levels over the course of the season, may reflect warmer water temperatures as well as increased human traffic

at beaches.

The posterior distribution of measurement error variance τ^2 had median 0.77 (log CE)^2 (95% CI, 0.74 to 0.8). For comparison, this is 53% of the variance in beach-day means (i.e. empirical estimates) of 1.45 (log CE)^2. It also corresponds to a measurement error standard deviation of $\tau = .88$ log CE which is 17% of the median measurement of 5.1 log CE.

We separately fit the model to *E. coli* culture data and estimated a measurement error τ^2 to be 0.37 (95% CI, 0.34 to 0.4) which is consistent with the estimate .401 of Whitman and Nevers 2004 (their table 2). We conclude that, as previously suggested by Whitman, Ge, et al. (2010), measurement error is greater for qPCR than culture tests, albeit with respect to different units.

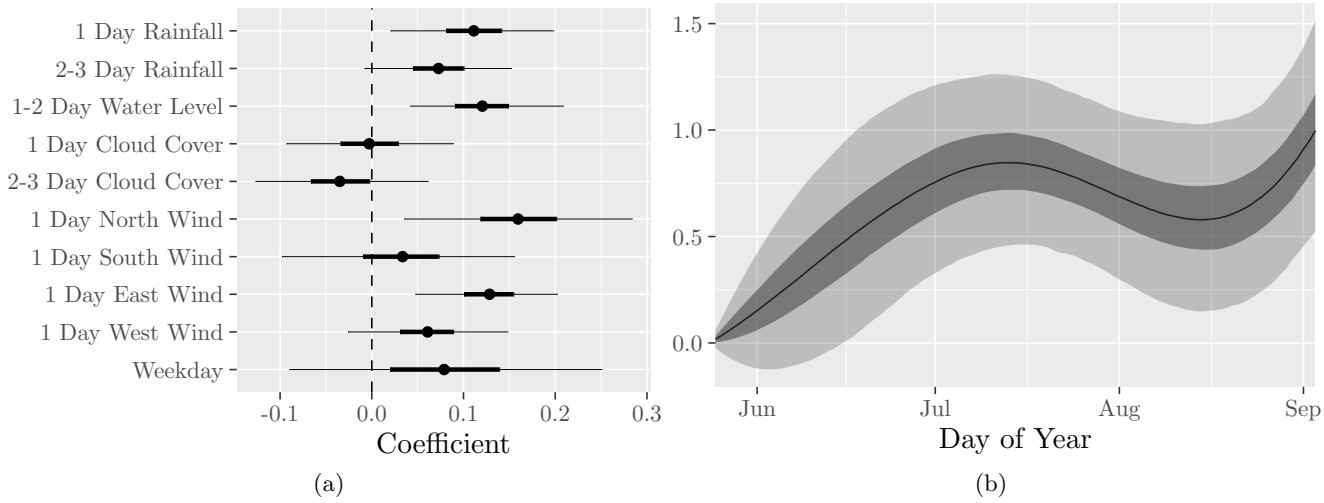

Figure 4: Bayesian model estimated (a) standardized coefficients $\mu_{\beta_k}/\text{sd}(x_k)$ and (b) day of year trend.

4.2 Cross-validation

During the 2019 season 3780 qPCR measurements were made over 102 days. We restricted the test period to those days with two samples at each of the 19 beaches so that all estimators could be evaluated. There were 67 such days.

According to empirical estimates (using both samples at each site site), the median level of indicator bacteria was 92 CE and 4.9% of these beach-days were in exceedance of the 1000 CE threshold. The Bayesian estimate of the median level was 93 CE (95% CI, 36 to 239 CE) and 4.0% (95% CI, 3.6% to 4.2%) of beach days exceeded the threshold.

4.2.1 Comparing cross-validation methods

We start by examining the three cross-validation methods on the estimators and metrics where they could all be compared, that is the empirical and Bayesian models with one sample per site and the MSE metric. Figure 5 shows the distribution of MSE values across the three methods.[5] Because the random forest estimator uses two samples per site at a subset of sites (and thus has no hold-out values), it cannot be evaluated under the non-parametric cross-validation method and is thus excluded from this analysis.

