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

Eric Potash[a†] Scott Steinschneider[b]

November 6, 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, model validation and comparison has largely omitted measurement error (defined here as variation in both sampling and the measurement technique) as an important source of uncertainty during validation. It is unknown how this omission affects estimates of model performance and comparisons between models. This study aims to fill this gap. First we derive a formula in terms of the measurement variance for the bias incurred when omitting measurement error in calculating a model's mean squared error. This leads us to develop a non-parametric method to correct estimates of mean squared error. To study other metrics of prediction performance (mean absolute error, sensitivity, precision, etc.) we develop a second method that uses simulations from a Bayesian model. These methods are applied to a comparison of two models (random forest and nearest neighbor) used to estimate the level of fecal indicator bacteria at 19 recreational beaches in the city of Chicago. We find that accounting for measurement error significantly changes estimates of model performance. Moreover it reveals substantial uncertainty underlying some of these estimates.

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). Here we define measurement error as the combined effect of error in the measurement process and sampling variability of in-situ FIB data.

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. In hydrology, for example, 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 errors in measurement techniques 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; Fancy 2013; 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 to a single metric of model performance, namely mean squared error (MSE). The second method is a Bayesian method that uses simulation from the posterior distribution of a Bayesian measurement error model. This method has the advantage of being applicable to 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, not to mention other domains. 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 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).

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

2 Materials and methods

In the sections below, we first describe the study site and data. Then we present two prediction models (estimators) of water quality. Next we present the proposed methods for inter-model comparison and cross-validation in the presence of measurement error. One of these, the Bayesian method, relies on a Bayesian model of FIB levels so we conclude this section by describing this Bayesian model.

2.1 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).

Figure 1: Map of the 19 recreational beaches on Lake Michigan in Chicago showing sample and prediction sites according to the proposed targeted sampling design of Lucius et al. 2019 described in section 2.2.2.

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 (https://data.cityofchicago.org/). 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 (https://darksky.net). Lake Michigan water levels at Calumet Harbor are taken from NOAA (https://tidesandcurrents.noaa.gov/).

Management decisions for each beach in Chicago are currently made by estimating FIB levels using the (geometric) mean of the samples at that site. Due to the cost of these measurements, the city has 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). 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.

2.2 Estimators

In this study we (re-)evaluate the estimates and management consequences of the random forest model. For a baseline comparison we consider a nearest neighbor model. The random forest and nearest neighbor models are referred to as estimators. We compare the performance of these estimators using both a naive validation method found in the literature and our proposed cross-validation methods that account for measurement error.

We denote the true (unobserved) level of *Enterococci* natural log cell equivalents per mL (log CE/mL) by θ_{jt} with $j = 1 \dots J$ a site index and $t = 1 \dots T$ a day index. Let Y_{ijt} be the observed measurements of θ_{jt}. In our case we typically have two measurements Y_{1jt}, Y_{2jt}.

Here we present two estimates of θ_{jt} at the prediction sites j. On day t, both estimators are based on the input vector of mean FIB measurements $\bar{Y}_{jt}$ at the proposed ten sampled sites (figure 1). The random forest estimator employs an additional input vector of K covariates

X_{jt} varying by date and site. The outputs are estimates of the FIB level at the proposed nine prediction sites.

2.2.1 Nearest neighbor estimator

For a prediction site j the nearest neighbor estimator simply predicts the FIB level to be equal to the mean level at the geographically nearest sampled site $n(j)$ on the same date (see figure 1):

$$\hat{\theta}^{\text{nn}}_{jt} = \bar{Y}_{n(j),t}.$$

2.2.2 Random forest estimator

Lucius et al. (2019) proposed a "hybrid nowcast model" using a random forest regression model with 400 trees (Breiman 2001). The outcomes used to fit the model were the mean levels at the prediction sites. The inputs to the model were the mean levels at the sampled sites together with covariates. Formally we can write the estimator as a vector of functions

$$\hat{\theta}^{\text{rf}}_{j}(\bar{Y}^{\text{sample}}_{t}, X_{jt})$$

where X_{jt} is a vector of $K = 11$ covariates (varying by site and date) and $\bar{Y}^{\text{sample}}_{t}$ is the vector of average measurements at the ten sample sites on date t. For this study we refit the random forest using our training set, which is larger than that of the original publication.

