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

E. Potash[1*] **and S. Steinschneider**[2]

[1]University of Illinois Urbana-Champaign, Illinois, USA
[2]Cornell University, Ithaca, New York, USA

Key Points:

- Recreational water quality models are typically validated and compared ignoring measurement errors
- New methods to account for such errors are developed and applied to fecal indicator bacteria models at Chicago beaches
- Accounting for measurement error significantly affects validation results and reveals high uncertainty

*1101 W. Peabody, Suite 350 (NSRC), MC-635, Urbana, IL 61801

Corresponding author: Eric Potash, epotash@illinois.edu

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 the bias incurred when omitting measurement error in calculating a model's mean squared error. We then develop a non-parametric validation 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 validation method that uses simulations from a Bayesian validation model. These methods are applied to a comparison of two prediction models (random forest and nearest neighbor) used to predict the level of fecal indicator bacteria at 9 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.

1 Introduction

Recreational waterways are subject to contamination by bacteria from various sources including stormwater, sewage, and wildlife (Whitman & 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 & Nevers, 2004; Whitman et al., 2010). Here we define measurement error as the combined effect of error in the measurement process and in-situ sampling variability.

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 et al., 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 et al., 2008; Gronewold & Borsuk, 2010; Gronewold et al., 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 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 validating prediction models, we found that researchers often ignored measurement error, simply assuming that a measurement (or mean of multiple measurements) represented the true bacteria level (Nevers & Whitman, 2011; Francy, 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 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 methods for model validation that 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 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 prediction of bacteria levels (Nevers & Whitman, 2011; Shively et al., 2016; Lucius et al., 2019).

2 Materials and methods

In our analysis we present both prediction models (section 2.2) and validation methods (2.3). The prediction models are used to predict FIB levels at unsampled beaches to support beach management decisions. The validation methods are used to evaluate and compare these prediction models and their resulting decisions. The focus of this study is on the methods for validation, not the specific prediction models being validated. We compare a commonly used method for validation (termed naive validation) against two new methods (a non-parametric method and Bayesian method) that account for measurement error. We note that Bayesian validation relies on an auxiliary model (termed the Bayesian validation model), which is used strictly to validate other prediction models and not for prediction itself (more detail given in section 2.3.4).

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 & Whitman, 2011).

Traditionally, administrators collected two samples per site for culture measurement of *E. coli* in terms of colony forming units (CFU) per 100 mL. Management decisions were made on the basis of the geometric mean measurement exceeding 235 CFU/100 mL. Culture measurements take at least 12-24 hours due to the bacteria incubation period. 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).

Starting in 2015, quantitative polymerase chain reaction (qPCR) measurements of *Enterococci* have been employed. This method quantifies indicator bacteria in less than two hours in terms of cell equivalents (CE) by comparing the sample to a calibrator with known number of *Enterococci* cells. A subset of these measurements are shown in figure 1. For details and comparisons of culture and qPCR measurements and their consequences see Noble et al. (2010), U.S. EPA (2012), and Dorevitch et al. (2017). Managers in Chicago currently estimate FIB levels at each beach each day using the geometric mean of two qPCR sample measurements. Management decisions are made, following U.S. EPA guidance (U.S. EPA, 2012), based on this estimate exceeding 1000 CE/mL.

Due to the cost of these measurements (Whitman et al., 2010) and the historical correlation of FIB levels between sites, the city has proposed reducing sampling to ten beaches and using a random forest (RF) model to predict levels at the remaining beaches (Lucius et al., 2019). The sampled beaches were chosen 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. Five beaches were selected to be sampled, one from each cluster. The five historically high FIB sites and five cluster representatives together give 10 sampled beaches, leaving 9 sites at which to make predictions (see figure 1). The prediction model uses daily meteorological and hydrological covariates collected between 2015-2019 for the months of May-September.

2.2 Prediction models of bacteria levels at unsampled sites

In this study we (re-)evaluate the predictions and management consequences of the RF model, and compare its performance against a benchmark NN model. The RF and NN models both serve as candidate *prediction* models. The purpose of this work is to assess how to validate and compare performance across prediction models given measurement error in the observations. We use the RF and NN prediction models to demonstrate our validation methods, but note that other prediction models could have been used for this purpose. Prior to describing the validation methods that are the focus of this work (see section 2.3), we first introduce notation and details of the specific predictive models used in the case study.

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} which are replicates, taken at the same time and location.

