On the implementation of reliable early warning systems at European bathing waters using multivariate Bayesian regression modelling

Seis, Wolfgang; Zamzow, Malte; Caradot, Nicolas; Rouault, Pascale

DOI
10.1016/j.watres.2018.06.057

Publication date
2018

Document Version
Final published version

Published in
Water Research

Water Research 143 (2018) 301–312

Contents lists available at ScienceDirect

Water Research

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

On the implementation of reliable early warning systems at European bathing waters using multivariate Bayesian regression modelling

Wolfgang Seis [a, b, *], Malte Zamzow [a], Nicolas Caradot [a], Pascale Rouault [a]

[a] Kompetenzzentrum Wasser Berlin gGmbH, Cicerostraße 24, 10709 Berlin, Germany
[b] Delft University of Technology, The Netherlands

ARTICLE INFO

Article history:
Received 14 February 2018
Received in revised form
13 June 2018
Accepted 24 June 2018
Available online 26 June 2018

Keywords:
Early waring
Bathing waters
Bathing water directive
Bayesian regression modelling

ABSTRACT

For ensuring microbial safety, the current European bathing water directive (BWD) (76/160/EEC 2006) demands the implementation of reliable early warning systems for bathing waters, which are known to be subject to short-term pollution. However, the BWD does not provide clearly defined threshold levels above which an early warning system should start warning or informing the population. Statistical regression modelling is a commonly used method for predicting concentrations of fecal indicator bacteria. The present study proposes a methodology for implementing early warning systems based on multivariate regression modelling, which takes into account the probabilistic character of European bathing water legislation for both alert levels and model validation criteria. Our study derives the methodology, demonstrates its implementation based on information and data collected at a river bathing site in Berlin, Germany, and evaluates health impacts as well as methodological aspects in comparison to the current way of long-term classification as outlined in the BWD.

1. Introduction

Ensuring microbial safety is one of the key objectives of bathing water management. The fecal indicator bacteria (FIB) *Escherichia coli* (*E. coli*) and intestinal enterococci are the most frequently used indicator organisms for the assessment of microbial safety at recreational waters. The European Bathing Water Directive (BWD) (76/160/EEC, 2006) uses both indicators for quality assessment of marine and inland waters. A major challenge regarding bathing water management is that concentrations of FIB may show spatial and temporal variability triggered by different causes and occurring on different temporal (e.g. seasonal, monthly, diurnal, hourly), and spatial scales (e.g. along-shore, longitudinal, depth) (US-EPA, 2010). With respect to temporal variability, the occurrence of some phenomena may follow regular and predictable patterns, like tidal changes at coastal beaches or within day variability caused by UV irradiation (Boehm et al., 2002). In contrast, other reasons, like event-scale variation due to heavy rainfall may show more stochastic patterns (Traister and Anisfeld, 2006). The latter may lead to discharges from multiple sources of urban drainage systems like combined sewer overflows (CSO) and stormwater discharges, which both may contain high amounts of FIB. Consequently, event-scale variability leads to the highest variations with respect to short-term temporal changes of FIB concentrations in surface waters (US-EPA, 2010). For this reason, the European BWD demands to elaborate so-called bathing water profiles for all bathing waters in order to identify the potential sources of contamination and to assess, whether the bathing water is subject to short-term pollution. If the assessment reveals that a bathing water is expected to be subject to short-term pollution, the current BWD (76/160/EEC, 2006, Article 12(c)) explicitly demands the implementation of early warning systems in order to prevent bathers from being exposed to contaminated water. However, the BWD neither provides guidance on how to implement early warning systems nor does it provide single sample, indicator based threshold levels above which an early warning system should start warning the population. Bathing water quality is assessed only in the long term by estimating parametric 90th and 95th percentiles based on the surveillance data of the previous four years (cf. section 2.1.1). The

Abbreviations: WWTP, wastewater treatment plant; LOO-IC, Approximate leave-one-out information criterion; PPD, Posterior predictive distribution; Q, Flow in [m^3/s]; MLE, Maximum likelihood estimate; MPN, most probable number; FIB, Fecal indicator bacteria; P, Precipitation in [mm/d]; BWD, Bathing water directive; EU, European Union; PI / CI, prediction interval/credible interval; CSO, combined sewer overflow.