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, we added a separate wind speed for each cardinal direction. Finally, to reduce correlation among covariates we reparameterized the multi-day aggregate covariates from 1 day and 3 day (e.g. 1 day total rainfall and 3 day total rainfall) to 1 and 2-3 day (e.g. 1 day total rainfall and 2-3 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
	Weekday indicator

Table 1: Covariates included in the random forest and Bayesian models.

2.2.3 Exceedance predictions

The above are 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 predictions of the continuous FIB level must be transformed into binary predictions of exceedance. This is done using a threshold decision rule, i.e. $D(\hat{\theta}) = 1(\hat{\theta} > C)$, where the threshold C may depend on the estimator. For the baseline nearest neighbor estimator, we simply used the 1000 CE 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 nearest neighbor estimator to be the reference model.

An alternative to predict exceedance would model this binary outcome directly, i.e. classification. However, we continue the standard practice in FIB prediction of modeling the continuous outcome, i.e. regression, as this uses all available information (rather than discretizing observations for training) and allows us to use a single model for both continuous and binary outcomes.

2.3 Cross-validation

The purpose of this study is to develop an approach to compare the performance of multiple models (i.e., the estimators above) in the presence of measurement error. In cross-validation we evaluate the fidelity of estimated states $\hat{\theta}$ to the true state θ by a function $L(\theta, \hat{\theta})$. Here L is one of various performance metrics (e.g. MSE) and θ and $\hat{\theta}$ are restricted to dates t in a *test period* which we choose to be the most recent beach season, 2019. The random forest model was fit using data from a *training period*, i.e. prior to 2019; the nearest neighbor estimator does not require any fitting so only uses data from the test period.

Our challenge in cross-validation is that we never observe θ_{jt}. In the literature, θ_{jt} is often assumed to be exactly equal to the mean measurement $\bar{Y}_{jt}$ (Nevers and Whitman 2011; Fancy 2013; Shively et al. 2016; Lucius et al. 2019). Note that it is because these sites were in fact sampled that we can conduct this validation.

However, this method does not account for measurement uncertainty and it is unclear what the consequences of this omission are regarding the overall performance assessment of an estimator or the comparison of multiple estimators. We term this method of cross-validation *naive*, and propose two additional methods: *non-parametric* and *Bayesian*. The cross-validation methods are described below and summarized in figure 2.

Note that there are two sources of variation accounting for the difference between the true FIB level θ_{jt} and an observation Y_{ijt}. The first is sampling variation due to the fact that a water sample is taken at a specific point in time and space (Whitman and Nevers 2004). The second is measurement variation due to the qPCR technology used to analyze the sample

(Whitman, Ge, et al. 2010).

2.3.1 Naive validation

In naive validation we simply assume that the mean measurement is true: $\theta_{jt} = \bar{Y}_{jt}$. Then we can evaluate a single number $L(\theta, \hat{\theta})$. The use of naive validation is potentially flawed since the mean observation does not account for measurement error.

When the metric L is mean squared error, we can explicitly analyze the effect of measurement error. Assume a measurement error model

$$Y_{ijt} = \theta_{jt} + \epsilon_{ijt} \tag{1}$$

where ϵ_{ijt} are independent identically distributed measurement errors. We do not assume a distribution for ϵ but assume they have (finite) variance τ^2. Then with $\mathbb{E}$ denoting expectation over the random measurement errors ϵ we have

$$\begin{aligned}
\mathbb{E}|\hat{\theta}_{jt} - \bar{Y}_{jt}|^2 &= \mathbb{E}|\theta_{jt} + \bar{\epsilon}_{jt} - \hat{\theta}_{jt}|^2 && (2)\\
&= \mathbb{E}[|\theta_{jt} - \hat{\theta}_{jt}|^2 + |\bar{\epsilon}_{jt}|^2 - 2\bar{\epsilon}_{jt}(\theta_{jt} - \hat{\theta}_{jt})] && (3)\\
&= |\hat{\theta}_{jt} - \theta_{jt}|^2 + \frac{1}{2}\tau^2 && (4)
\end{aligned}$$

where we used the fact that the ϵ are independent of each other and both θ and $\hat{\theta}$ and there are two errors $\epsilon_{1jt}, \epsilon_{2jt}$ so that $\mathbb{E}[\bar{\epsilon}_{jt}^2] = \frac{1}{2}\tau^2$.

This formula means that under these mild assumptions, the naive estimate of MSE *overestimates* the true MSE by a multiple of the measurement error variance. Thus, the greater the measurement error variance τ^2, the larger the distortion of naive validation for the particular metric of MSE.

