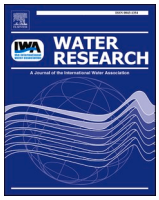
Water Research 252 (2024) 121186

Contents lists available at ScienceDirect

Water Research

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



A new Bayesian approach for managing bathing water quality at river bathing locations vulnerable to short-term pollution

Wolfgang Seis [a,b,*], Marie-Claire Ten Veldhuis [b], Pascale Rouault [a], David Steffelbauer [a], Gertjan Medema [b,c]

[a] KWB Kompetenzzentrum Wasser Berlin gGmbH, Cicerostraße 24, Berlin 10709, Germany
[b] Water Management Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, Delft 2628 CN, the Netherlands
[c] KWR Water Research Institute, Groningenhaven 7, Nieuwegein 3433PE, the Netherlands

ARTICLE INFO

Keywords:
Probabilistic modelling
Recreational waters
Dirichlet Process Mixture Model

ABSTRACT

Short-term fecal pollution events are a major challenge for managing microbial safety at recreational waters. Long turn-over times of current laboratory methods for analyzing fecal indicator bacteria (FIB) delay water quality assessments. Data-driven models have been shown to be valuable approaches to enable fast water quality assessments. However, a major barrier towards the wider use of such models is the prevalent data scarcity at existing bathing waters, which questions the representativeness and thus usefulness of such datasets for model training. The present study explores the ability of five data-driven modelling approaches to predict short-term fecal pollution episodes at recreational bathing locations under data scarce situations and imbalanced datasets. The study explicitly focuses on the potential benefits of adopting an innovative modeling and risk-based assessment approach, based on state/cluster-based Bayesian updating of FIB distributions in relation to different hydrological states. The models are benchmarked against commonly applied supervised learning approaches, particularly linear regression, and random forests, as well as to a zero-model which closely resembles the current way of classifying bathing water quality in the European Union. For model-based clustering we apply a non-parametric Bayesian approach based on a Dirichlet Process Mixture Model. The study tests and demonstrates the proposed approaches at three river bathing locations in Germany, known to be influenced by short-term pollution events. At each river two modelling experiments ("longest dry period", "sequential model training") are performed to explore how the different modelling approaches react and adapt to scarce and uninformative training data, i.e., datasets that do not include event pollution information in terms of elevated *FIB* concentrations. We demonstrate that it is especially the proposed Bayesian approaches that are able to raise correct warnings in such situations (> 90 % true positive rate). The zero-model and random forest are shown to be unable to predict contamination episodes if pollution episodes are not present in the training data. Our research shows that the investigated Bayesian approaches reduce the risk of missed pollution events, thereby improving bathing water safety management. Additionally, the approaches provide a transparent solution for setting minimum data quality requirements under various conditions. The proposed approaches open the way for developing data-driven models for bathing water quality prediction against the reality that data scarcity is common problem at existing and prospective bathing waters.

1. Introduction

The fecal indicator bacteria (FIB), *Escherichia coli* and intestinal enterococci, are the most important water quality parameters for managing microbial safety at recreational waters worldwide. The European Bathing Water Directive (EU-BWD) (2006/7/EC, 2006) requires the collection of at least monthly surveillance samples during the bathing season and annually assesses bathing water quality by calculating the 90^{th} and 95^{th} percentiles based on the data collected over the four previous years. Additionally, the EU-BWD requires the development of early warning systems as a measure for exposure prevention at bathing waters vulnerable to short-term pollution episodes (e.g. from discharges from outlets of combined sewer overflows (CSO)), since it has been experienced that due to the low sampling frequency of regular

* Corresponding author at: KWB Kompetenzzentrum Wasser Berlin gGmbH, Cicerostraße 24, Berlin 10709, Germany.
E-mail address: wolfgang.seis@kompetenz-wasser.de (W. Seis).

https://doi.org/10.1016/j.watres.2024.121186
Received 6 November 2023; Received in revised form 21 December 2023; Accepted 23 January 2024
Available online 25 January 2024

surveillance monitoring, pollution episodes caused by storm events are rarely detected (Kay et al., 2005). Moreover, the turn-over times of standard laboratory methods between 24 and 48 h from sampling to results are too long to timely inform swimmers about impaired water quality.

In this context, data-driven models based on readily available environmental data, like rainfall and flow data, have been shown to be suitable approaches for predicting short-term pollution episodes (Francy et al., 2020). Previously published literature applied different modeling approaches, like multiple linear regression, generalized least squares (GLS), logistic regression, random forests (RF), support vector machines, and artificial neural networks (ANN) (Francy, 2009; Francy et al., 2020; Mälzer et al., 2016; Searcy and Boehm, 2021; Thoe et al., 2015). While algorithms differ, many models share general input parameters, especially rainfall. Moreover, all mentioned modeling approaches are based on the principle of supervised learning. That means that predictions are based on models and parameters, which summarize the information contained in the training data about the relation of the outcome variable (FIB concentration) and selected predictor variables. Thus, the quality of the predictions highly depends on a representative dataset for FIB used for model training. This poses a major limitation on the wider use of such modeling approaches, since rich historical FIB datasets, for training advanced machine learning algorithms are often missing, given the low sampling frequencies of routine FIB surveillance monitoring, like for example according to the EU-BWD. This aspect is especially relevant for regions with prolonged dry periods and only sporadically occurring but intense rain events, which might become more pronounced against progressive climate change. Such circumstances may result in highly imbalanced datasets, meaning datasets with no or only very low numbers of high-concentration observations for FIB.

Such unbalanced datasets pose a challenge for training statistical models since the training data do not provide the necessary information to learn the relationships required to make accurate predictions regarding peak events. In such situations, collecting event-based samples, explicitly targeting rain weather conditions, may be a suitable solution but requires resource intensive sampling setups, with flexible equipment and personal standing-by and reacting to rain events. Against this background, the World Health Organisation (WHO) identified the minimum amount of datapoints and appropriate sampling strategies necessary for model training as a key research need for managing bathing waters (WHO, 2018).

There are only a few studies which directly address the problem of finding a practical approach towards the development of data-driven models in data scarce situations for recreational waters. A study by Searcy and Boehm (2021) tried to address the problem by collecting samples at high frequencies (< 1 h) over a limited period (1-2 days) to collect enough data to train various machine learning models (RF, ANN, GLS). While the authors showed that the models trained with high frequency data improved bathing water quality management under dry weather conditions in comparison to the "common method", they underline that the lack of rain weather conditions might negatively influence predictions, if rain events occur, making targeted rain weather sampling again necessary.

The fact that we know (a.) that the presence of pollution sources, like combined sewer overflows, close to a bathing site will potentially cause rainfall induced pollution, and that we further know that (b.) infrequent statutory monitoring of FIB will not necessarily detect these events, raise doubts about the general use of supervised data-driven models which solely rely on paired data between outcome variable and the selected covariates.

Against this background, the present study compares different data-driven modeling approaches to predict periods of high risk of fecal indicator bacteria in recreational waters. We especially focus on the ability of two state-based Bayesian approaches, in comparison to previously used approaches (quantile random forest, multiple linear regression, intercept only) to handle situations where prolonged dry periods lead to a lack of informative training data regarding rainfall-induced pollution episodes. We propose incremental validation by state-based Bayesian updating of a weakly informative, precautionary prior distribution as a potential method to cope with data scarcity and data imbalance.

2. Method

The study proposes new Bayesian methods to support risk-based bathing water quality management and compares it to existing data-driven modeling techniques. Special attention is given to the ability each modeling approach to correctly predict pollution episodes even if the training data lacks information about event-scale variability of FIB due to prolonged dry periods or non-detected pollution episodes.