Two findings emerge from figure 5. First, while naive cross-validation shows the empirical estimator performs best, non-parametric cross-validation shows the Bayesian estimator performs best. We anticipated above that naive validation would overestimate the performance of the empirical estimator since the same measurement errors are unaccounted for in both the true state in naive validation and the predictions of the empirical estimator. Conversely, non-parametric validation accounts for measurement error, so we are inclined to trust its results (preference of the Bayesian model) and dismiss naive validation, which is a priori flawed. We re-emphasize here that the non-parametric approach to cross-validation made no distributional assumptions, and therefore does not unfairly favor the Bayesian model.

Our second finding is a remarkable agreement between non-parametric and simulated MSE estimates. Because non-parametric validation makes so few assumptions, this agreement provides strong evidence to support our use of the simulated validation method to further explore the performance of estimators and metrics for which we do not have a non-parametric method.

4.2.2 Comparing estimators

Next we use the simulated and naive cross-validation methods to evaluate all error metrics and estimators (i.e., all model types and sampling designs).[6] The results are shown in figure 6.

Using simulated cross-validation, the Bayesian model is expected to outperform both the empirical and random forest models at each level of sampling resources, with the greatest improvement in continuous predictions as measured by MSE and MAE. The degree of uncertainty varies across the performance metrics, with estimates of MAE the most certain and sensitivity the least. The quantification of this uncertainty is an asset of the simulated validation method.

The discrepancy between simulated and naive validation, first documented in section 4.2.1 above, continues here across more estimators and metrics. In most cases the naive estimates are more optimistic (e.g. lower

[5] In naive validation of the empirical estimator with one observation per site, uncertainty collapses to zero due to symmetry between the observation and hold-out.

[6] The empirical estimator with two samples per site, which uses a binary classification threshold of 1000 CE, had an expected specificity of 97.2%. Classification thresholds for all other estimators were calibrated to match this (see section 3.2.4).

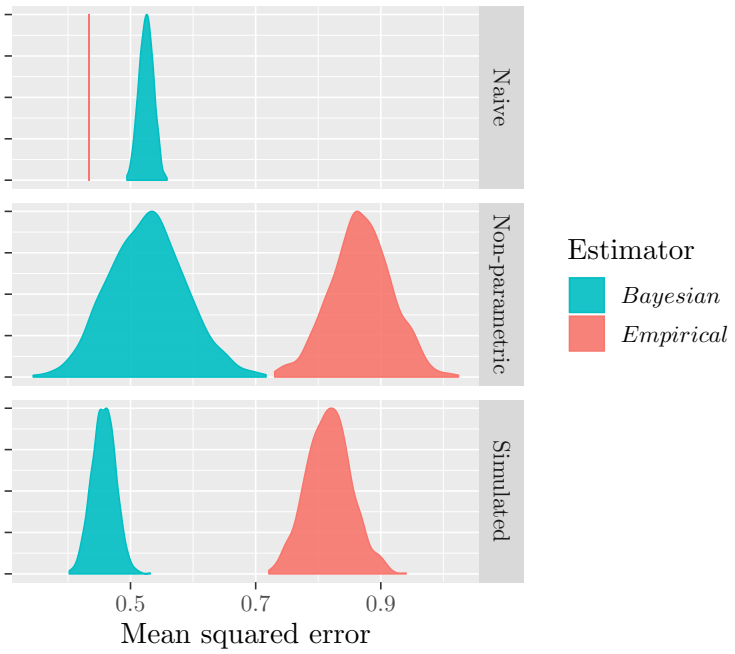

Figure 5: Mean squared error distributions estimated using naive, non-parametric, and simulated cross-validation methods for empirical and Bayesian estimators each using one sample per site.

MSE, higher precision, etc.), as we anticipated above. Figure 7 summarizes the optimistic nature of naive cross-validation, averaging the difference between naive and simulation-based prediction metrics across all five estimators. This is explained by the fact that the empirically estimated state θ in naive validation includes some of the same measurement error which is used by the estimators to make their predictions.