However, since the distortion does not depend on which estimator $\hat{\theta}$ is being evaluated (e.g. random forest or nearest neighbor), the naive validation will give an unbiased estimate of the difference in performance, i.e.

$$\mathbb{E}[MSE(\bar{Y}, \hat{\theta}^{\text{rf}}) - MSE(\bar{Y}, \hat{\theta}^{\text{nn}})] = MSE(\theta, \hat{\theta}^{\text{rf}}) - MSE(\theta, \hat{\theta}^{\text{nn}}) \tag{5}$$

2.3.2 Non-parametric validation

Equation 2 shows that if we can estimate the measurement error τ then we can correct the bias of the naive estimate of MSE by subtracting it off. If we have more than one FIB measurement per beach-day in the data (as is the case in our data set), we can estimate τ using the standard sample variance estimator, and then average across beaches and days.

Namely if Y_{1jt} and Y_{2jt} are the two observations with measurement error ϵ_{1jt} and ϵ_{2jt} as in equation 1, we define the estimate

$$\hat{\tau}_{jt}^2 = \frac{1}{2}|Y_{1jt} - Y_{2jt}|^2 \tag{6}$$

which is unbiased because

$$\mathbb{E}[\hat{\tau}_{jt}^2] = \mathbb{E}\left[\frac{1}{2}|\epsilon_{1jt} - \epsilon_{2jt}|^2\right] \tag{7}$$

$$= \tau^2. \tag{8}$$

Combining 2 and 7 we have the following estimate for the mean-squared error of $\hat{\theta}_{jt}$:

$$\mathbb{E}[|\hat{\theta}_{jt} - \bar{Y}_{jt}|^2 - \frac{1}{2}\hat{\tau}_{jt}^2] = |\hat{\theta}_{jt} - \theta_{jt}|^2 \tag{9}$$

We average across sites j and dates t to estimate $MSE(\hat{\theta}, \theta)$. We bootstrap this estimate across t to estimate the sampling distribution.

We emphasize 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 1). However, this approach is limited to the specific error metric of MSE.[1]

2.3.3 Bayesian validation

The approaches to cross-validation presented above either: assume the true FIB level θ_{jt} is equal to the mean of available observations (naive); or indirectly 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 (Rubin 2004) of the outcome θ using simulations from a Bayesian model.

Given a Bayesian model of (θ, Y, X) we sample θ at the prediction sites from the posterior distribution $\theta|Y, X$. (The details of the model and the Monte Carlo software used to perform this sampling are described in section 2.4 below.) Then we can simply evaluate $L(\theta, \hat{\theta})$. Here, as in naive and non-parametric validation, $\hat{\theta}$ are estimates at the prediction sites based on covariates at those sites and measurements at the sampled sites. By repeatedly sampling $\theta|Y, X$ and evaluating $L(\theta, \hat{\theta})$ we sample the target distribution $L(\theta, \hat{\theta})|Y, X$. We emphasize that Y includes all sites. Thus, unlike the prediction estimators which only use observations from the sample sites, the Bayesian simulations use observations from the prediction sites themselves (figure 1).

The Bayesian cross-validation method assumes that the Bayesian model is correct. However, it takes into account both uncertainty in model parameters and uncertainty in observations due to measurement error. In addition, the MSE error metric inferred under the Bayesian cross-validation method can be validated against that of the non-parametric approach, which makes fewer assumptions.

Yet compared to the non-parametric validation above, which can only estimate MSE, Bayesian simulation has the advantage that it can be used to estimate any prediction performance metric. We consider several (table 2), including MSE, mean absolute error (MAE), and the

[1] In principle the bootstrap could be used to estimate θ_{jt}, providing a non-parametric approach to cross-validation for any metric. But with only two samples of each θ_{jt} this approach is not viable here.

area under the receiver operating curve (AUC) to evaluate predictions of the continuous FIB level. The remaining metrics use binary classifications which are obtained from continuous predictions using a threshold (see section 2.2.3): precision measures the proportion of exceedance predictions which are correct; sensitivity measures the proportion of exceedances which are correctly predicted; specificity measures the proportion of non-exceedances which are correctly predicted. These latter metrics are particularly relevant to decision-making in recreational water quality management, where binary decisions (e.g. site closure) are often based on water quality predictions exceeding a predetermined threshold.