2.1. Datasets for modelling experiments

2.1.1. FIB data

For model comparison, we used datasets of *E. coli* from three existing or prospective river bathing locations. *E. coli* was chosen since it is the major cause of threshold violations at the selected rivers in comparison to intestinal enterococci. All three locations are known to be vulnerable to short-term pollution episodes, as outlets from the combined sewer network (CSOs) are located upstream or discharge in direct vicinity of the bathing site. All bathing sites were subject to previous research activities. Therefore, comparably rich datasets with $N = 281$ (*River I*), $N = 1191$ (*River II*) and $N = 251$ (*River III*) are available for *E. coli*. *E. coli* data were analyzed using standard laboratory methods according to ISO 9308-3 applied by an accredited laboratory. Table 1 summarizes the sampling periods and setups at the different rivers. More detailed information is provided in the SI.

For the modeling experiments, we generated experimental datasets by applying several assumptions to the available data:

1. Results from composite sampling were treated as if they were collected using grab sampling.
2. If multiple samples were collected on the same day, data were treated as independent observations for model training.
3. Values below or above the LOQ were set to the lower or upper LOQ, respectively.

2.1.2. Hydraulic predictor variables

Daily average stream flow data were taken from the closest flow station of each bathing site. Cumulative daily rainfall data was used as spatial averages of rain stations upstream of the river bathing site covering the relevant area of the combined sewer systems upstream of each bathing site (see SI). Temporal lag variables were created by averaging rainfall and flow data over multiple days prior to sampling following Cyterski et al. (2012). As potential predictor variables the averages from 1 to 7 days prior to the *FIB* sampling date were created, i.e. [1-2 days, 1–3 days … 1–7 days]. Moreover, the data from each individual day [1…7] were included.

Table 1
Overview of available dataset at the three different rivers.

River	Total	Daily grab sample (surveillance)	Event based composite samples	Event-based hourly grab samples
River I (2014–2017)	281	281	0	0
River II (2010–2019)	1191	147 (2010–2019)	267 (2016–2019)	777 (2018, 2019)
River III (2018–2021)	251	116 (2018–2021)	135 (2018–2019)	0

2.2. Modeling experiments

To investigate the response of the modeling approaches to differences in input data and how they cope with situations of longer periods of non-informative, low FIB concentrations, we conducted two modeling experiments, namely *"longest period of low FIB"* (cf. Section 2.2.1), and *"sequential model training"*(cf. Section 2.2.2).

2.2.1. Training on "low" FIB concentrations

For the modeling experiment "*longest period of low FIB"* we intentionally trained the different models (cf. Section 2.3) with a subset of data, which represents the longest period of days indicating suitable bathing water quality. Thus, the training data did not include any confirmed pollution episodes, yet the models should be able to detect them in the test set. We chose this approach as realistic scenario, where longer dry periods led to data series which do not indicate any sign of serious contamination. As a threshold for splitting the data we used 500 MPN / 100 mL, which is the legal threshold (95th percentile) for "excellent" water quality for freshwater in Europe. The periods and number of data-points are summarized in Table 2. Note, that the EU-BWD demands a minimum of 16 samples per bathing season, which is exceeded in all the training scenarios.

2.2.2. Sequential training and evaluation

During the modeling experiment "*sequential training*", we focused on how different model performance indicators (cf. Section 2.4) change depending on different proportions of the data used for model training. To this end we trained the selected models with an incrementally increasing proportion of available data. For *Rivers I* and *III* we increased the *training ratio* in 13 steps from 10 % to 70 % by 5 % increments. At *River II* we increased the training ratio also in 13 steps from 4 % to 52 % using 4 % increments, due to the larger dataset at *River II*. Model predictions were calculated for the remaining dataset as described in Section 2.3. For all three *Rivers X,* we performed the modeling experiment as observed in forward chronological order. Moreover, we created scenarios ("reverse") in which we keep the chronological order of the lag variables in relation to the FIB data (e.g. rainfall → FIB) but assume that the last observation would have been observed first. The reason for this is, that prolonged dry periods at *River II* and *III* occurred at the end of the sampling period. Technically, we basically reversed the indices 1…n, after the lag variables were created. Thereby, we construced scenarios which first only include dry periods data, while informative data on rainfall become incrementally available.

2.3. Modeling approaches

The selected models are divided into two groups. The focus lies on the first group, which separates the rivers into different hydrologic states and applies a Bayesian updating procedure based on the data collected under each individual hydrologic state (cf. Section 2.3.1). The second group of models consists of supervised regression modeling approaches, which either autonomously select/weigh predictor variables, or which do not include any predictor variables at all (cf. Section 2.3.2). The latter closely resembles the currently way of long-term bathing water classification in Europe.

Table 2
Overview of the training and test data used for modeling experiment "longest period of low FIB concentration".

River	Longest period of "low" measurements < 500 MPN/ 100mL	Number of Sampling days / Number of training samples	Number of Test days / Number test samples
River I	2014-09-03 – 2015-06-22	25 / 25	257 / 257
River II	2016-06-03 – 2016-08-16	25 / 42	213 / 1149
River III	2020-06-16 – 2020-08-26	44 / 52	113 /199

2.3.1. Bayesian state-based models

In Bayesian statistics, observed data y are used to update the prior distributions $p(\theta)$ of the model parameters of the likelihood $p(y|\theta)$, to estimate the posterior distribution $p(\theta|y)$, using Bayes law (Gelman et al., 2014):

$$p(\theta|y) \propto p(y|\theta)\ p(\theta) \tag{1}$$

The latter represents the conditional probability of the model parameters given the data. If the prior distributions are proper, i.e., if they integrate to 1, it is possible to simulate from the prior predictive distribution (PriorPD), which represents the initial uncertainty about the expected distribution of possible outcomes before the data was collected. The two investigated Bayesian modeling approaches (*manual state setting, Dirichlet process clustering*) exploit these mechanisms by sequentially updating a precautionary prior distribution for FIB data collected under a set of predefined (*manual state setting*) or algorithmically identified (*Dirichlet process clustering*) states. The two approaches follow a common general approach. The steps are described in more detail in the referenced sections and are illustrated in Figs. 1 and 2.

For model calibration / training we:

1. used the available hydrological information to define distinct hydrologic states K, defined by river flow and rainfall (cf. Sections 2.3.1.1 and 2.3.1.2),
2. assigned the same weakly informative precautionary prior distribution for a $\log_{10}$-normal likelihood of FIB for each created state K (cf. Section 2.3.1.3),
3. assigned the collected FIB data to the hydrologic states at which they were collected, and
4. updated the state specific prior distribution independently with the FIB data collected at the associated state to obtain state-specific posterior distributions (cf. Section 2.3.1.4).

For prediction:

1. For each new prediction case, we identified the associated hydrological state K based on the hydrological information.
2. If the prior distribution of the FIB data had been updated during model training, we used the state specific posterior predictive distribution (PostPD) as the predictive distribution for new FIB data. If the FIB distribution was not updated, we used the PriorPD.

For bathing water quality assessment, we compared the daily predicted 90th percentile of the used predictive FIB distribution against the percentile threshold of 900 MPN / 100 mL for "poor" bathing water quality as a decision point for raising warnings due to elevated risk or not. While the conceptual structure is similar, the two approaches differ in how they determine hydrological states:

1. The first approach defines hydrological states manually based on historical hydraulic information (*manual state setting*, Section 2.3.1.1).
2. The second approach defines hydrological states using a Dirichlet-Process Gaussian Mixture Model (DPMM) for model-based clustering. The approach only relies on the hydrological information provided in the training dataset and thus, does not rely on longer term hydrological information (*Dirichlet process clustering,* Section 2.3.1.4).