In figure 5 we saw that when moving from naive to non-parametric validation, the relative standing in terms of MSE of the empirical and Bayesian estimators using one sample per site switched. Using simulated validation we see a similar phenomenon for metrics besides MSE and additional estimators, including all three estimator types using two samples per site. This is also explained by the fact that naive validation erroneously favors empirical and random forest estimators which do not account for measurement error.

We used simulated cross-validation to estimate the difference in each performance metric between the Bayesian model with one sample per site and the currently used empirical model with two samples per site. Note that estimates of prediction performance using simulated validation for different estimators are jointly distributed. This is taken into account to estimate differences in performance. The results are displayed in figure 8.

The Bayesian estimator with one sample per site performs almost as well as the empirical estimator with two samples per site across metrics. This suggests that using the Bayesian estimator in 2019, beach administrators would have been able to cut sampling resources by half with minor changes in the fidelity of their estimates, and in turn the utility of their decisions based on those estimates.

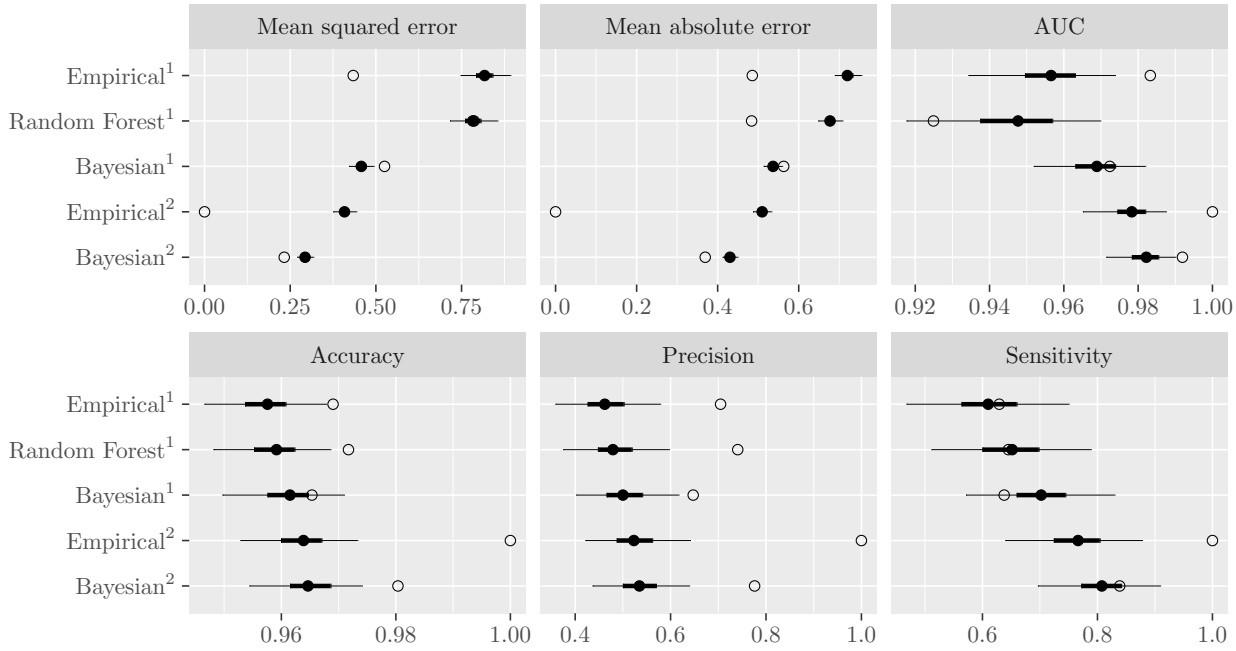

Figure 6: Prediction performance metrics. Solid dots and intervals show median and 50% and 95% credible intervals using simulation to account for measurement error. Open dots show naive estimates without accounting for measurement error. Estimator superscripts indicate average number of samples per site. Abbreviations: area under receiver operating characteristic curve (AUC).