Another advantage of the Bayesian cross validation approach, though we do not take it here, is that it can be used to investigate scenarios where the data are insufficient for the naive or non-parametric approaches. For example, we could simulate more measurements $Y_{ijt}|\theta$ than are in the data to study the performance of estimators using sample designs that are not possible to study directly.

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

Table 2: Applicability of cross-validation methods to prediction performance metrics. Abbreviations: area under receiver operating characteristic curve (AUC).

2.4 Bayesian model

In order to implement Bayesian cross validation (section 2.3.3) we need a Bayesian model of (θ, X, Y). We use a linear regression (on the log scale) model with coefficients varying by site:

$$\theta_{jt} = X_{jt}\beta_j + \eta_{jt} \tag{10}$$

where X_{jt} is a vector of K covariates and for each site j and β_j is a vector of K regression coefficient. We use the same covariates as in the random forest above but add an intercept and parameterize the day of year as a B-spline with 4 degrees of freedom (since this model is linear as opposed to the non-linear random forest). Thus $K = 15$.

On top of this regression we add three components. First we add a multivariate normal error distribution with covariance matrix Σ to model correlation in the errors across beaches on a given day t:

$$\eta_t \sim \text{Normal}(0, \Sigma) \tag{11}$$

Input: Test period measurements Y and covariates X
Estimator $\hat{\theta}$
Performance metric L
Result: Estimate of $L(\theta, \hat{\theta})$

1 Split Y between sample sites Y^{sample} and prediction sites Y^{predict}
2 Predict $\hat{\theta}$ using measurements at sample sites Y^{sample} and covariates X
3 Assume θ at the prediction site j equals mean measurement $\bar{Y}_{jt}^{\text{predict}}$
4 Evaluate $L(\theta, \hat{\theta})$

(a) Naive cross-validation

Input: Test period measurements Y and covariates X
Estimator $\hat{\theta}$
Result: Estimate of $\text{MSE}(\theta, \hat{\theta})$

1 Produce naive estimate of MSE using algorithm (a)
2 Estimate measurement error $\hat{\tau}_{jt}^2$ using equation 6
3 Correct bias of naive estimate using equation 9

(b) Non-parametric cross-validation

Input: Test period measurements Y, covariates X
Estimator $\hat{\theta}$
Performance metric L
Fitted Bayesian model of (θ, X, Y)
Result: Estimate of $L(\hat{\theta}, \theta)$

1 Simulate states $\theta|Y, X$ at prediction sites
2 Predict $\hat{\theta}$ using measurements Y at sample sites and covariates X
3 Evaluate $L(\theta, \hat{\theta})$

(c) Bayesian cross-validation

Figure 2: Cross validation methods.

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:

$$\beta_{jk} \sim \text{Normal}(\mu_{\beta_k}, \sigma^2_{\beta_k}) \tag{12}$$

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).

The final component is an additive and normally distributed measurement error with variance τ^2 (Gronewold, Qian, et al. 2009):

$$Y_{ijt} \sim \text{Normal}(\theta_{jt}, \tau^2). \tag{13}$$

Note that, unlike the random forest model which is fit to beach-day mean levels $\bar{Y}_{jt}$, the Bayesian model is fit to the individual observations Y_{ijt}.

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

$$\Sigma = \text{diag}(\sigma) \cdot \Omega \cdot \text{diag}(\sigma) \tag{14}$$

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.

(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 of 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).

With the uninformative prior on Σ we are making relatively weak assumptions about the covariance structure. This is possible in our application because of the relatively small number (19) of sites and the efficiency of Hamiltonian Monte Carlo. In applications with more sites, it may be useful to model the covariance in terms of the distance between sites j and k using a Gaussian process model or in terms of an adjacency matrix using a conditional autoregressive model (Gelfand et al. 2010).

3 Results

We first present details of the fitted Bayesian model. We then compare all three cross-validation methods estimates' of MSE since this is the only metric where the non-parametric method is applicable. Finally we compare Bayesian and naive estimates of all prediction performance metrics.

3.1 Bayesian model fit

The Bayesian model was fit using qPCR measurements from the 2015 to 2019 seasons. There were 13109 such observations made at the 19 beaches on 430 days. A subset of these measurements at a cluster of beaches are shown in figure 3.