2.3.1.1. Manual state setting. In the first approach, the states K were defined manually, based on availably historical hydrological data

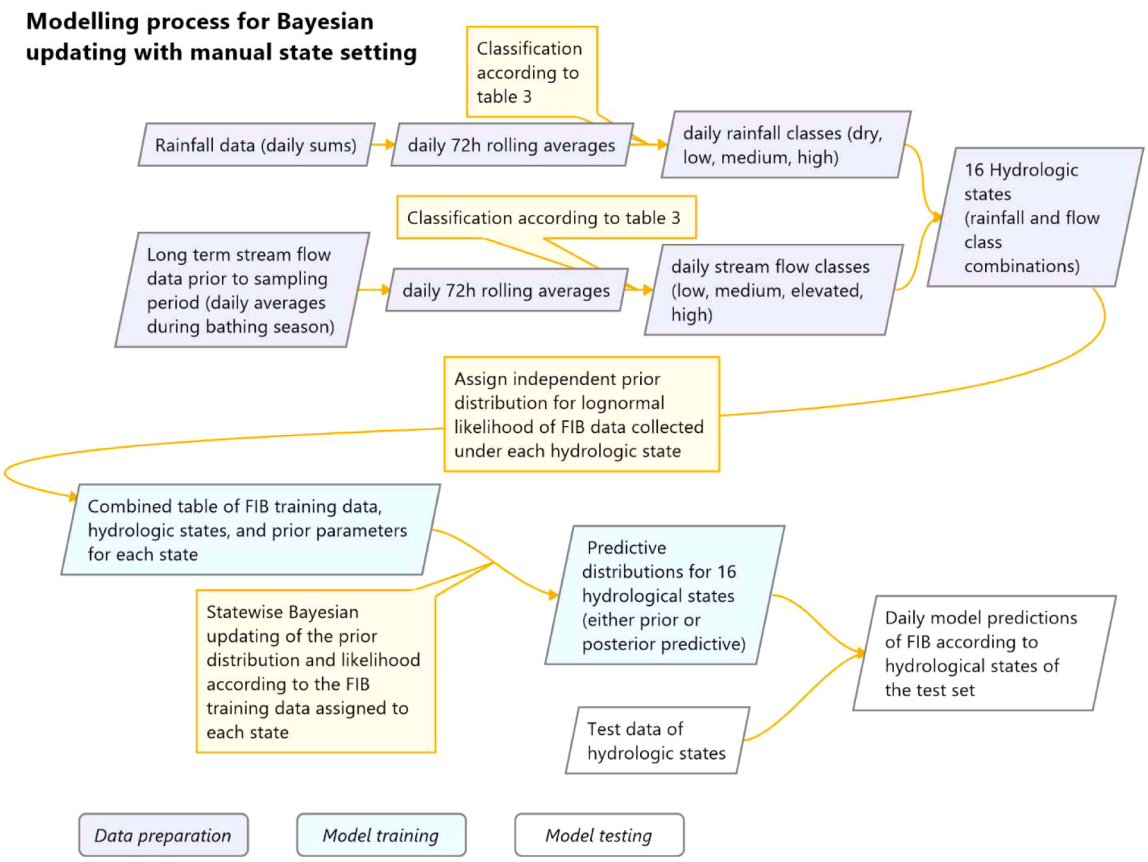


Fig. 1. Overview of the modelling process according to Bayesian updating with manual state setting.

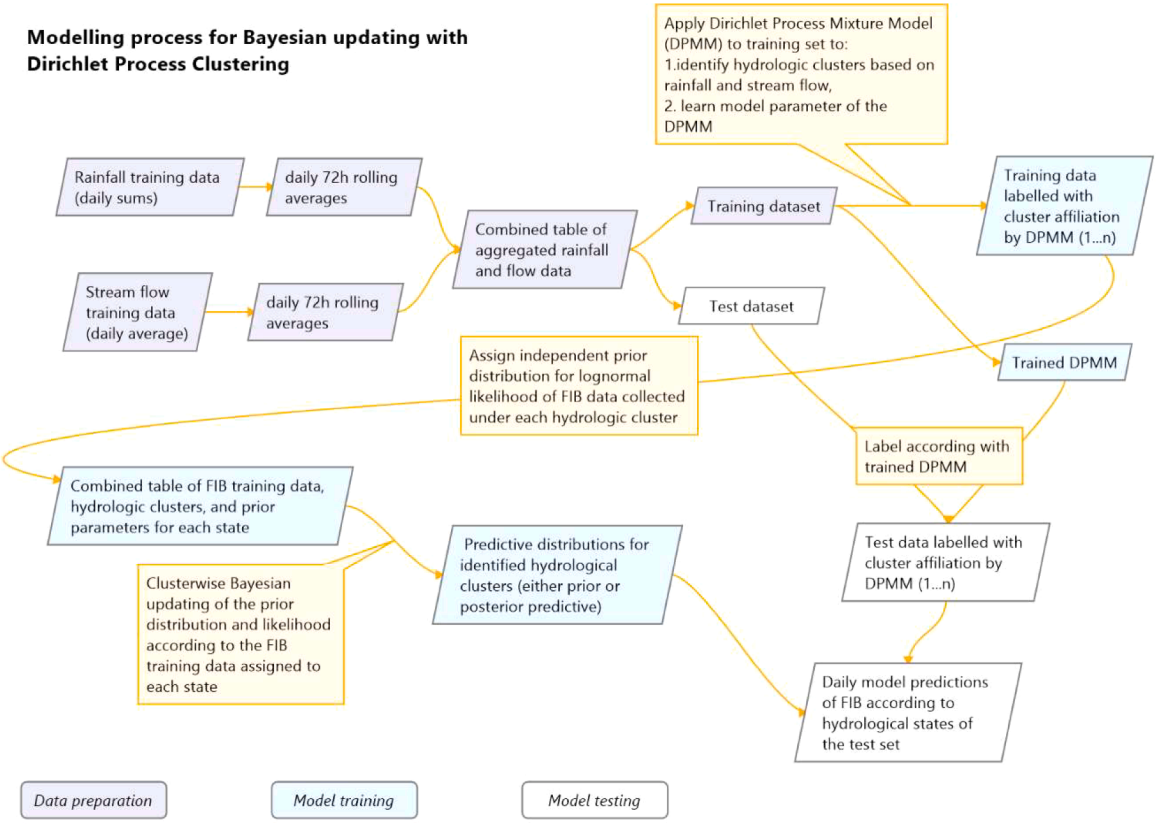


Fig. 2. Modelling process according to Bayesian updating with Dirichlet process clustering.

Table 3
Rainfall and flow category criteria for manual state setting.

Rainfall (avg. 72 h sum)	Rain category	Flow (avg.72 h)	Flow category
Rainfall [0 mm, 1 mm]	Dry	Flow (<25 %)	Low
Rainfall]1 mm, 5 mm]	Medium	Flow (25–50 %)	medium
Rainfall]5 mm, 10 mm]	Elevated	Flow (50–75 %)	elevated
Rainfall]10 mm, ∞[	High	Flow (>75 %)	High

(Table 3). For categorization, rainfall and flow data were averaged over 72 h before FIB sampling, since Cyterski et al. (2012) showed that temporal synchronization of up to 3-day lags led to better predictions of FIB. While this time window could be optimized in data rich situations depending on the watershed, averaging over a window of 72 h covers the most recent rain-events, which were considered most relevant for short-term peak contaminations. Historical flow data were used up to the year where the first FIB sampling occurred. For defining flow categories, daily flow data during the bathing season (May-September) were evaluated. From these data the 25th, 50th and 75th percentiles calculated. Flow categories were created as outlined in Table 3. For rainfall, 5 mm increments were used (see Table 3). While this categorization is arbitrary, the discretization into 5 mm increments of continuous rainfall data is considered intuitive and is also a common format provided for example by meteorological forecast services (e.g., German Weather Service MOSMIX forecast).

Each individual combination of rainfall and flow category defined a specific hydrological state. Unusual combinations like "rainfall = high" and "river flow = low" were kept in the model, since some rivers may be highly regulated and thus hydrologic dependencies may be disconnected. Keeping an "empty" state in the model, i.e., a state that is not populated by any data, is acceptable since it will not influence the modelling process.