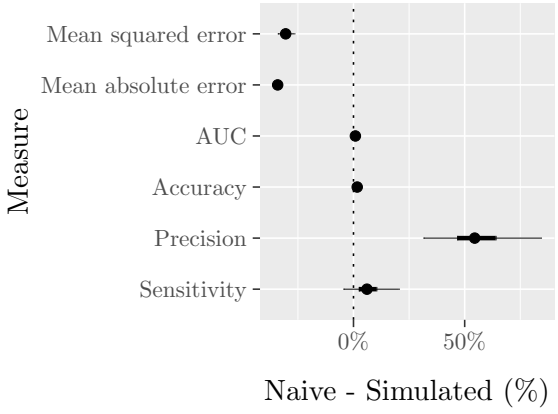

Figure 7: Comparison between naive and simulated estimates. Solid dots and intervals show median and 50% and 95% credible intervals for percentage difference between naive and simulated estimates of prediction performance averaged across the five estimators. Abbreviations: area under receiver operating characteristic curve (AUC).

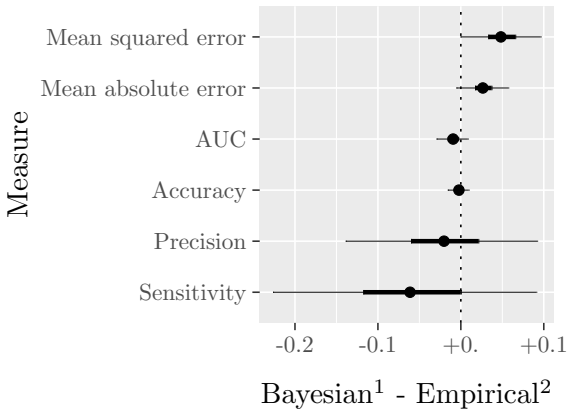

Figure 8: Relative prediction performance. Solid dots and intervals show median and 50% and 95% credible intervals for difference between the Bayesian model with one sample per site and the empirical model with two samples per site using simulation to account for measurement error. Abbreviations: area under receiver operating characteristic curve (AUC).

5 Discussion

5.1 Limitations and possible advances of the Bayesian model

One natural way to extend the Bayesian model is through autoregression. These models were considered but initial testing (not shown) confirmed previous findings that system dynamics are too fast for memory effects of daily samples to provide substantial explanatory power (Dorevitch et al. 2017). The parameters in our model are also stationary in time (though they may be updated by periodically refitting the model). If the true parameters are changing, the performance we estimated here may not be representative of future performance. This could be overcome to some extent by modeling the parameters as varying in time (Petris, Petrone, and Campagnoli 2009).

There is some evidence suggesting that measurement error may vary with the bacteria level (Whitman, Ge, et al. 2010). The model could be extended with a heterogeneous measurement error which varies with the bacteria level, as well as other factors such as turbidity.

The data for this study came from Chicago's existing water quality monitoring program. This program relies on two samples per beach per day with each pair collected at the same location and time. However it has been shown that there is substantial variation spatially within each beach and temporally within each day (Whitman and Nevers 2004). With the relevant data, our model could be extended to these finer scales.

One innovation of the proposal by Lucius et al. (2019) was to target the allocation of water samples. Our Bayesian model may also be used with a targeted sampling design. Moreover, compared to their random forest model, the Bayesian model is more flexible: it can be used with any sampling design, or combination of

designs, without refitting. However, we evaluated the Bayesian model under the targeted design of Lucius et al. 2019 and found that it underperformed the uniform sampling design (appendix figure B.1). It is possible that a different targeted design may be superior to the uniform one. On the other hand, uniform sampling ensures that measurements are made at every beach which would prove useful if and when a given model is invalidated by non-stationarity.

Finally, while our model may give better estimates of bacteria levels, to make decisions based on such estimates beach managers will need to incorporate additional information on the consequences of their decisions. This includes information on the effects of human exposure to elevated bacteria levels as well the effects of mitigating actions such as swimming advisories and beach closures.

5.2 Implications for cross-validation and inter-model comparison

In this study we showed that the relatively common omission of measurement error from evaluations of predictive models of FIB levels for recreational water quality (Nevers and Whitman 2011; Shively et al. 2016; Lucius et al. 2019) can have a substantial effect on estimates of predictive performance. We saw that a naive approach to cross-validation fails to quantify the uncertainty of results and can substantially overestimate performance, favoring models that themselves do not account for measurement error.