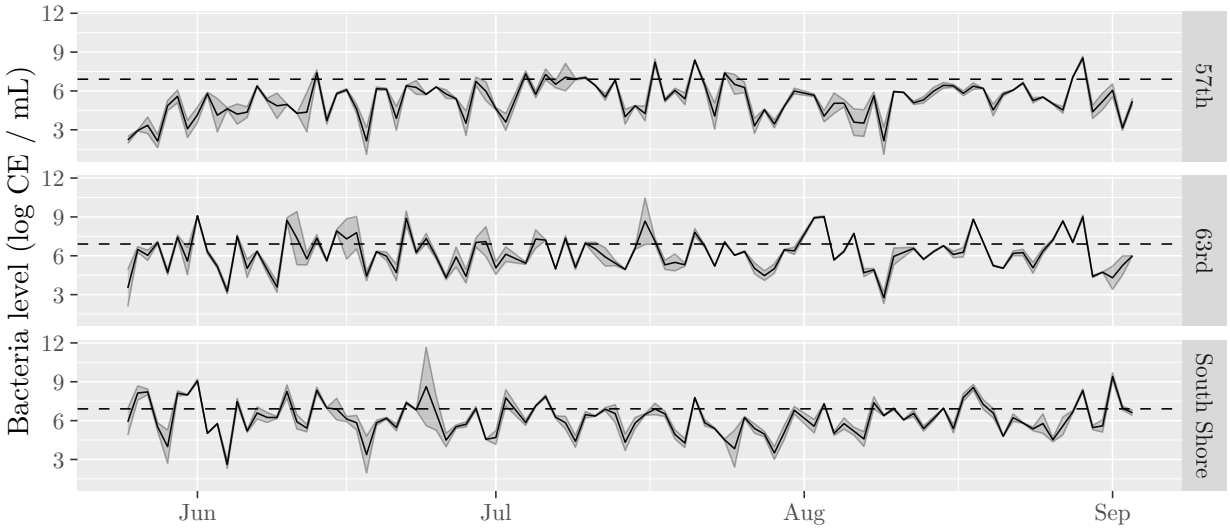

Figure 3: Fecal indicator bacteria measurements at a cluster of 3 of the 19 Chicago beaches during the 2018 beach season. Gray region spans minimum and maximum measurements, solid line connects daily means. Dashed lines indicate action threshold of log 1000 cell equivalents (CE).

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, but this 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 (95% CI, 0.74 to 0.8). This is 30% of the variance of daily means $\bar{Y}_{jt}$ of 2.5. We separately fit the Bayesian model to *E. coli* culture data and estimated a measurement error τ^2 of 0.37 (95% CI, 0.34 to 0.4) which is consistent with the estimate .401 of Whitman and Nevers 2004 (their table 2). This is just 12% of the variance of daily means of 3.0, supporting the suggestion of Whitman, Ge, et al. (2010) that measurement error is greater for qPCR than culture tests.