Here we present NN-based and RF-based predictions of θ_{jt} at the prediction sites j. On day t, both predictions are based on the input vector of mean FIB measurements

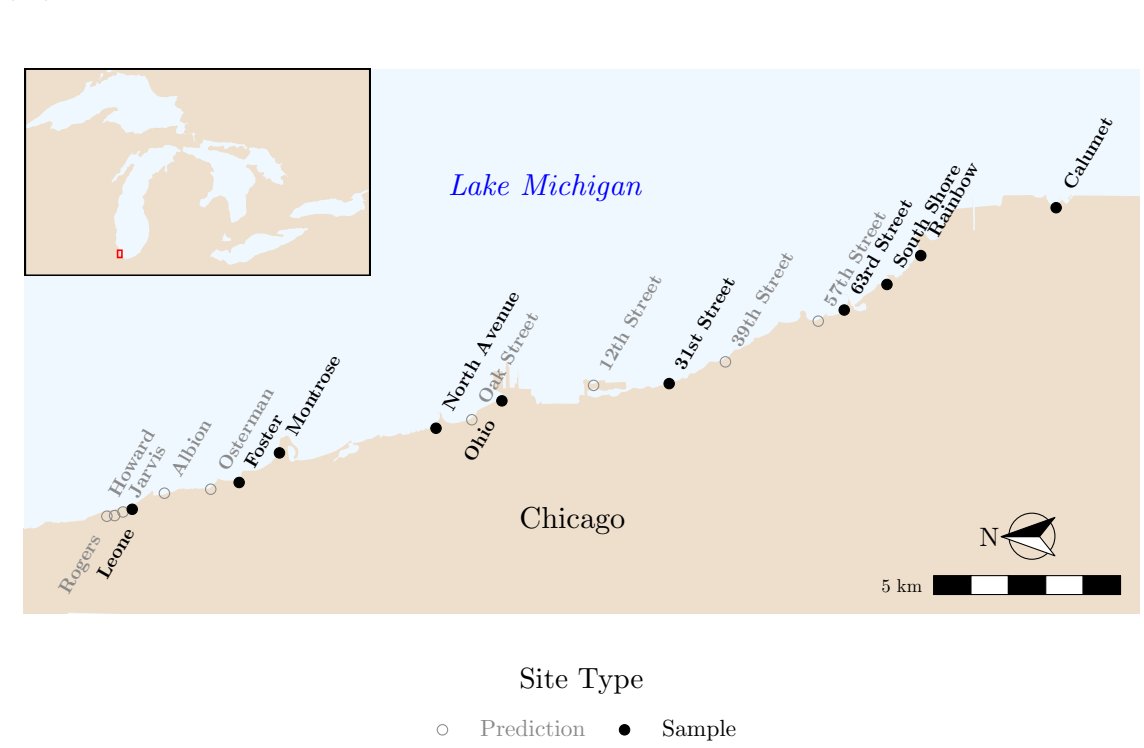

Figure 1: (A) Map of the 19 recreational beaches in Chicago on Lake Michigan (see inset for location within the Great Lakes) showing sample and prediction sites according to the proposed targeted sampling design of Lucius et al. (2019) described in section 2.2.2 and (B) daily fecal indicator bacteria levels during the 2019 beach season at three nearby sites. Gray circles are qPCR measurements (typically 2 per site per day), black line connects daily mean measurements, red region is 95% interval of Bayesian validation model posterior, and dashed line indicates action threshold of log 1000 cell equivalents (CE) per mL.

$\bar{Y}_{jt}$ at the proposed ten sampled sites (figure 1). The RF prediction employs an additional input vector of K covariates X_{jt} varying by date and site. The outputs are predictions of the true FIB level at the proposed nine prediction sites.

2.2.1 Nearest neighbor (NN) prediction model

For prediction site j, let $n(j)$ be the index $(1, \ldots, J)$ of the geographically nearest sampled site (figure 1A). The NN model predicts the FIB level at site j on date t to be equal to the mean level at this neighbor on the same date:

$$\hat{\theta}^{\text{nn}}_{jt} = \bar{Y}_{n(j),t}. \tag{1}$$

This model serves as a simple but practical benchmark.

2.2.2 Random forest (RF) prediction model

Lucius et al. (2019) proposed a "hybrid nowcast model" using a RF regression model with 400 trees (Breiman, 2001). The outcomes used to fit the model were the mean FIB 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 prediction model as a vector of functions

$$\hat{\theta}^{\text{rf}}_{jt}(\bar{Y}^{\text{sample}}_{t}, X_{jt}) \tag{2}$$

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 RF using our training set, which is larger than that of Lucius et al. (2019). The covariates X_{jt} mirror those of the original publication:

- Precipitation: 1 and 2-3 day total rainfall, 1-2 day change in water level
- Sunlight: 1 and 2-3 day average cloud cover
- Wind: 1 day average N/S/E/W wind speed
- Time: day of year, weekday indicator

where the weekday indicator is an indicator for whether the date is a weekday or weekend and 1 day, 1-2 day, and 2-3 day covariates aggregate over the period 24 hours, 24-48 hours, and 48-72 hours prior.

2.2.3 Calibration of exceedance predictions

The above are prediction models of continuous FIB levels, but decisions are based on the binary event of exceeding 1000 CE/mL (section 2.1), for which an FIB level prediction must be transformed into a binary exceedance prediction. Some of our performance measures (e.g. precision) evaluate these exceedance predictions. For the baseline NN prediction model, we predicted an exceedance whenever the predicted FIB was greater than the 1000 CE/mL threshold. Since the RF is known to produce biased predictions, Lucius et al. (2019) calibrated a custom threshold so that the resulting specificity (equivalently false positive rate) matched that of a reference model. We follow this approach, taking NN as the reference model.