2.3.1.2. Dirichlet Process Mixture Model. In a second approach, hydrologic states were derived by applying a Dirichlet Process Gaussian Mixture Model (DPMM) for clustering the hydrologic data. The DPMM is a non-parametric Bayesian model, which clusters the available data into clusters of pre-defined base distributions (here: multivariate Gaussians). What makes the DPMM special, in comparison to many other clustering algorithms, is that the DPMM identifies the number of clusters algorithmically from the provided dataset, with the number of clusters increasing with the number of datapoints. Thereby, model complexity increases with the number of available data. During training and testing, the DPMM estimates whether a new observation belongs to an existing cluster or should be considered a new cluster since the new observation deviates strongly from the data in any existing cluster. In the present study we made use of this behavior to first, calibrate the number of clusters and cluster parameters in the training set. Secondly, we identified whether a specific observation in the test set belongs to a hydraulic cluster already identified during model training or whether it constitutes a new, i.e. unobserved situation. In the latter case, the PriorPD for *E. coli* is used as the predictive distribution as the cluster has not been updated during model training. There are multiple ways to describe and construct a DPMM. For detailed explanations, please see e.g. Teh (2011), Li et al. (2019).

As for the manual state-setting approach, we only used two predictor variables, namely the average flow and rainfall averaged over 3 days before FIB sampling, which led to clusters, consisting of 2-dimensional multivariate normal distributions. In contrast to the manual state stetting approach, we only used the hydrological data associated with the training data set, i.e., no historical data was used. The hydrological training and test data were standardized by subtracting the mean and dividing by the standard deviation of the training set. For fitting the DPMM we used the R package *dirichletprocess* (Ross and Markwick, 2023), which uses collapsed Gibbs sampling for Markov-Chain-Monte Carlo for fitting the model. For each chain we simulated 10,000 iterations, with a burn-in phase of 5000 iterations. Model-convergence was assessed by checking whether the Gelman-Rubin statistics $\widehat{R}$ for each data point's cluster assignment chain had converged to 1. As final cluster assignment we used the posterior mode of assigned cluster labels for each data point. For predicting the cluster assignment of any new observation, we used the function *clusterLabelpredict* provided by the *dirichletprocess* package. The function generates a single draw from the posterior distribution $p\ (z_i \mid z_{i\text{-}1}, \theta)$, where z_i describes the cluster assignment of a new observation i, given the cluster assignments of the remaining data $z_{i\text{-}1}$ and the parameter vector θ, which represents the cluster parameters of the existing clusters. For obtaining the most likely cluster assignment we used the most frequently occurring cluster assignment generated by 1000 repetitions.

2.3.1.3. Distributional assumptions for FIB data. For each hydraulic cluster/state K, an independent $\log_{10}$-normal likelihood was assumed for the *E. coli* concentration. The $\log_{10}$-normal likelihood was used because the EU-BWD bases bathing water criteria on the percentiles of a $\log_{10}$-normal model. A $\log_{10}$-normal likelihood is defined by the geometric mean and standard deviation. For each state the same precautionary and weakly informative prior distributions (Eqs. (2)–(4)) were used. We preferred a lognormal prior for the standard deviation over the often-used conjugate inverse-gamma, because of the ability to define location and scale parameters separately, without further re-parameterization. Both priors are very vague, claiming that the geometric mean lies between 10 and 10^5 *E. coli*/100 mL, and the residual standard deviation lies between 0.35 and 2.7 (95 %) on a $\log_{10}$-scale, allowing for additional variations of multiple orders of magnitude. Thereby, these priors ensure that future observations will fall inside the PriorPD, as the PriorPD covers the whole measurement scale of common methods for bathing water quality surveillance. Thus, they essentially provide the weak information that pollution episodes should be considered a possible event unless data demonstrate otherwise.

$$\lg(E.coli)_{i,\ K} \sim Normal(\mu_k, \sigma_k) \tag{2}$$

$$\mu_k \sim Normal(3,\ 1) \tag{3}$$

$$\sigma_K \sim Lognormal(0,\ 0.5) \tag{4}$$

In Eqs. (2)–(4) $E.coli_k$, μ_k and σ_k refer to the *E. coli* data i and its distributional parameters, collected under hydrologic state K, and the related parameters of the $\log_{10}$-normal likelihood.

2.3.1.4. Applying Bayesian updating and prediction. For updating the prior distribution of FIB data, we used the software package *brms* (Bürkner, 2017), which is an interface between the programming languages R (R Development Core Team, 2008), and Stan (StanDevelopmentTeam, 2017). Each state, defined either by manual state setting or clustering, is updated separately. Markov chains ran for 10, 000 iterations using Hamiltonian Monte Carlo for Markov Chain Monte Carlo. Model convergence was checked by checking whether the Gelman-Rubin statistics converged to a value of 1. Samples from the PostPD were generated using the function *posterior predict* of the *brms* package from which the 10th, 50th and 90th quantiles were calculated (cf. Section 2.4).

2.3.2. Supervised regression-based approaches

As benchmark models we used three different types of regression approaches, a "zero-model", "stepwise regression", and "quantile random forest".

Zero-model: The zero-model does not consider any predictor variables but only infers the geometric mean and geometric standard deviation from the available FIB data in the training set. We included this approach as it is similar to the approach applied in the EU-BWD, even though we use all available data instead of only data collected over the

four previous years. For prediction, we calculated the 90th percentile, by:

$$90^{th}\ percentitle = m + 1.282s$$

where m is the geometric mean and s is the geometric standard deviation.

Stepwise regression: Multivariate linear regression is the most commomly applied method for bathing water quality management (Francy et al., 2020). Thus, we applied a stepwise regression approach for variable selection. In stepwise regression, predictor variables are added and removed stepwise from the set of available predictor variables, keeping only those variables for which a significant improvement, i.e. a reduction in the *Akaike-Information Criterion (AIC)* can be observed. To this end, we applied the function *stepAIC* provided by the R-package "*MASS*". Model predictions of new data were simulated by using the function *predict,* using a 80% confidence level to derive estimates of the 10^{th}, 50^{th}, and 90^{th} percentiles.

Quantile random forest: QRF have been published by Meinshausen (2006). QRF, are a generalization of the classical random forest (RF) published by Breiman (2001). QRF extend RF by the ability to introduce uncertainty into model predictions by predicting quantiles of the expected range of future values. The model is trained with all available predictor variables. Variable "selection" happens implicitly as different predictor variables are weighted according to their predictive power. In contrast to linear regression, QRF represents an ensemble method (ensemble of decision trees), which makes no distributional assumption. Therefore, the estimated percentiles are "non-parametric", which differs from the estimates provided by the other regression models. For the implementation of QRF, we used the function *ranger* of the r-package *ranger*. Tree-based ensemble methods have been at the top with regard to predictive accuracy in previous studies (e.g. Searcy and Boehm 2021), and were therefore included in the present study. Predictions are simulated by using the function *predict* to calculate the 10^{th}, 50^{th}, and 90^{th} percentiles.

2.4. Model evaluation and assessment

Data driven models do not necessarily aim at providing a true representation of the physical world by explicitly modelling their underlying physical processes, but rather justify their use by the utility they provide for prediction and decision-support. Validating model predictions against the probabilistic character of the EU-BWD is challenging, since it cannot be validated by a single reference value. To evaluate and compare quality and utility of the different models we used a combination of "model-related" (cf. Section 2.4.1) and "decision-related" (cf. Section 2.4.2) metrics.

2.4.1. Model-related metrics

As model-related metrics we calculated the explained variance R^2 for *E. coli* both "in-sample" for the training data and "out-of-sample" for the test data. Additionally, we calculated the out-of-sample root-mean-squared error (RMSE) to assess the average distance between the predicted geometric mean and the test data. As probabilistic metrics we calculated the percentage coverage rates of the 90^{th} percentile and the 80 % prediction interval. Ideally, 90 % and 80 % of the test data fall below and inside the chosen test-intervals, respectively.