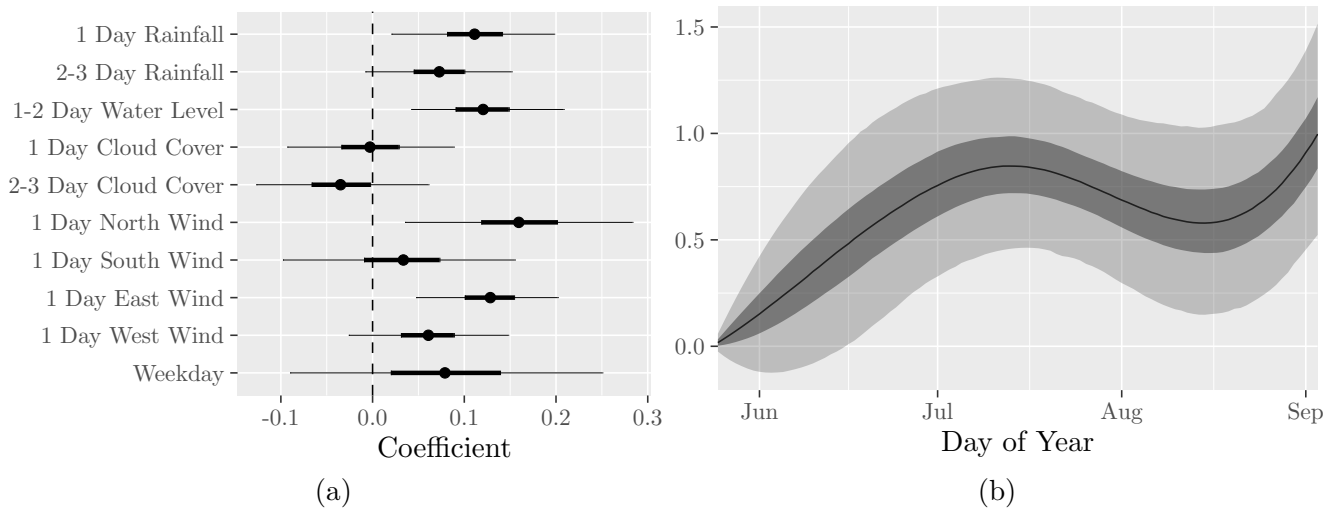

Figure 4: Bayesian model estimated (a) standardized coefficients $\mu_{\beta_k}/\text{sd}(x_k)$ and (b) day of year trend.

3.2 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.

3.2.1 Mean squared error under all cross validation methods

We start by examining the three cross validation methods on the metric where they can all be compared, namely MSE. Figure 5 presents these estimates. While naive validation gives point estimates, non-parametric and Bayesian validation give distributions. Moreover, the latter give joint distributions of MSE for the two estimators and so yield distributions for the difference in MSE between the two estimators.

There are four findings here. First, we anticipated that naive validation would give a positively biased estimate of mean-squared error (2) and we see that it does give larger estimates than both non-parametric and Bayesian validation. For both random forest and nearest neighbor estimators, the naive estimates of MSE lie above the 95% intervals estimated by non-parametric validation. Because non-parametric validation accounts for measurement error, we are inclined to trust its results and dismiss naive validation which is a priori flawed.

Second, we find as expected (equation 5) that while naive validation overstates the MSE of both estimators, the estimated *difference* between the estimators agrees with the difference given by non-parametric validation.

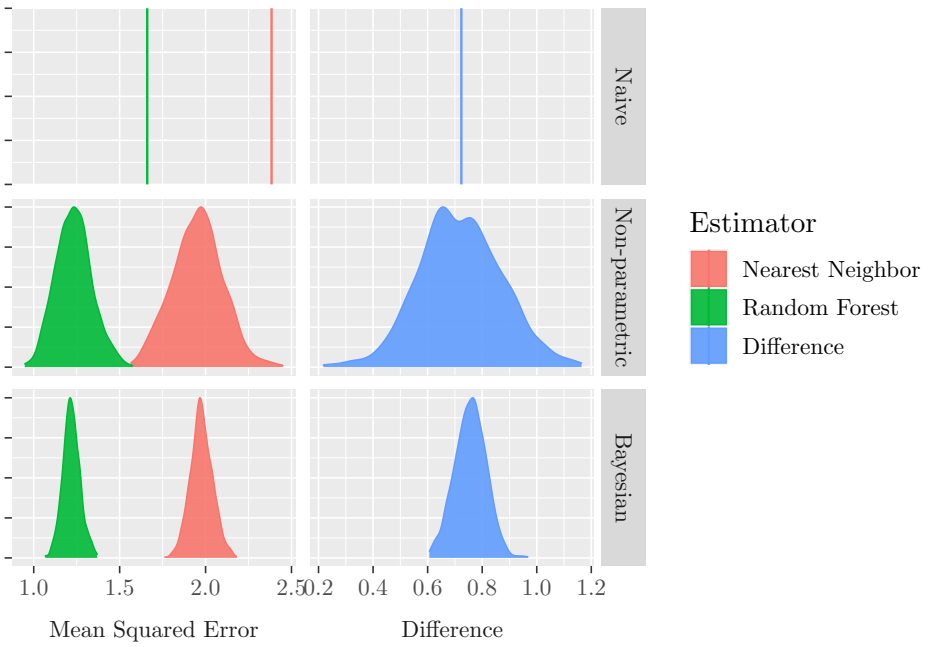

Figure 5: Mean squared error of nearest neighbor and random forest predictions and their difference estimated using naive, non-parametric, and Bayesian cross validation methods.

Third, we find remarkable agreement between non-parametric and Bayesian MSE estimates. Because non-parametric validation makes few assumptions, this agreement provides evidence to support our use of the Bayesian validation method to further explore the performance of estimators and metrics for which we do not have a non-parametric method (discussed next).

Fourth, the Bayesian method provides narrower uncertainty around its estimates than the non-parametric method. This may be explained by the fact that when estimating the mean squared error of a given prediction $\hat{\theta}_{jt}$, the non-parametric method only uses the measurements Y_{ijt} at the site while the Bayesian model uses the additional information of covariates and measurements at other sites.