The omission of measurement error in cross-validation is ubiquitous in the water resources literature (e.g. Dawson and Wilby 2001; Berenguer et al. 2005; Biondi et al. 2012; Lohani, Kumar, and Singh 2012; Shortridge, Guikema, and Zaitchik 2016). In instances when measurement error is small, the effects on cross-validation and inter-model comparison are also likely to be small. However, for applications like recreational water quality modeling where measurement error is substantial, we have seen that the effects can be large and so recommend that future evaluations use the methods we have proposed to account for this source of error in model evaluation and comparison.

Using these methods we produced the first evaluation of the performance of the empirical estimator using two samples per site that is currently used by administrators in Chicago. Such estimates are essential to understanding the public health consequences of management decisions made on their basis. We found that with a single sample per site, model-based estimates were a substantial improvement over empirical estimates. The Bayesian estimator was best, with performance approaching that of the empirical estimator that uses twice as many sampling resources. With an estimated processing cost of $30 per qPCR sample (Griffith and Weisberg 2011), we estimate that processing of two samples per beach-day at 19 beaches and 100 days per season to cost about $114k per season. Switching to the Bayesian estimator using a single sample per site

could therefore reduce costs by about $57k per season with minimal changes in the utility of management decisions. Outside of Chicago there are many other locations where FIB monitoring resources are more scarce or non-existent. Future work should apply our model to those locations.

6 Conclusions

In this study we showed that the conventional approach for validating and comparing FIB prediction models is naive in taking observed levels as truth and not accounting for measurement error. We made two contributions to this literature. First, we proposed new methods for performing cross-validation and inter-model comparison while accounting for measurement uncertainty: a non-parametric approach for MSE and a parametric approach for any performance metric using simulation from a Bayesian model.

Second, to implement the parametric method we developed a Bayesian model of FIB levels at 19 recreational beaches in Chicago with the following innovations: a state space approach to explicitly model measurement error; a multilevel structure to efficiently estimate regression coefficients that varied across beaches; a multivariate normal error distribution whose estimated covariance structure enabled incorporating measurements at other beaches in estimating FIB levels at a given beach.

We used the non-parametric cross-validation approach to validate the parametric, simulation-based approach. We found that, across several estimators and metrics, estimates of performance using the naive method differed substantially from estimates using the non-parametric and simulation methods that account for measurement error. Moreover, we showed that the Bayesian model, when used to make predictions, outperformed empirical and recent machine-learning based estimators under the proposed cross-validation strategies when using comparable numbers of samples. Overall, the results of this work suggest that future studies conducting inter-model comparison should consider alternative methods for cross-validation that better account for measurement error, especially in the recreational water quality literature where measurement error is significant.

Acknowledgements

Thanks to Dan Black, Steven Durlauf, Jeff Johnston, Andrew Gelman, Jim Savage, Jackie Shadlen, and Rob Trangucci for useful conversations. Thanks also to Lucius et al. (2019) for transparency about their model and data and assistance in replicating their results.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

A Proofs

Lemma 1. *Let $Y_t = \theta_t + \nu$ and $Y_t^{hold} = \theta_t + \nu^{hold}$ where ν, ν^{hold} are independent vectors of length J with mean 0 and variance τ^2 in each component. Then*

$$\mathbb{E}\left[|\hat{\theta}(Y_t) - Y_t^{hold}|^2\right] = \mathbb{E}\left[|\hat{\theta}(Y_t) - \theta|^2\right] + J\tau^2 \tag{14}$$

Proof. We have

$$\begin{aligned}|\hat{\theta}(Y_t) - Y_t^{\text{hold}}|^2 &= |\hat{\theta}(Y_t) - \theta_t - \nu^{\text{hold}}|^2 \\ &= |\hat{\theta}(Y_t) - \theta_t|^2 + |\nu|^2 - 2\nu^{\text{hold}} \cdot (\hat{\theta}(Y_t) - \theta_t).\end{aligned}$$

Next we take expectation over ν, ν^{hold}. The second term becomes $\sum_j \text{Var}[\nu_j^{\text{hold}}] = J\tau^2$ and independence of ν and ν^{hold} implies the third term is zero. □