2.4.2. Decision-related metrics

Since threshold levels in the EU-BWD are defined in terms of percentile values, we base the decision-making on whether swimming is advised or not on the predicted 90^{th} percentile of the predictive uncertainty intervals of the different modelling approaches. If the predicted percentile is larger than 900 MPN/100 mL bathing is "not advised" otherwise bathing is considered "safe". As a single sample threshold, we used a value of 1800 MPN/100 mL as an incidence of "confirmed pollution", following the single sample threshold applied in most of Germany. It follows the rational that if the "true" 90^{th} percentile would be below 900 MPN/100 mL, a value of >1800 MPN/100 mL is very unlikely, and thus can be regarded as sufficient evidence that bathing water quality standard are violated. The number of correctly and incorrectly predicted confirmed pollution events were considered "true positives", and "false negatives", respectively.

3. Results

3.1. Training on low FIB concentrations

A comparison of the different performance indicators for the five models at the three river sites is shown in Fig. 3. For *River I* the Bayesian approaches achieve true positive values of 97.5 % (39 / 40) and bathing rates of 30 % (61 / 257) (*manual*), and 36 % (88 / 257) (*dirichlet*); for *River II* true positive rates of 96 % (24/25) and bathing rates of 48 % (103/213) (*manual*), and 64 % (136/213) (*dirichlet*); for *River III,* the approach with manual-state setting performs slightly better than the one using the DPMM with a true positive rate 82 % (9 / 11) and a bathing rate of 61 %, in comparison to 72 % (8 / 11) and bathing rate of 43 %. Note, that due to the lower number of pollution episodes at *River III*, percentage values are more sensitive to small differences. The zero-model and the QRF model do not predict any of the confirmed pollution episodes correctly, and consistently achieve true positive rates of 0 % at all three rivers, while indicating a bathing rate of 100 %, indicating the false assumption of permanent low risk. The results underline the inability of these modelling approaches to predict high values in the response variable (*E. coli*) if the training data does not include this information.

The stepwise regression model achieves true positive rates of 78 % (31 / 40), and 82 % (9 / 11) at *Rivers I,* and *III*, respectively, but only predicts 44 % (11 / 25) at *River II*. These results indicate that the datasets at *River I* and *III*, already included a positive correlation between the selected predictors and *E. coli* even in the narrow range of *E. coli* < 500 MPN/100 mL (*River I:* flow averaged over the past 24 h, rainfall averaged over 7 days before sampling, *River III:* rainfall variables averaging over 24 h and 72 h before sampling). If extrapolated, the existing correlation allows the model to achieve a noticeable rate of correctly predicted pollution episodes. However, the fact that the model performs worse at *River II* underlines that this behavior is not robust against different situations. Moreover, while extrapolating a linear model is possible in principle, it is generally not advised, since an observed linear relationship may not persist as the model extends beyond its calibration range. For stepwise regression there is no relation between the in-sample R^2 and the out-of-sample R^2, or the RMSE. At *River III* an in-sample R^2 of 0.16 in the training set, leads to the highest out-of-sample R^2 of 0.44 and the lowest out-of-sample RMSE. Indeed, the high true positive rate and high bathing rate indicate that despite the low in-sample R^2 the model is of high practical value. On the other hand, an in-sample R^2 of 0.54 at *River II*, leads to an out-of-sample R^2 of 0.07 at *River II.* This shows that the in-sample R^2 is not a suitable metric to assess the predictive accuracy of a given model.

Regarding the percentage coverage criteria that 80 % of test data should fall within the 80 % prediction interval and 90 % below 90^{th} percentile, only the two Bayesian approaches consistently achieve the percentage coverage rates close to the target values of 80 and 90 % respectively. This is achieved by the precautionary prior distribution, which leads to high uncertainty intervals for "unpopulated" states / clusters, which cover even the highest observations.

Figs. 4 and 5 allow for additional insights regarding the updating mechanisms of the applied Bayesian modeling / management approaches. First, focusing on the four flow categories, where *"Rain = dry"* at *River I.* At its current training state the model indicates low risk for medium and elevated flow conditions, whereas it raises warnings for low and high flow conditions. This is because for low and high flow

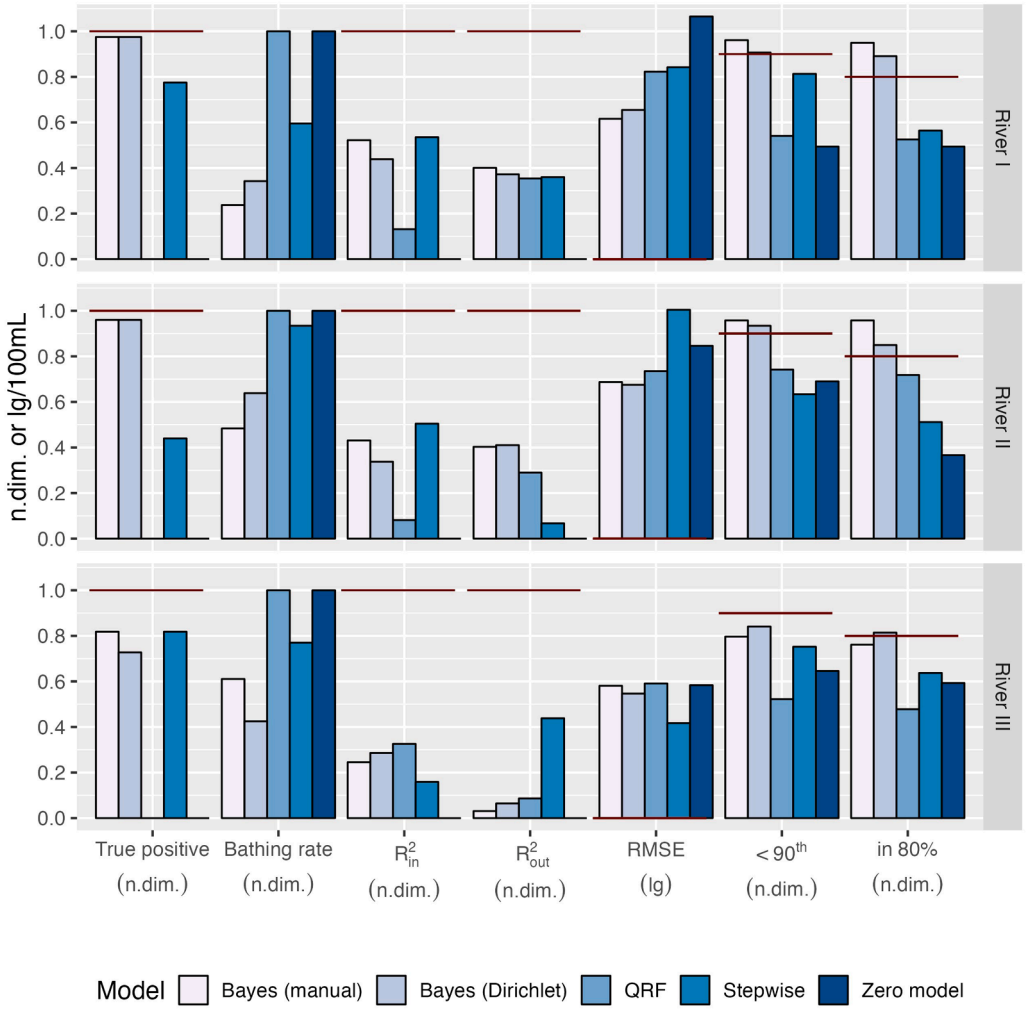


Fig. 3. Comparison of performance indicators for the different modelling approaches. Target values are visualized by horizontal lines. n.dim.: non dimensional metrics (y-axis), lg: metric on measurements scale lg/100 mL. Performance metrics from left to right: True positive rate (target: 1), Predicted bathing rate (no target), in-sample R^2 (target: 1), out of sample R^2 (target: 1), root mean squared error for FIB concentrations (target: 0), ratio of predicted data points below 90th percentile (target: 0.9), ratio of data points within 80 % prediction interval for FIB concentrations (target: 0.8).