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

https://doi.org/10.1016/j.watres.2018.06.057
0043-1354/

lack of specified thresholds makes it difficult for the responsible authorities to justify and defend short-term decisions about closures of or warnings on bathing sites.

An additional challenge regarding the implementation of early warning systems for managing event-scale variability, or short-term pollution, is the quality of available surveillance data. Bathing water surveillance in Europe is only based on periodic (at least monthly) grab samples. Accordingly, event-scale variability is detected only by chance as pollution events may occur between sampling intervals (Kay et al., 2005). Even if grab sampling takes place same day a pollution event occurs the sampling time and the occurrence of the pollution event do not necessarily coincide. In such cases, the grab sample would indicate low FIB concentrations, which would be correct for that specific location and time. If, however, it is intended to manage the bathing site on a daily basis, as e.g. outlined in (Stidson et al., 2012), the measurement will not contain the correct information regarding the management objective. Finally, even if a grab sample indicates short-term pollution no information is given about the duration of the contamination.

Due to the lack of a single sample threshold and the limited information given by periodic/single grab samples, the prediction of bathing water quality as well as its validation remain a challenge. Studies which used statistical modelling, regardless of using classification or regression approaches, often validated the *accuracy* of the made predictions by using a contingency matrix approach (Brooks et al., 2013; Heberger et al., 2008; Herrig et al., 2015; Mälzer et al., 2016; Motamarri and Boccelli, 2012; Stidson et al., 2012; Brady, 2007). This approach determines specificity and sensitivity by the rate of *false positives* (FP), *false negatives* (FN), *true positives* (TP) and *true negatives* (TN). If, however, the validation data cannot be trusted to contain the correct information regarding the management objective (e.g. daily bathing water quality), a high false positive rate may be either the result of an overly conservative model or simply the result of random sampling error.

Against this background the objectives of the present study are:

a.) To derive a decision criterion for early warning as well as to develop model validation criteria in view of the probabilistic character of European bathing water legislation and uncertain data
b.) To develop statistical models, which allow for day-to-day bathing water management based on official surveillance data fulfilling the derived criteria
c.) To use event-based sampling to check whether models are able to predict the duration of contamination in time scales relevant for daily bathing water management
d.) To compare the proposed management approach to the current way of long term classification, both in terms of microbial safety and methodological aspects.

2. Material and methods

In order to be accepted by European authorities and beach managers, alert levels for early warning should be related as closely as possible to the thresholds outlined in the European BWD. Therefore, we first elaborate a general methodological approach for using the current numerical standards of the BWD as alert levels for early warning systems as well as for validating the quality of model predictions. In a second step, we apply the derived methodology at a river bathing site, known to be recurrently affected by short term pollution. Finally, we compare the suggested approach to the current way of long-term classification in terms of microbial safety.

Since *E. coli* dominates bathing water classification in Berlin, meaning that elevated concentrations of intestinal enterococci without simultaneous increases of *E. coli* are rare, the study focuses on the prediction of *E. coli* concentrations. However, the suggested method can be readily applied to other FIB.

2.1. Methodological approach for deriving alert levels and model validation criteria

2.1.1. Current approach of long-term classification as set in the EU BWD

The European BWD defines numerical standards only for long-term bathing water classification. To this end, a probabilistic approach based on parametric 90th and 95th percentiles is applied. The approach assumes that the measured data are $\log_{10}$-normally distributed and derives these percentiles by a) $\log_{10}$-transformation of the measured data, b) calculation of arithmetic mean and standard deviation c) calculation of the parametric 90th and 95th percentiles by:

$$95\text{th percentile} = \mu_s + 1.65^{*}\sigma_s \quad (1)$$

$$90\text{th percentile} = \mu_s + 1.282^{*}\sigma_s \quad (2)$$

and, d) back-transformation from the $\log_{10}$ to the linear scale. μ_s and σ_s stand for the sample mean and the sample standard deviation of a normal distribution N (μ_s, σ_s^2). Bathing waters are classified *once a year* in view of the upcoming bathing season. For this purpose, the surveillance data of the four previous bathing seasons are used with a minimum of 16 samples in total. Table 1 summarizes the different quality classes ranging from “excellent” to “poor”.