In particular when using the empirical estimator, $\hat{\theta}^{\text{emp}}(Y_t) = Y_t$ we have

Corollary 2.

$$\mathbb{E}\left[|Y_t - Y_t^{hold}|^2\right] = 2J\tau^2 \tag{15}$$

Combining Lemma 1 and Corollary 2 we have

Corollary 3.

$$MSE(\hat{\theta}(Y_t), \theta_t) = \mathbb{E}\left[|\hat{\theta}(Y_t) - Y_t^{hold}|^2 - \frac{1}{2}|Y_t - Y_t^{hold}|^2\right] \tag{16}$$

Taking the expectation over t, we have

Corollary 4.

$$MSE(\hat{\theta}(Y), \theta) = \mathbb{E}\left[\frac{1}{T}\sum_t |\hat{\theta}(Y_t) - Y_t^{hold}|^2 - \frac{1}{2}|Y_t - Y_t^{hold}|^2\right] \tag{17}$$

B Figures

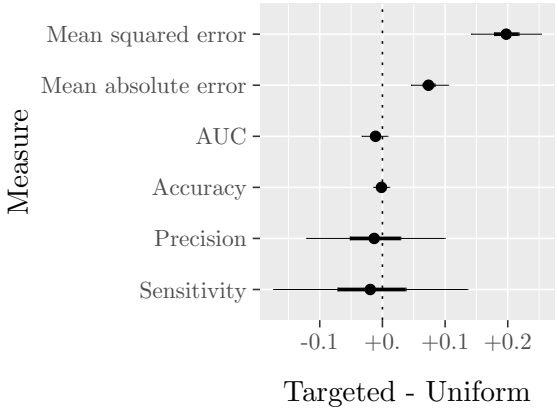

Figure B.1: Prediction performance of Bayesian model with one sample per site, comparison of targeted and uniform sampling designs. Solid dots and intervals show median and 50% and 95% credible intervals for difference between the targeted design and the uniform design using simulation to account for measurement error.

References

Berenguer, Marc et al. (2005). "Hydrological validation of a radar-based nowcasting technique". In: *Journal of Hydrometeorology* 6.4, pp. 532–549.

Biondi, Daniela et al. (2012). "Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice". In: *Physics and Chemistry of the Earth, Parts A/B/C* 42, pp. 70–76.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Carpenter, Bob et al. (2017). "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1.

Cha, YoonKyung et al. (2010). "Phosphorus load estimation in the Saginaw River, MI using a Bayesian hierarchical/multilevel model". In: *Water research* 44.10, pp. 3270–3282.

Dawson, CW and RL Wilby (2001). "Hydrological modelling using artificial neural networks". In: *Progress in physical Geography* 25.1, pp. 80–108.

Dorevitch, Samuel et al. (2017). "Monitoring urban beaches with qPCR vs. culture measures of fecal indicator bacteria: Implications for public notification". In: *Environmental Health* 16.1, p. 45.

Dotto, Cintia Brum Siqueira et al. (2014). "Impacts of measured data uncertainty on urban stormwater models". In: *Journal of hydrology* 508, pp. 28–42.

Duane, Simon et al. (1987). "Hybrid monte carlo". In: *Physics letters B* 195.2, pp. 216–222.

Eaton, Morris L, Alessandra Giovagnoli, and Paola Sebastiani (1996). "A predictive approach to the Bayesian design problem with application to normal regression models". In: *Biometrika* 83.1, pp. 111–125.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Griffith, John F and Stephen B Weisberg (2011). "Challenges in implementing new technology for beach water quality monitoring: lessons from a California demonstration project". In: *Marine Technology Society Journal* 45.2, pp. 65–73.

Gronewold, Andrew D and Mark E Borsuk (2010). "Improving water quality assessments through a hierarchical Bayesian analysis of variability". In: *Environmental science & technology* 44.20, pp. 7858–7864.

Gronewold, Andrew D, Mark E Borsuk, et al. (2008). *An assessment of fecal indicator bacteria-based water quality standards*.

Gronewold, Andrew D, Luke Myers, et al. (2011). "Addressing uncertainty in fecal indicator bacteria dark inactivation rates". In: *Water research* 45.2, pp. 652–664.