Exceedance could alternatively be predicted by modeling the 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 and allows us to use a single model for both continuous and binary outcomes.

2.3 Validation methods for estimating prediction model performance

The purpose of this study is to develop an approach to compare the performance of the above prediction models in the presence of measurement error. In validation we evaluate the fidelity of predicted 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 the prediction sites (figure 1) and dates t in a *test period* which we choose to be the most recent beach season, 2019. The RF model was fit using data from a *training period*, i.e. prior to 2019; the NN model does not require any fitting. We used a single training and test set, known as hold-out validation (Schneider & Moore, 2000), for simplicity. All of the methods below can be easily adapted for cross-validation with multiple folds.

Our challenge in 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 & Whitman, 2011; Francy, 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 a prediction model or the comparison of multiple models. We term this method of validation *naive*, and propose two additional methods: *non-parametric* and *Bayesian*. The 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 & Nevers, 2004). The second is measurement variation due to the qPCR technology used to analyze the sample (Whitman et al., 2010). As stated earlier, we define measurement error as the combination of these sampling and analysis variations.

2.3.1 Naive validation not accounting for measurement error

In naive validation we simply assume that the mean measurement is true: $\theta_{jt} = \bar{Y}_{jt}$. Then we can evaluate $L(\theta, \hat{\theta})$ as a point estimate of the performance (in contrast to non-parametric and Bayesian validation which estimate performance distributions). The use of naive validation is potentially flawed since the mean observation does not account for measurement error.

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

$$Y_{ijt} = \theta_{jt} + \epsilon_{ijt} \tag{3}$$

where ϵ_{ijt} are independent identically distributed measurement errors, defined as the sum of sampling and analysis errors. We do not assume a measurement error distribution but assume the errors have fixed (finite) variance $\mathrm{Var}(\epsilon_{ijt}) = \tau^2$ independent of the measurement number i, site j, and date t. Then with $\mathbb{E}$ denoting expectation over the random measurement errors ϵ we have

$$\begin{aligned}
\mathbb{E}|\hat{\theta}_{jt} - \bar{Y}_{jt}|^2 &= \mathbb{E}|\theta_{jt} + \bar{\epsilon}_{jt} - \hat{\theta}_{jt}|^2 && (4)\\
&= \mathbb{E}[|\theta_{jt} - \hat{\theta}_{jt}|^2 + |\bar{\epsilon}_{jt}|^2 - 2\bar{\epsilon}_{jt}(\theta_{jt} - \hat{\theta}_{jt})] && (5)\\
&= |\hat{\theta}_{jt} - \theta_{jt}|^2 + \frac{1}{2}\tau^2 && (6)
\end{aligned}$$

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 $\mathbb{E}[\bar{\epsilon}_{jt}^2] = \frac{1}{2}\tau^2$.