conditions the training data provide only a single observation (grey points). Thus, the remaining uncertainty expressed by the PostPD is still high, and the predicted 90th percentile exceeds the threshold levels of 900 MPN/100 mL. The figure further shows that the white test data (which are unknown during model training) indicate that swimming would most likely be possible. Thus, at its current training state the precautionary warnings, which the model raises in the cases of low and high flow conditions are *false alarms*. However, in real applications this would not be known as the test data would not be available, yet. While *false alarms* are generally unwanted, a model evaluation as presented in Fig. 4 clearly identifies the lack of validation data a potential reason for these "alarms" and allows for adapting future sample collections to these situations. Focusing on category *"Rain = elevated, Flow = medium"* at *River I* shows that such precautionary warnings may also be justified. Again, a single low measurement is not able to reduce the predictive uncertainty of the prior enough to expect "low risk". However, in this specific case, the test data show that 4/8 (50 %) data points exceed the single sample threshold of 1800 MPN/100 mL. Thus, while raising precautionary warning due to limited data availability may lead to *false alarms (false negatives)* it also leads to correct warnings *(true positives)*. Overall, the method follows a consistent and transparent assessment approach.

The comparison of the two Bayesian modeling approaches shows that the model based *Dirichlet* approach led to a lower number of populated clusters for *River I* and *II*, and a slightly higher amount for *River III* in comparison to *manual bayes* approach. This underlines that even though the DPMM tends to identify an increasing number of clusters with an increasing number of datapoints, the number of clusters, keeps within reasonable limits.

3.2. *Sequential fitting*

Results from reverse modeling experiments are shown in Fig. 6. The results from forward modeling are provided in the SI. The datasets from forward modeling are relatively balanced and indicate similar outcomes as reverse modelling for *River I,* i.e. the case of balanced datasets. Briefly, the QRF and manual Bayesian approaches achieve overall the most reliable predictions of pollution episodes, while the manual Bayesian approach behaves more conservative regarding the number of identified bathing days in the test set for low training ratios compared to the QRF model.

For the reverse modelling scenarios, the dataset from *River I* is relatively balanced even for low training ratios (0.05–0.2). In contrast, the datasets from *Rivers II* and *III* are characterized by prolonged dry periods at the beginning (Training ratios < 0.4) of the dataset. For the balanced dataset at *River I*, all models achieve high true positive rates between 90 and 100 %. A special case is the zero-model which consistently achieves a true positive rate of 100 % but does this at the costs of a

Fig. 4. Model analysis for manual state setting. Solid error bars: 80 % prediction interval of the PosteriorPD derived for each state. Dotted error bars: PriorPD used for cases where the incoming hydraulic data has no support in the training set. Error bars color: indicator whether the 90th percentile exceeds a value of 900 MPN/100 mL and risk is predicted. Red vertical dashed lines: single sample threshold for confirmed contamination of 1800 MPN / 100 mL. Black vertical dashed line: model decision threshold of 900 MPN/100 mL.

bathing rate of 0 %, making swimming impossible. The Bayesian Dirichlet, stepwise and QRF approaches achieve bathing rate of $>$ 50 % from the beginning, while the manual Bayesian approach shows an increasing tendency, due to the "states" which must be validated incrementally. The lowest RMSE can be observed for the QRF model, which also achieves "true positive rates" of larger than 95 % for training ratios above 0.15 and similar bathing rates as the manual Bayesian approach. The manual Bayesian approach leads to comparable results at training ration above 0.2 but provides a more precautionary approach at low training ratios $<$ 0.2, where it predicts 100 % of confirmed pollution in contrast to the QRF approach.

At *Rivers II* and *III* the differences between modeling approaches are more pronounced. At both rivers, the training data contain mostly dry weather information up to a training ratio between 0.34 and 0.4. Below ratios of 0.25 only the two Bayesian approaches (manual and DPMM) lead to prediction rates of 100 % for river II and $>$ 90 % for river III. From the remaining modeling approaches the zero-model and the QRF predict no relevant percentage of the pollution episodes correctly, while indicating high bathing rates (100 % for the zero-model). In such situations the use of such models, thus, produces a false indication of safety, potentially leading to acute health risk. The stepwise approach leads to highly fluctuating results at *River II* and to similar results as the Bayesian approaches for the true positive rate at *River III* above a training ration of 0.25. Only after information about rainfall induced pollution becomes available at training ratios of 0.34 (*River II*) and 0.4 (*River III*) the QRF can reliably predict pollution episodes. In both locations QRF and stepwise regression react faster to the new information than the zero-model, which changes from a prediction rate of 0 % to 100 % only at a training rate of 0.6 for *River III* and stays at 0 % for *River II*. The results of the sequential learning experiment confirm the finding that the two Bayesian approaches reliably detect pollution episodes against the background of extended dry periods with no observed contaminations in the training set. An additional insight is that the poor true positive rate of the QRF and zero-modeling approach may persist even when there is a high number of training data available ($>$ 400 for *River II*, $>$ 100 for *River III)*.

4. Discussion

In the present study, we placed a weakly informative precautionary prior distribution on the parameters of the lognormal likelihood of the concentration of FIB at different hydrological states. These hydrological states are either pre-defined manually or identified algorithmically. First, each state is empty and gets incrementally updated as more FIB data become available. Thereby it is ensured that bathing water quality is independently validated for each defined state. If the training data do not provide sufficient evidence for indicating acceptable water quality, precautionary warnings are raised. While the general procedure of Bayesian updating of a normal likelihood distribution can be regarded as straightforward, its application for sequentially validating bathing water

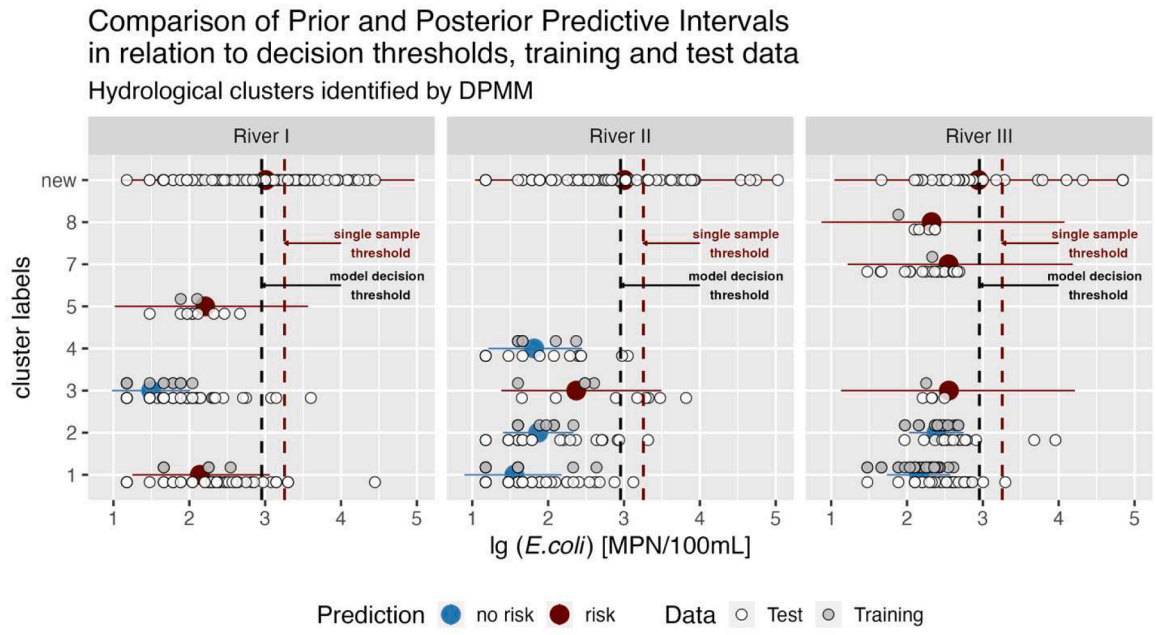


Fig. 5. Model analysis of the DPGMM for model-based clustering. Error bars: 80 % prediction interval of the PostPD derived for each state. Error bar labeled "new": PriorPD used for cases where the incoming hydraulic data has no support in the training set and a new cluster is identified. Error bars color: indicator whether the 90th percentile exceeds a value of 900 MPN/100 mL and risk is predicted. Red vertical dashed lines: single sample threshold for confirmed contamination of 1800MPN / 100 mL. Black vertical dashed line: model decision threshold of 900MPN/100 mL.