2.1.2. Using BWD quality standards as alert levels for early warning

In order to use the numerical quality standards of the BWD for early warning systems, we need to a) define appropriate alert levels and b) increase the timely resolution from *once per year* to time scales relevant for bathing water management.

2.1.2.1. Defining appropriate alert levels. Regarding appropriate alert levels, we are especially interested in the quality class “poor”, since it is the quality we want the population to be warned of. According to the BWD, bathing water quality is classified as “poor” when the point estimate of the 90th percentile based on a minimum of 16 samples collected over four years and assuming a lognormal statistical model indicates that the probability of measuring values larger than 900 MPN/100 mL exceeds 10%. Consequently, if we want to use these standards for early warning we have to predict a probability based on a lognormal model. A way of estimating the “probability of exceeding”, which accounts for the lognormality condition, is to fit a linear regression model on the $\log_{10}$-transformed data and to use the predicted mean and residual standard deviation to construct a lognormal probability density function (PDF). In the USA, Heberger et al., 2008 used such an approach to first calibrate the “probability of exceeding” of the single-sample swimming standard of Massachusettes of 61 enterococci cfu/100 mL. The calibrated probability was used

Table 1
Bathing water quality requirements for inland waters according to (76/2006/EC).

Indicator	Excellent	Good	Sufficient	Poor
E. coli [MPN/100 mL]	<500*	<1000*	<900**	>900**
Intestinal enterococci [MPN/100 mL]	<200*	<400*	<330**	>330**

* Based on the 95th percentile, **based on the 90th percentile.

subsequently as a decision criterion above which the model would indicate impaired water quality.

In a European setting, we argue that bathing water quality standards are already defined as "probabilities of exceeding", so there is actually no need to calibrate these values as they are already given by the BWD, namely 10% for sufficient/poor quality and 5% for good/excellent quality. Thus, we propose to estimate the "probability of exceeding" based on statistical regression modelling on the $\log_{10}$-transformed FIB data and use these estimates as decision criteria against the percentile thresholds as outlined in the BWD. In the present study, we applied a Bayesian approach for regression modelling. Therein, the constructed PDF is referred to as the posterior predictive distribution (PPD) and the 95% prediction interval as the 95% credible interval (CI) of the PPD. In general, the equivalent prediction interval could be computed using a frequentist approach.

2.1.2.2. Increasing the timely resolution. In Scotland, information systems are in place, which inform the public about potentially impaired water quality on a daily basis (Stidson et al., 2012). Following this approach, we use correlations between FIB and readily available data (cf. section 2.2.4), to update our estimates of the probability of exceeding on a daily basis.

2.1.3. Validating predicted probabilities by percentage coverage

A challenge regarding the prediction of probabilities is that they cannot be validated by comparing predicted classes to single measurements using e.g. a contingency matrix approach. As an illustration, if a regression model predicts a probability of exceedance of 20% the classification would be "poor". However, if the prediction were correct, we would expect 80% of the validation data to fall below the threshold and 20% to fall above it. Thus, a single measurement gives no information about the correctness of the prediction. Moreover, official surveillance data deliver only one data point per day, so we cannot validate the true distribution for each day using only daily grab samples. For this reason, we focus on evaluating overall model consistency by looking at all validation data points at once. If the model is able to capture reality, from a probabilistic perspective, 95% of the validation data should fall within the 95% credible intervals (CI) of the PPD, 95% should fall below the 95th percentiles and 90% should fall below the 90th percentiles. So, e.g. from one hundred validation data points collected on one hundred days, the constructed one hundred 95% CIs of the PPDs should cover 95%.