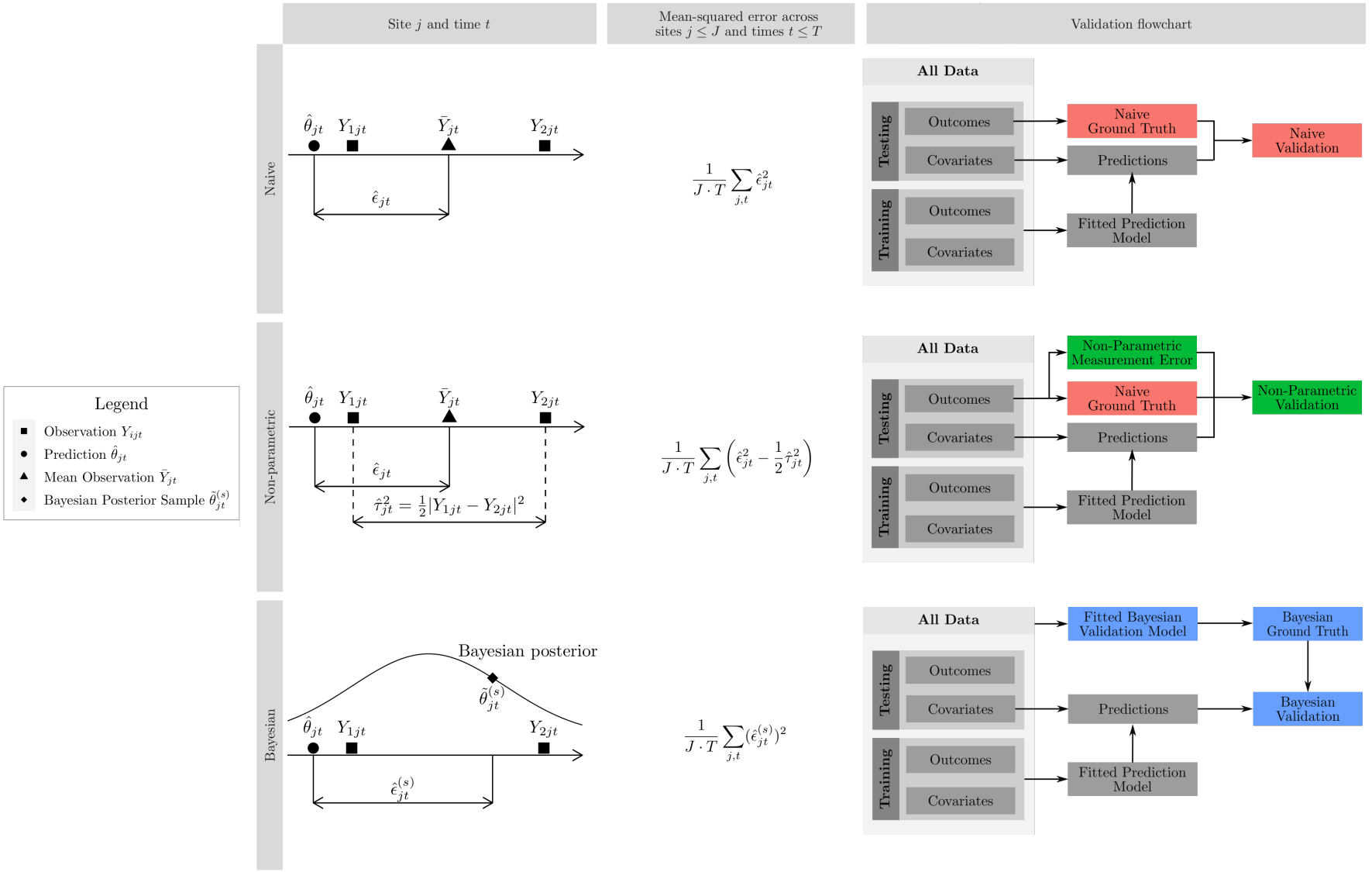

Figure 2: Validation methods. Prediction models are fit using training (2015-2018) covariates and outcomes, and use testing (2019) covariates to predict testing outcomes. For site j and time t the naive method simply compares the prediction $\hat{\theta}_{jt}$ with the mean observation $\bar{Y}_{jt}$ to give a point estimate $\hat{\epsilon}_{jt} = \hat{\theta}_{jt} - \bar{Y}_{jt}$ of the error, which is averaged across sites j and times t to estimate MSE. The non-parametric method corrects the naive method by incorporating an estimate $\hat{\tau}$ of the measurement error. This is bootstrapped to estimate a sampling distribution. The Bayesian method compares the prediction to a sample of the posterior distribution of the FIB state in a Bayesian validation model that is fit to and conditioned on all the data. For each sample of the posterior $\tilde{\theta}^{(s)}_{jt}$ (indexed by s) there is a corresponding sample of MSE. Abbreviations: mean squared error (MSE).

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 prediction model is being evaluated (e.g. RF or NN), the naive validation will give an unbiased estimate of the difference in performance, i.e.

$$\mathbb{E}[MSE(\bar{Y}, \hat{\theta}^{\text{rf}}) - MSE(\bar{Y}, \hat{\theta}^{\text{nn}})] = MSE(\theta, \hat{\theta}^{\text{rf}}) - MSE(\theta, \hat{\theta}^{\text{nn}}) \tag{7}$$

2.3.2 Non-parametric validation accounting for measurement error

Equation 4 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 3, we define the estimate

$$\hat{\tau}^2_{jt} = \frac{1}{2}|Y_{1jt} - Y_{2jt}|^2 \tag{8}$$

which is unbiased because

$$\mathbb{E}[\hat{\tau}^2_{jt}] = \mathbb{E}\left[\frac{1}{2}|\epsilon_{1jt} - \epsilon_{2jt}|^2\right] \tag{9}$$

$$= \tau^2. \tag{10}$$

Combining 4 and 9 we have the following estimate for the mean-squared error of $\hat{\theta}_{jt}$:

$$\mathbb{E}[|\hat{\theta}_{jt} - \bar{Y}_{jt}|^2 - \frac{1}{2}\hat{\tau}^2_{jt}] = |\hat{\theta}_{jt} - \theta_{jt}|^2 \tag{11}$$

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 (unlike naive validation which gives a point estimate).

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 3). However, this approach is limited to the specific error metric of MSE.

2.3.3 Bayesian validation accounting for measurement error

The approaches to validation presented above either: 1) assume the true FIB level θ_{jt} is equal to the mean of available observations (naive); or 2) indirectly estimate a specific error metric, MSE, under a specific sampling design (non-parametric). An alternative and more general approach is to use a Bayesian model to simulate the latent FIB state θ in validation. This is a form of Monte Carlo uncertainty propagation (ISO, 2009).

We start with a Bayesian model (details of which are in section 2.3.4) of the FIB states and measurements given covariates $(\theta, Y|X)$ fit to all the data (i.e. covariates X and measurements Y). Then we sample θ at the prediction sites from the posterior distribution $\theta|Y, X$ conditioned on all the data. For convenience we write this posterior as $\tilde{\theta}$ and samples indexed by s as $\tilde{\theta}^{(s)}$. Then we can simply evaluate $L(\tilde{\theta}^{(s)}, \hat{\theta})$. Here, as

in naive and non-parametric validation, $\hat{\theta}$ are RF or NN predictions at the prediction sites based on covariates at those sites and measurements at the sampled sites. By repeatedly sampling $\tilde{\theta}^{(s)}$ and evaluating $L(\tilde{\theta}^{(s)}, \hat{\theta})$ we sample the target distribution $L(\theta, \hat{\theta})|Y, X$.