quality and the use of the prior predictive distribution for unknown hydrologic situations is novel and innovative as it, to the best of our knowledge, has not been proposed to manage bathing water quality before. We argue that our approach contributes to addressing the key research need identified by WHO (2018) about the minimum number of datapoints and sampling strategies, required for developing predictive models. First, we showed that the poor predictive performance of supervised regression models, especially the zero-model and QRF, persist even for relatively large datasets of $>$ 100 at *River III* and $>$ 400 at *River II*, showing that the question on the minimum amount of datapoints might be too narrowly framed because it lacks aspects on data quality and information content. Second, the two Bayesian approaches can be considered feasible and particularly transparent solutions for ensuring that data quality requirements under various conditions are met. In this context the approach of manual state definition seems especially appealing due to its simplicity regarding both state-definition and updating procedure. Model evaluations like provided in Fig. 4 clearly identify insufficiently populated states and allow for adapting sampling strategies accordingly. In comparison, the use of a non-parametric DPMM introduces an additional level of complexity. While having the advantage of reducing "subjectivity" about how specific hydrological "states" are defined, the DPMM makes the general approach much more difficult to apply.

From a risk management perspective, the provided approaches clearly take a risk averse precautionary approach, by accepting high false positive rates, for the sake of achieving high true positive rates, in the absence of sufficient information. This is achieved by applying a wide precautionary prior for unknown situations which introduces the concept that pollution episodes should be considered a *possible* event unless data indicate otherwise. In the presented cases, we know about the presence of outlets from the combined sewer system in the direct vicinity and upstream of the bathing waters. Therefore, this choice can easily be justified. However, applying the approach to pristine water bodies, without indication of any major pollution source (point or non-point from diffuse sources), might cause unnecessary disqualification of such water bodies und unnecessary efforts to manage them. A potential solution to avoid this, could be to link the application of the presented approach to the information on pollution sources, like the presence of CSO outlets, collected during the elaboration of bathing water profiles which is mandatory for European bathing waters. This would allow such kind of qualitative information to play a more prominent and quantitative role in bathing water quality assessments. Moreover, by linking the choice of such a precautionary approach to the qualitative evidence provided by the bathing water profile, would follow the reasoning of the precautionary principle, which is a cornerstone of European environmental decision making.

Regarding the predictive accuracy, our results confirm the high accuracy of tree-based ensemble methods reported by others (Searcy and Boehm, 2021; Thoe et al., 2014) in the case of balanced datasets. However, the calculated performance indicators show that especially the manual Bayesian approach, often performed similarly well. Our results also confirm the known problem that supervised learning approaches only perform well if the test data resembles the training data (Meyer and Pebesma, 2021). Only the two Bayesian approaches are able to achieve high true positive rates in cases of uninformative training data. Especially the QRF and zero-model perform extremely poorly in these situations. This is noteworthy, since the zero-model closely resembles current way of bathing water classification in Europe, and our results indicate that this approach may lead to unmanaged risks. A real example where four consecutive dry years (2013–2016) led to a classification of "excellent" followed by an extremely rainy year (2017) leading to severe pollution has been reported by Seis et al. (2018), showing the relevance of such situations.

Regarding the assessment of the predictive accuracy is further noteworthy, that there is no relation between the in-sample R^2 and out-of-sample R^2, or the out-of-sample RMSE. This is noteworthy, since it is recommended by WHO (2018) that a minimum explained variance of 50–60 % should be achieved for ensuring the quality of model predictions. While the recommendations do not specify whether these values refer to the in-sample or out-of-sample R^2, we would like to stress that the in-sample R^2, or explained variance, is not a suitable indicator for assessing the predictive performance of a model.

Regarding the broader application, the benefits of using hydrologic predictor variables together with supervised learning approaches for

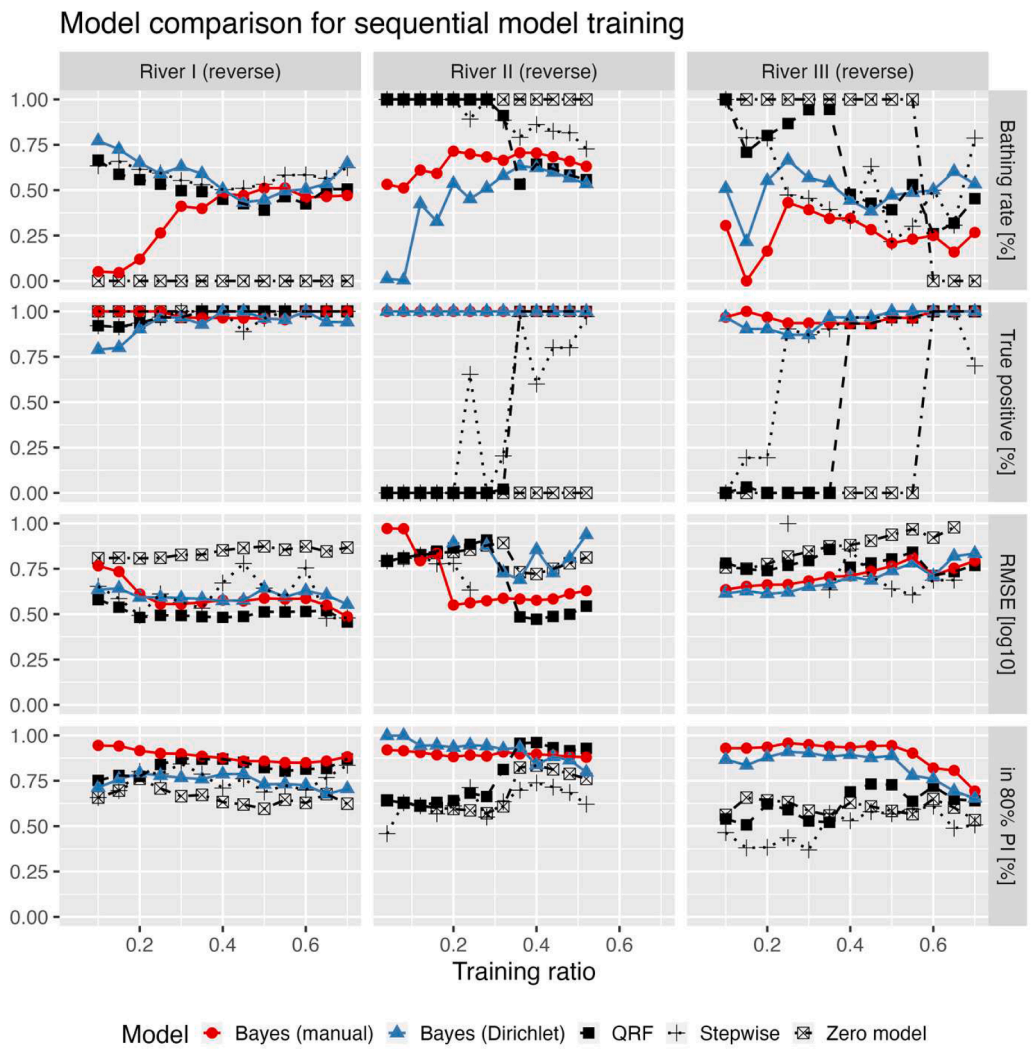


Fig. 6. Model indicators for sequential model training. Bathing rate: ratio of data points in the test set classified as suitable for swimming, in 80 % PI: ratio of test data inside 80 % prediction interval (target: 0.8), RMSE: root mean squared error (log10), values > 1 not shown, True positive: ratio of correctly predicted pollution episodes (*E. coli* < 1800 MPN/100 mL).