3.2.2 All performance measures using naive and Bayesian cross validation

We proceed to evaluate the full set of performance metrics using Bayesian and naive cross validation methods. We started by using Bayesian validation to estimate the expected sensitivity of the nearest neighbor estimator with a binary classification threshold of 1000 CE. The estimate was 95.6%, and to match this (section 2.2.3), a threshold of 440 for the random forest was calibrated. Estimates for all prediction performance metrics are shown in figure 6.

According to the both naive and Bayesian validation, random forest outperforms nearest neighbor in all metrics (except specificity, where they are calibrated to match). However, the discrepancy between Bayesian and naive validation, first documented for MSE in section 3.2.1 above, continues here across more metrics. Unlike MSE which was systematically overestimated (i.e. pessimistic) using naive validation, other measures are variously pessimistic as well as optimistic, such as when naive validation estimates the precision of the random

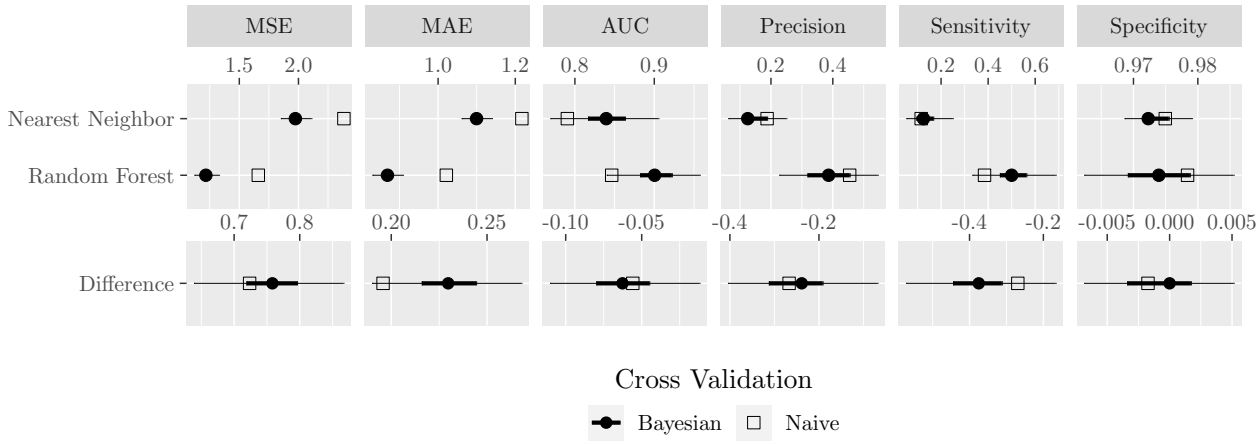

Figure 6: Prediction performance estimates using naive and Bayesian validation. Solid dots and intervals show median and 50% and 95% credible intervals using simulation to account for measurement error. Open squares show naive estimates without accounting for measurement error. Abbreviations: area under receiver operating characteristic curve (AUC).

forest to be 0.45 while the Bayesian point estimate is 0.39.

However, as for MSE above, for each metric the bias of naive validation (relative to Bayesian validation) is consistently in a fixed direction for both random forest and nearest neighbor. Naive estimates of the *difference* in predictive performance benefit from this consistency in that the biases of the difference estimates are not larger than those of the absolute performance estimates. However, the discrepancy between naive and Bayesian validation of these estimated differences can still be quite large, as for example sensitivity where naive validation estimates random forest to be an improvement of 0.27 while the Bayesian estimate is 0.38.

The discrepancy between naive and Bayesian validation turns out to be greatest for MSE and MAE which measure performance of continuous FIB level predictions, and less for the others which measure binary prediction performance.

While the Bayesian estimates of MSE and MAE estimates are relatively precise, the estimates of precision and sensitivity are very uncertain. This is explained by the fact that exceedances (positive events) form the denominator in the sensitivity metric and contribute to the numerator in precision. As a tail event defined by a sharp threshold, exceedance is difficult to measure and model precisely.

These metrics as well as their uncertainty have important consequences for management decision making. Sensitivity for example is the proportion of exceedances that are correctly predicted, which for the random forest is 0.50 (95% CI, 0.33 to 0.69). Thus it measures the proportion of exceedances for which a protective action is taken. Precision on the other hand is the proportion of predicted exceedances that are correct, which for the random forest is 0.39 (95% CI, 0.23 to 0.55). Thus it measures the proportion of protective actions which are taken when there is in fact an exceedance.