We emphasize that Y here includes all sites and times. That is, unlike the prediction models which only use observations from the sample sites during the testing period, the Bayesian simulations use measurements from the prediction sites themselves (figure 1). In this way, the Bayesian model can support the validation of other prediction models, but should not be considered a prediction model itself.

Strictly speaking, the Bayesian validation method assumes that the Bayesian validation model is correct. However, it takes into account uncertainty in both model parameters and in observations due to measurement error. Moreover, the method is useful under the weaker assumption that the Bayesian simulations of θ provide a more realistic representation of the true FIB levels against which to compare the prediction models (and infer the target distribution $L(\theta, \hat{\theta})|Y, X$). Additionally, the MSE error metric inferred under Bayesian validation can be validated against the non-parametric approach, which makes fewer assumptions.

Yet compared to the non-parametric validation above, which only estimates MSE, Bayesian validation can be used to estimate any prediction performance metric. We consider several, including MSE, mean absolute error (MAE), and the area under the receiver operating curve (AUC) to evaluate predictions of the continuous FIB level. The remaining metrics (precision, sensitivity, specificity) use binary classifications that 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.

2.3.4 Bayesian validation model to facilitate Bayesian validation

In order to implement Bayesian validation (section 2.3.3) we need a model of $\theta, Y|X$. Again, this model is not used for prediction, but rather uses all measurements (including those at prediction sites) to simulate the true FIB level θ and thus support validation of other prediction models that only use measurements at the sampling sites. We use a linear regression (on the log scale) model with coefficients varying by site:

$$\theta_{jt} = X_{jt}\beta_j + \eta_{jt} \tag{12}$$

where X_{jt} is a vector of K covariates (see section 2.2.2) and for each site j, β_j is a vector of K regression coefficients, and η_{jt} are model structural errors (distinct from measurement errors modeled below) capturing variation in the true FIB level not explained by the linear regression. We use the same covariates as in the RF (see section 2.2.2) 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 RF). 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 structural errors across beaches on a given day t:

$$\eta_t \sim \text{Normal}(0, \Sigma) \tag{13}$$

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 regression:

$$\beta_{jk} \sim \text{Normal}(Z\gamma_k, \sigma^2_{\beta_k}) \tag{14}$$

where Z is a $J \times L$ matrix of site-level covariates, γ_k is a vector of L second-level regression coefficients, and σ_{β_k} is the second-level residual standard deviation. Specifically we use $L = 3$ site level covariates including an intercept, breakwater length, and latitude (which is highly correlated with longitude, see figure 1). This second level regression allows the first level regression (equation 12) between FIB and the observation-level predictors (e.g. precipitation) to vary by site. Moreover, it allows us to partially pool information across sites and incorporate site-level information to more efficiently estimate the coefficients at a given beach (Stow et al., 2009; Cha et al., 2010). Incorporating site-level covariates also supports the (conditional) exchangeability of the sites in the model (Gelman et al., 2013).

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

$$Y_{ijt} \sim \text{Normal}(\theta_{jt}, \tau^2). \tag{15}$$

Note that, unlike the RF 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

$$\Sigma = \text{diag}(\sigma) \cdot \Omega \cdot \text{diag}(\sigma) \tag{16}$$

we put a uniform prior over Ω and a $\text{Cauchy}_+(0, 1)$ prior on the components of σ. The second-level parameters γ_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 fit the Bayesian validation model using the Markov Chain Monte Carlo software Stan (Carpenter et al., 2017), which uses No-U-Turn sampling (Hoffman & 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 $S = 2000$ samples 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

The Bayesian validation model was fit using 13,109 observations at the 19 beaches on 430 days between 2015 and 2019. The posterior distribution of measurement error variance τ^2 had median 0.77 (95% CI, 0.74 to 0.8). This was 30% of the variance of daily means $\bar{Y}_{jt}$ of 2.5. Complete results of the fit are included in the supplementary material.

During the 2019 season 3,780 qPCR measurements were made over 102 days. The Bayesian posterior FIB level θ is displayed in figure 1. We restricted the test

period to those days with two samples at each of the 19 beaches. There were 67 such days.

Using the geometric mean FIB level on each beach-day, the median FIB level was 92 CE/mL 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, 88 to 99 CE) with 3.9% (95% CI, 3.1% to 4.7%) of beach days exceeded the threshold.

We first compare all three validation methods' estimates of MSE since this is the only metric where the non-parametric method is applicable. Then we compare Bayesian and naive estimates of all prediction performance metrics.

3.1 Mean squared error under all validation methods

We start by examining the three validation methods on the metric where they can all be compared, namely MSE. Figure 3(A) 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 prediction models and so yield distributions for the difference in MSE between the two prediction models.