predicting FIB concentrations is widely recognized (Francy et al., 2020). However, data scarcity is one of the major bottlenecks for its broader implementation. Therefore, Searcy and Boehm (2021) investigated high resolution sampling as a potential way to cope with this practical problem. Our study addresses the same problem, but instead of collecting many datapoints over a short period, it starts with an empty but generative model which is incrementally updated. Thus, it provides a slower but also less resource intensive approach. Indeed, both approaches could easily complement each other, by using the model to identify unpopulated states/clusters i.e. unknown situations and using techniques for rapid analysis for closing the identified knowledge gaps, thus making the data collection process more efficient.

For now, the benefits of our approach have been shown for river bathing sites affected by short-term pollution, which is the major driver of short-term changes of FIB concentrations (US EPA 2010). For other locations, like bathing waters impacted by more distant, and/or constant pollution sources or marine bathing waters additional predictors, like e.g. tidal changes, global irradiation, or "hour-of-day" might be necessary to account for (Boehm et al., 2002; Wyer et al., 2018), and would need to be further validated.

From a high-level conceptual perspective, it is the ability of including external information quantitatively into a statistical evaluation, its formalized updating procedures and its provision of a generative model based on the prior predictive distribution, which makes Bayesian approaches especially suitable for risk-based environmental decision-support under data-scare situations. In combination with state definitions and cluster assignments, respectively, the two Bayesian approaches not only exploit the information about the correlation of predictor and response variables in the training set, like other supervised learning procedures, but also the information about the similarity of the covariates in the test set relative to the training set, expressed as state/cluster affiliation. Thereby the presented approaches can achieve high true positive rates even under non-informative datasets.

Since in Bayesian inference uncertainty is quantified by means of probability, and probability in turn is regarded as the current *state of knowledge*, also a direct conceptual connection can be drawn to the risk management guidance principle *know your system* which focuses on incremental and continuous improvement procedures. Our assessment approach follows this principle by identifying *known unknowns* indicated by insufficiently populated states and/or new clusters. In such cases precautionary warnings are raises since the *lack of knowledge* is too large to ensure microbial safety.

In summary, our results provide a highly innovative and practically feasible approach to risk-based management of recreational waters and potentially further areas of applied water management, as it supports environmental decision making according to the precautionary principle

by identifying and exploiting information on *known unknowns*.

5. Conclusion

- State-based Bayesian updating may support bathing water quality management in situation of data-scarcity and imbalanced datasets.
- Prior-supported approaches may be a suitable approach to include qualitative catchment information into quantitative statistical analyses and thus can identify high-risk situations, outperforming data-only supervised learning approaches.
- In the absence of informative data, the application of supervised ML models as well as the zero-model may pose significant health risk the bathers.

CRediT authorship contribution statement

Wolfgang Seis: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Marie-Claire Ten Veldhuis:** Writing – review & editing. **Pascale Rouault:** Writing – review & editing. **David Steffelbauer:** Writing – review & editing. **Gertjan Medema:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The data used for this study was collected during the research projects iBATHWATER, and FLUSSHYGIENE, “Hygienically relevant microorganism and pathogens in multifunctional surface water and water cycles: sustainable management of different river types in Germany”. The projects were funded by the European LIFE program and the German Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung, BMBF) under sponsorship numbers LIFE17 ENV/ES/000396, and 02WRS1278A.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.121186.

References

2006/7/EC, 2006. Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC.

Bürkner, P.-C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. J. Stat. Softw. 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Boehm, A.B., Grant, S.B., Kim, J.H., Mowbray, S.L., McGee, C.D., Clark, C.D., Foley, D. M., Wellman, D.E., 2002. Decadal and shorter period variability of surf zone water quality at Huntington Beach, California. Environ. Sci. Technol. 36 (18), 3885–3892.

Breiman, L., 2001. Random Forests. Mach Learn 45 (1), 5–32.

Cyterski, M., Zhang, S., White, E., Molina, M., Wolfe, K., Parmar, R., Zepp, R., 2012. Temporal synchronization analysis for improving regression modeling of fecal indicator bacteria levels. Water Air Soil Pollut. 223 (8), 4841–4851.

Francy, D.S., Brady, A.M.G., Cicale, J.R., Dalby, H.D., Stelzer, E.A., 2020. Nowcasting methods for determining microbiological water quality at recreational beaches and drinking-water source waters. J. Microbiol. Methods 175, 105970.

Francy, D.S., 2009. Use of predictive models and rapid methods to nowcast bacteria levels at coastal beaches. Aquat. Ecosyst. Health Manag. 12 (2), 177–182.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D, 2014. Bayesian Data Analysis, 3rd ed. Chapman and Hall/CRC (Chapman & {Hall/CRC} Texts in Statistical Science).

Kay, D., Wyer, M., Crowther, J., Stapleton, C., Bradford, M., McDonald, A., Greaves, J., Francis, C., Watkins, J., 2005. Predicting faecal indicator fluxes using digital land use data in the UK's sentinel Water Framework Directive catchment: the Ribble study. Water Res. 39 (16), 3967–3981.

Li, Y., Schofield, E., Gönen, M., 2019. A tutorial on Dirichlet process mixture modeling. J. Math. Psychol. 91, 128–144.

Mälzer, H.J., der Beek, T., Müller, S., Gebhardt, J., 2016. Comparison of different model approaches for a hygiene early warning system at the lower Ruhr River, Germany. Int. J. Hyg. Environ. Health 219 (7), 671–680. Part B.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Meyer, H., Pebesma, E., 2021. Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633.

R Development Core Team 2008 A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ross, G.J. and Markwick, D. 2023. Dirichletprocess: build Dirichlet process objects for Bayesian modelling. R-Package.

Searcy, R.T., Boehm, A.B., 2021. A day at the beach: enabling coastal water quality prediction with high-frequency sampling and data-driven models. Environ. Sci. Technol. 55 (3), 1908–1918.

Seis, W., Zamzow, M., Caradot, N., Rouault, P., 2018. On the implementation of reliable early warning systems at European bathing waters using multivariate Bayesian regression modelling. Water Res. 143, 301–312.

StanDevelopmentTeam 2017 Stan Modeling language users guide and reference manual, Version 2.17.0. http://mc-stan.org.

Teh, Y., 2011. Dirichlet Process. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_219.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L., Boehm, A.B., 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. 67, 105–117. Supplement C.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L., Boehm, A.B., 2015. Sunny with a chance of gastroenteritis: predicting swimmer risk at California beaches. Environ. Sci. Technol. 49 (1), 423–431.

WHO 2018. WHO recommendations on scientific, analytical and epidemiological developments relevant to the parameters for bathing water quality in the Bathing Water Directive (2006/7/EC).

Wyer, M.D., Kay, D., Morgan, H., Naylor, S., Clark, S., Watkins, J., Davies, C.M., Francis, C., Osborn, H., Bennett, S., 2018. Within-day variability in microbial concentrations at a UK designated bathing water: implications for regulatory monitoring and the application of predictive modelling based on historical compliance data. Water. Res. X 1, 100006.