Gronewold, Andrew D, Song S Qian, et al. (2009). "Calibrating and validating bacterial water quality models: A Bayesian approach". In: *Water research* 43.10, pp. 2688–2698.

Gronewold, Andrew D, Mark D Sobsey, and Lanakila McMahan (2017). "The compartment bag test (CBT) for enumerating fecal indicator bacteria: basis for design and interpretation of results". In: *Science of the Total Environment* 587, pp. 102–107.

Gronewold, Andrew D, Craig A Stow, et al. (2013). "Differentiating Enterococcus concentration spatial, temporal, and analytical variability in recreational waters". In: *Water research* 47.7, pp. 2141–2152.

Hoffman, Matthew D and Andrew Gelman (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Kinzelman, Julie et al. (2003). "Enterococci as indicators of Lake Michigan recreational water quality: comparison of two methodologies and their impacts on public health regulatory events". In: *Appl. Environ. Microbiol.* 69.1, pp. 92–96.

Kuczera, George et al. (2006). "Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters". In: *Journal of Hydrology* 331.1-2, pp. 161–177.

Leube, PC, A Geiges, and W Nowak (2012). "Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design". In: *Water Resources Research* 48.2.

Liu, Xiaoyi et al. (2012). "Value of information as a context-specific measure of uncertainty in groundwater remediation". In: *Water resources management* 26.6, pp. 1513–1535.

Lohani, AK, Rakesh Kumar, and RD Singh (2012). "Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques". In: *Journal of Hydrology* 442, pp. 23–35.

Lucius, Nick et al. (2019). "Predicting E. coli concentrations using limited qPCR deployments at Chicago beaches". In: *Water research X* 2, p. 100016.

Nevers, Meredith B and Richard L Whitman (2011). "Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches". In: *Water research* 45.4, pp. 1659–1668.

Noble, Rachel T et al. (2010). "Comparison of rapid quantitative PCR-based and conventional culture-based methods for enumeration of Enterococcus spp. and Escherichia coli in recreational waters". In: *Appl. Environ. Microbiol.* 76.22, pp. 7437–7443.

Olyphant, Greg A and Richard L Whitman (2004). "Elements of a predictive model for determining beach closures on a real time basis: the case of 63rd Street Beach Chicago". In: *Environmental monitoring and assessment* 98.1-3, pp. 175–190.

Petris, Giovanni, Sonia Petrone, and Patrizia Campagnoli (2009). "Dynamic linear models". In: Springer.

Prüss, Annette (1998). "Review of epidemiological studies on health effects from exposure to recreational water". In: *International journal of epidemiology* 27.1, pp. 1–9.

Rabinovici, Sharyl JM et al. (2004). *Economic and health risk trade-offs of swim closures at a Lake Michigan beach*.

Renard, Benjamin et al. (2011). "Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation". In: *Water Resources Research* 47.11.

Rubin, Donald B (2004). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

Shively, Dawn A et al. (2016). "Prototypic automated continuous recreational water quality monitoring of nine Chicago beaches". In: *Journal of environmental management* 166, pp. 285–293.

Shortridge, Julie E, Seth D Guikema, and Benjamin F Zaitchik (2016). "Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds." In: *Hydrology & Earth System Sciences* 20.7.

Stow, Craig A et al. (2009). "Bayesian hierarchical/multilevel models for inference and prediction using cross-system lake data". In: *Real World Ecology*. Springer, pp. 111–136.

United States Environmental Protection Agency (2012). *Recreational water quality criteria*.

Vrugt, Jasper A et al. (2008). "Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation". In: *Water Resources Research* 44.12.

Whitman, Richard L, Zhongfu Ge, et al. (2010). "Relationship and variation of qPCR and culturable enterococci estimates in ambient surface waters are predictable". In: *Environmental science & technology* 44.13, pp. 5049–5054.

Whitman, Richard L and Meredith B Nevers (2004). *Escherichia coli sampling reliability at a frequently closed Chicago beach: monitoring and management implications*.

– (2008). "Summer E. coli patterns and responses along 23 Chicago beaches". In: *Environmental science & technology* 42.24, pp. 9217–9224.