There are four findings here. First, we anticipated that naive validation would give a positively biased estimate of mean-squared error (4) and we see that it does give larger estimates than both non-parametric and Bayesian validation. For both RF and NN prediction models, 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 7) that while naive validation overstates the MSE of both prediction models, the estimated *difference* between the prediction models agrees with the difference given by non-parametric validation.

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 prediction models 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 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 method uses the additional information of covariates and measurements at other sites.

3.2 All performance measures under naive and Bayesian validation

We proceed to evaluate the full set of performance metrics using Bayesian and naive validation methods. We started by using Bayesian validation to estimate the expected sensitivity of the NN prediction model 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 RF was calibrated. Estimates for all prediction performance metrics are shown in figure 3(B).

According to both naive and Bayesian validation, RF outperforms NN in all metrics. However, the discrepancy between Bayesian and naive validation, first documented for MSE in section 3.1 above, continues here across more metrics. Unlike MSE which was systematically overestimated (i.e. pessimistic) using naive validation,

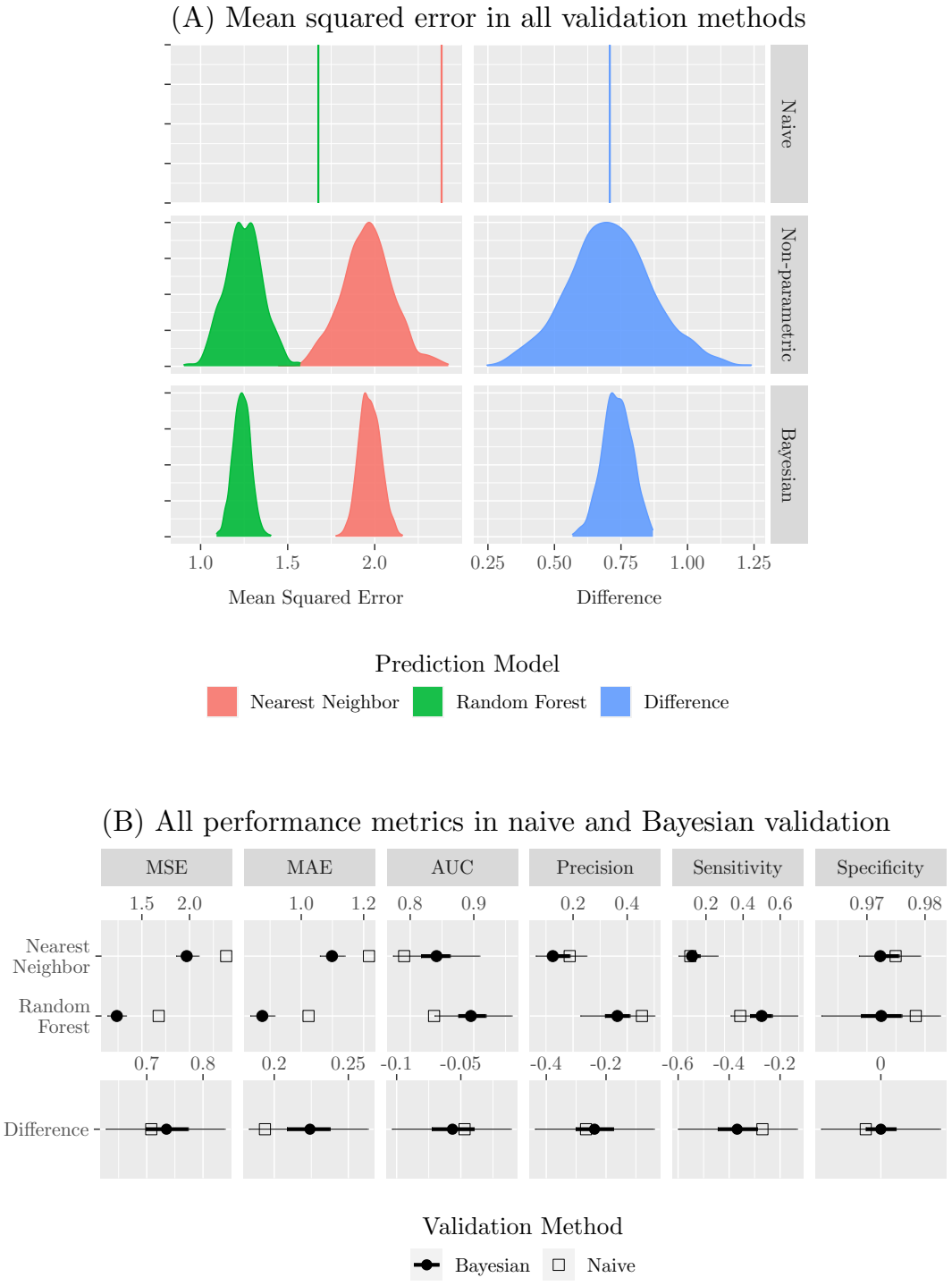

Figure 3: (A) Mean squared error of nearest neighbor and random forest predictions and their difference estimated using naive, non-parametric, and Bayesian validation methods. (B) Prediction performance estimates using naive and Bayesian validation. Solid dots and intervals show median and 50% and 95% credible intervals, respectively, using simulation to account for measurement error. Open squares show naive estimates without accounting for measurement error. Abbreviations: mean squared error (MSE), mean absolute error (MAE), area under receiver operating characteristic curve (AUC).

other measures are variously pessimistic and optimistic, such as when naive validation estimates the precision of the RF to be 0.45 while the Bayesian point estimate is 0.36.