4 Discussion

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), including predictive models for recreational water quality (e.g. Nevers and Whitman 2011; Fancy 2013; Shively et al. 2016; Lucius et al. 2019). In this study we examined the effect of this omission.

For the specific prediction performance metric of MSE we showed that ignoring measurement error biases validation results (equation 2). The size of the bias depends on the size of the measurement error, which is very large in our context of recreational water quality. Next we proposed two methods for model validation and inter-comparison that account for measurement error. The first was a non-parametric method making few assumptions but limited to the metric of mean squared error. The second was a Bayesian method that uses simulations from a parametric model to estimate any performance metric. We applied these methods to the evaluation of prediction models of FIB levels at beaches in Chicago and found that not accounting for measurement error significantly mis-estimated model performance across a range of metrics. Moreover it failed to quantify the uncertainty of prediction performance. Our non-parametric and Bayesian approaches overcame these issues.

Accurate model skill assessments are important. These estimates are required by water quality managers to understand the utility of model predictions for decision-making. Bias in estimated performance metrics (e.g. MSE or sensitivity, see Figure 6) could skew how decision-makers interpret model predictions or select among competing models, as could the presentation (or lack thereof) of performance uncertainty. More generally, performance estimates and their uncertainty are essential to understanding the public health consequences of management decisions made on the basis of these models. Measures of model performance are also used to inform decisions about additional sampling (if deemed necessary to improve performance), which could be costly. For example, if the city of Chicago were to conclude based on an assessment of model performance that they needed two samples per beach-day at the 19 beaches, rather than the current proposal which includes roughly half the number of samples, the additional sampling cost would total about $57,000 per season.[2]

For reasons explored above, estimates of binary prediction performance metrics (e.g. sensitivity and precision) are especially uncertain. If decision makers wish to reduce the uncertainty of these estimates they will need to reduce the uncertainty of estimates of the true FIB level θ. This could be achieved through increased sampling intensity or improvements to the Bayesian model. On the other hand, this uncertainty stems from events which, while uncertain to be above the threshold, are still quite likely to be near it (albeit below), in which case a protective action may still be desirable depending on its costs.

Both the non-parametric and Bayesian approaches to cross-validation proposed in this study help overcome limitations of a naive approach. However, the Bayesian approach to cross-validation is more flexible in its ability to compare models across a variety of metrics, some of

[2]Assuming 100 days per season and an estimated processing cost of $30 per qPCR sample (Griffith and Weisberg 2011)

which might be particularly relevant to decision-making (e.g., predictions of exceedances over a water quality threshold). In addition, the Bayesian model could also be used as a predictive model in its own right, instead of just as a tool to support cross-validation and inter-model comparison of other models. However, care would be required to validate Bayesian model predictions based on simulations from that very same model.

Additional model improvements may also be needed to improve prediction, and even cross validation, using the Bayesian model. For instance, one natural way to extend the Bayesian model is through temporal 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. 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 also 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. These efforts are left for future work. Importantly though, even if the Bayesian model is not the best predictive model of FIB levels (as compared to, say, a machine learning model), it can still provide realistic simulations of uncertain bacteria concentrations. Our work shows how to use these Bayesian simulations to cross-validate other predictive models for more realistic assessments of model skill compared to a naive approach.

5 Conclusion

Our study shows that, in the evaluation of predictive models of FIB levels for recreational water quality, the omission of measurement error can substantially distort estimates of prediction performance. To address this issue, we contributed two new methods for performing cross-validation and inter-model comparison while accounting for measurement uncertainty: a non-parametric method for MSE and a Bayesian simulation method for any performance metric. We used the non-parametric method to validate the Bayesian method, and compared both to a naive method found in the literature that doesn't account for measurement error. We found that, across several estimators and metrics, the naive approach can be systematically biased and fails to quantify the occasionally large uncertainty in performance. These represent significant issues for water resource managers using such estimators for decision support. Our non-parametric and Bayesian methods overcame these issues and so we recommend that future evaluations use our methods to account for measurement error in model validation and comparison.

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.

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.

Fancy, Donna S (2013). *Developing and implementing predictive models for estimating recreational water quality at Great Lakes beaches*. US Department of the Interior, US Geological Survey.

Gelfand, Alan E et al. (2010). *Handbook of spatial statistics*. CRC press.

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, interpretabil-

ity, 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.