However, as for MSE above, for each metric the bias of naive validation (relative to Bayesian validation) is the same for both RF and NN. Naive estimates of the *difference* in prediction performance benefit from this consistency in that the biases of the difference estimates are not larger than those of the absolute 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 RF to be an improvement of 0.27 while the Bayesian estimate is 0.37.

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 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 capture the efficacy of the predictions for binary management decision making such as issuing a swim advisory or closing a site. Sensitivity for example is the proportion of elevated FIB events that are correctly predicted (and hence acted on), which for the RF is 0.50 (95% CI, 0.33 to 0.69). Precision on the other hand is the proportion of predicted elevated FIB events (hence actions taken) that are correct, which for the RF is 0.36 (95% CI, 0.23 to 0.50).

4 Conclusion

The omission of measurement error, specifically in validation, is ubiquitous in the water resources literature (e.g. Dawson and Wilby (2001); Berenguer et al. (2005); Biondi et al. (2012); Lohani et al. (2012); Shortridge et al. (2016)), including prediction models for recreational water quality (e.g. Nevers and Whitman (2011); Francy (2013); Shively et al. (2016); Lucius et al. (2019)). In this technical note 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 4). 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 contributed two new 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 MSE. 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 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 perfor-

mance 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 assuming an estimated qPCR analysis cost of $30 per sample (Griffith & Weisberg, 2011).

Both the non-parametric and Bayesian approaches to validation proposed in this study help overcome limitations of a naive approach. However, the Bayesian approach to validation is more flexible in its ability to compare models across a variety of metrics, some of which might be particularly relevant to decision-making (e.g., predictions of exceedances over a water quality threshold).

Additional developments could improve the Bayesian model. Temporal autocorrelation was considered but initial testing (not shown) confirmed previous findings that system dynamics are too fast (Dorevitch et al., 2017). The regression coefficients β_{jk} could be modeled as correlated using a multivariate normal distribution. A more sophisticated approach to spatial correlation is also possible, e.g. using an adjacency matrix within a conditional autoregressive model (Gelfand et al., 2010). There is some evidence suggesting that measurement error may vary with the bacteria level (Whitman et al., 2010), which could be modeled using a heteroskedastistic measurement error. While the measurements in our dataset are only at a single location and time for each beach and date, it has been shown that there is substantial variation spatially within each beach and temporally within each day (Whitman & Nevers, 2004; Boehm, 2007). With the relevant data, our model could be extended to these finer scales. In applications where predictions models produce probabilistic forecasts, the Bayesian validation method could be further developed with performance measures that compare the forecast distribution with the Bayesian posterior. These efforts are left for future work. Importantly though, even if the Bayesian model is not the best prediction model of FIB levels (as compared to, say, a machine learning model), it enables us to incorporate all available information into simulations of uncertain bacteria concentrations. This study shows how those simulations can be used to validate prediction models for more realistic assessments of skill compared to a naive approach.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. 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. for transparency about their model and data and assistance in replicating their results. And thanks to the associate editor and reviewers for helpful feedback.

Datasets used in this research are available from the following websites: bacteria test results (`https://data.cityofchicago.org/`); site-specific rainfall, cloud cover, and wind speed data (`https://darksky.net`); and Lake Michigan water levels at Calumet Harbor (`https://tidesandcurrents.noaa.gov/`).

References

Berenguer, M., Corral, C., Sánchez-Diezma, R., & Sempere-Torres, D. (2005). Hydrological validation of a radar-based nowcasting technique. *Journal of Hydrometeorology*, *6*(4), 532–549.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., & Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and

a proposal for a code of practice. *Physics and Chemistry of the Earth, Parts A/B/C*, *42*, 70–76.

Boehm, A. (2007). Enterococci concentrations in diverse coastal environments exhibit extreme variability. *Environmental Science & Technology*, *41*(24), 8227–8232.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Cha, Y., Stow, C. A., Reckhow, K. H., DeMarchi, C., & Johengen, T. H. (2010). Phosphorus load estimation in the Saginaw River, MI using a Bayesian hierarchical/multilevel model. *Water Research*, *44*(10), 3270–3282.

Dawson, C., & Wilby, R. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, *25*(1), 80–108.

Dorevitch, S., Shrestha, A., DeFlorio-Barker, S., Breitenbach, C., & Heimler, I. (2017). Monitoring urban beaches with qPCR vs. culture measures of fecal indicator bacteria: Implications for public notification. *Environmental Health*, *16*(1), 45.

Dotto, C. B. S., Kleidorfer, M., Deletic, A., Rauch, W., & McCarthy, D. T. (2014). Impacts of measured data uncertainty on urban stormwater models. *Journal of Hydrology*, *508*, 28–42.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, *195*(2), 216–222.

Francy, D. 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, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Griffith, J. F., & Weisberg, S. B. (2011). Challenges in implementing new technology for beach water quality monitoring: lessons from a california demonstration project. *Marine Technology Society Journal*, *45*(2), 65–73.

Gronewold, A. D., Borsuk, M., Wolpert, R., & Reckhow, K. (2008). An assessment of fecal indicator bacteria-based water quality standards. *Environmental Science & Technology*, *42*(13), 4676–4682.

Gronewold, A. D., & Borsuk, M. E. (2010). Improving water quality assessments through a hierarchical Bayesian analysis of variability. *Environmental Science & Technology*, *44*(20), 7858–7864.

Gronewold, A. D., Myers, L., Swall, J. L., & Noble, R. T. (2011). Addressing uncertainty in fecal indicator bacteria dark inactivation rates. *Water Research*, *45*(2), 652–664.

Gronewold, A. D., Qian, S. S., Wolpert, R. L., & Reckhow, K. H. (2009). Calibrating and validating bacterial water quality models: A Bayesian approach. *Water Research*, *43*(10), 2688–2698.

Gronewold, A. D., Sobsey, M. D., & McMahan, L. (2017). The compartment bag test (CBT) for enumerating fecal indicator bacteria: basis for design and interpretation of results. *Science of the Total Environment*, *587*, 102–107.

Gronewold, A. D., Stow, C. A., Vijayavel, K., Moynihan, M. A., & Kashian, D. R. (2013). Differentiating enterococcus concentration spatial, temporal, and analytical variability in recreational waters. *Water Research*, *47*(7), 2141–2152.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

ISO. (2009). *Uncertainty of measurement — part 1: Introduction to the expression of uncertainty in measurement* (Vol. 98-1; Tech. Rep.). Geneva, CH.

Kinzelman, J., Ng, C., Jackson, E., Gradus, S., & Bagley, R. (2003). Enterococci as indicators of Lake Michigan recreational water quality: comparison of two methodologies and their impacts on public health regulatory events. *Appl. Environ. Microbiol.*, *69*(1), 92–96.

Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1-2), 161–177.

Leube, P., Geiges, A., & Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, *48*(2).

Liu, X., Lee, J., Kitanidis, P. K., Parker, J., & Kim, U. (2012). Value of information as a context-specific measure of uncertainty in groundwater remediation. *Water Resources Management*, *26*(6), 1513–1535.

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

Lucius, N., Rose, K., Osborn, C., Sweeney, M. E., Chesak, R., Beslow, S., & Schenk Jr, T. (2019). Predicting E. coli concentrations using limited qPCR deployments at Chicago beaches. *Water Research X*, *2*, 100016.

Nevers, M. B., & Whitman, R. L. (2011). Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches. *Water Research*, *45*(4), 1659–1668.

Noble, R. T., Blackwood, A. D., Griffith, J. F., McGee, C. D., & Weisberg, S. B. (2010). Comparison of rapid quantitative PCR-based and conventional culture-based methods for enumeration of Enterococcus spp. and Escherichia coli in recreational waters. *Appl. Environ. Microbiol.*, *76*(22), 7437–7443.

Prüss, A. (1998). Review of epidemiological studies on health effects from exposure to recreational water. *International Journal of Epidemiology*, *27*(1), 1–9.

Rabinovici, S. J., Bernknopf, R. L., Wein, A. M., Coursey, D. L., & Whitman, R. L. (2004). Economic and health risk trade-offs of swim closures at a lake michigan beach. *Environmental Science & Technology*, *38*(10), 2737–2745.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, *47*(11).

Schneider, J., & Moore, A. (2000, February). *A locally weighted learning tutorial using Vizier 1.0* (Tech. Rep. No. CMU-RI-TR-00-18). Pittsburgh, PA: Carnegie Mellon University.

Shively, D. A., Nevers, M. B., Breitenbach, C., Phanikumar, M. S., Przybyla-Kelly, K., Spoljaric, A. M., & Whitman, R. L. (2016). Prototypic automated continuous recreational water quality monitoring of nine Chicago beaches. *Journal of Environmental Management*, *166*, 285–293.

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology & Earth System Sciences*, *20*(7).

Stow, C. A., Lamon, E. C., Qian, S. S., Soranno, P. A., & Reckhow, K. H. (2009). Bayesian hierarchical/multilevel models for inference and prediction using cross-system lake data. In *Real World Ecology* (pp. 111–136). Springer.

U.S. EPA. (2012). *Recreational water quality criteria* (Tech. Rep.).

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources*

Research, *44*(12).

Whitman, R. L., Ge, Z., Nevers, M. B., Boehm, A. B., Chern, E. C., Haugland, R. A., ... others (2010). Relationship and variation of qPCR and culturable Enterococci estimates in ambient surface waters are predictable. *Environmental Science & Technology*, *44*(13), 5049–5054.

Whitman, R. L., & Nevers, M. B. (2004). *Escherichia coli sampling reliability at a frequently closed Chicago beach: monitoring and management implications.* ACS Publications.

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