As these criteria are in turn subject to sample size-based uncertainty, we define a one-sided cut-off criterion based on a beta distribution to exclude all models for which the probability that the true coverage rate is 95% and 90%, respectively, is less than 5%. A Beta (1, 1) is used as a uniform prior distribution. We underline that fulfilling these criteria is no proof that the PPD predicted for each day is correct, it is rather an additional minimum requirement, which checks for overall model consistency with the test/validation data.

2.1.4. Differences between suggested regression-based approach and the approach for long-term classification

Although both the current way of long-term classification and the suggested regression-based approach estimate the probability of exceeding based on a lognormal model there are some conceptual differences. First, the current way of calculating upper percentiles according to the BWD does not account for parameter uncertainty regarding μ and σ^2. From a human health perspective, we prefer including parameter uncertainty, which leads to a more conservative estimate of the upper percentiles. Secondly, current long-term bathing water classification is based on the data from only four years. In order to increase the chances that the training data include data influenced by short-term pollution as well as to reduce parameter uncertainty we include all available surveillance data. Thirdly, according to the classification methods as outlined in the BWD all variance in the empirical data is described by the variance parameter σ^2_s of the lognormal PDF. In regression, the variance in the training data is partly explained by the linear model, expressed as R^2. Thus, the residual standard deviation decreases with an increase in R^2. This relation will become important for model selection and validation (cf. section 3.3).

2.2. Application of the described methodology to a river-bathing site in Berlin, Germany

2.2.1. Bathing site

The bathing site "Kleine Badewiese" is situated on the River Havel, downstream of the city center of Berlin, Germany. The bathing site is an official European bathing water and is regularly monitored for FIB during the bathing season (May 15 to September 15). In Berlin, the usual sampling interval is 14 days. From 2012 to 2015, bathing water quality was classified as "poor", while in 2016 and 2017 it was classified as "good" and "excellent", respectively. Due to water quality problems from 2012 to 2015, the responsible health authorities increased sampling frequencies to weekly samplings during 2016 and 2017.

2.2.2. Catchment description

The sewer system in Berlin consists of a separate sewer system in the outer parts of the city and a combined system in the city center (Fig. 1). There are approximately 170 combined sewer overflow (CSO) outlets in Berlin. Heavy rainfall can therefore lead to discharges from CSOs and CSO-impacted river water, which eventually reaches the bathing location. In Berlin, flow velocities are usually very low. Depending on the flow (Q) in the river, travel times from the city center to the bathing site may range between 32 h and 14 days (Schumacher and Storz, 2016).

In addition to the CSO outlets, the largest of Berlin's six wastewater treatment plants (WWTP) is located upstream of the bathing location. In dry weather, the WWTP treats approximately 250000 m^3/d. In order to protect bathing water quality during dry periods the majority of the secondary effluent of the WWTP is not discharged into the rivers Spree and Havel but is pumped south and discharged into a bypassing canal. The proportion of secondary effluent, which is not pumped south, is treated by an additional UV disinfection. Therefore, only disinfected wastewater is discharged into the adjacent river in dry weather. During rainy weather, however, water volumes are too high to be treated completely by the disinfection unit or to be pumped southwards. Consequently, non-disinfected secondary effluent is discharged into the river with a maximum flow of approximately 3 m^3/s. Water impacted by these discharges will eventually reach the bathing site.

2.2.3. Data availability of FIB and event-based monitoring

Official surveillance data, based on periodic grab samples from 2010 to 2017 ($N = 114$) were provided by the Berlin health authorities (LAGeSo). Grab samples are measured using *most probable number* methods (ISO 9308-3) with a 2-fold dilution and a lower detection limit (LOD) of 15 MPN/100 mL. Since biweekly grab samples only detect short-term pollution by chance and do not give any information about the duration of a contamination, refrigerated automated samplers (HYDREKA, Sigma AS 950, see supplementary material) were installed for event-based sampling during the bathing seasons 2016 and 2017. The main purpose of event-based sampling was to identify periods of major fecal pollution from CSO and WWTP discharges relevant for daily bathing water

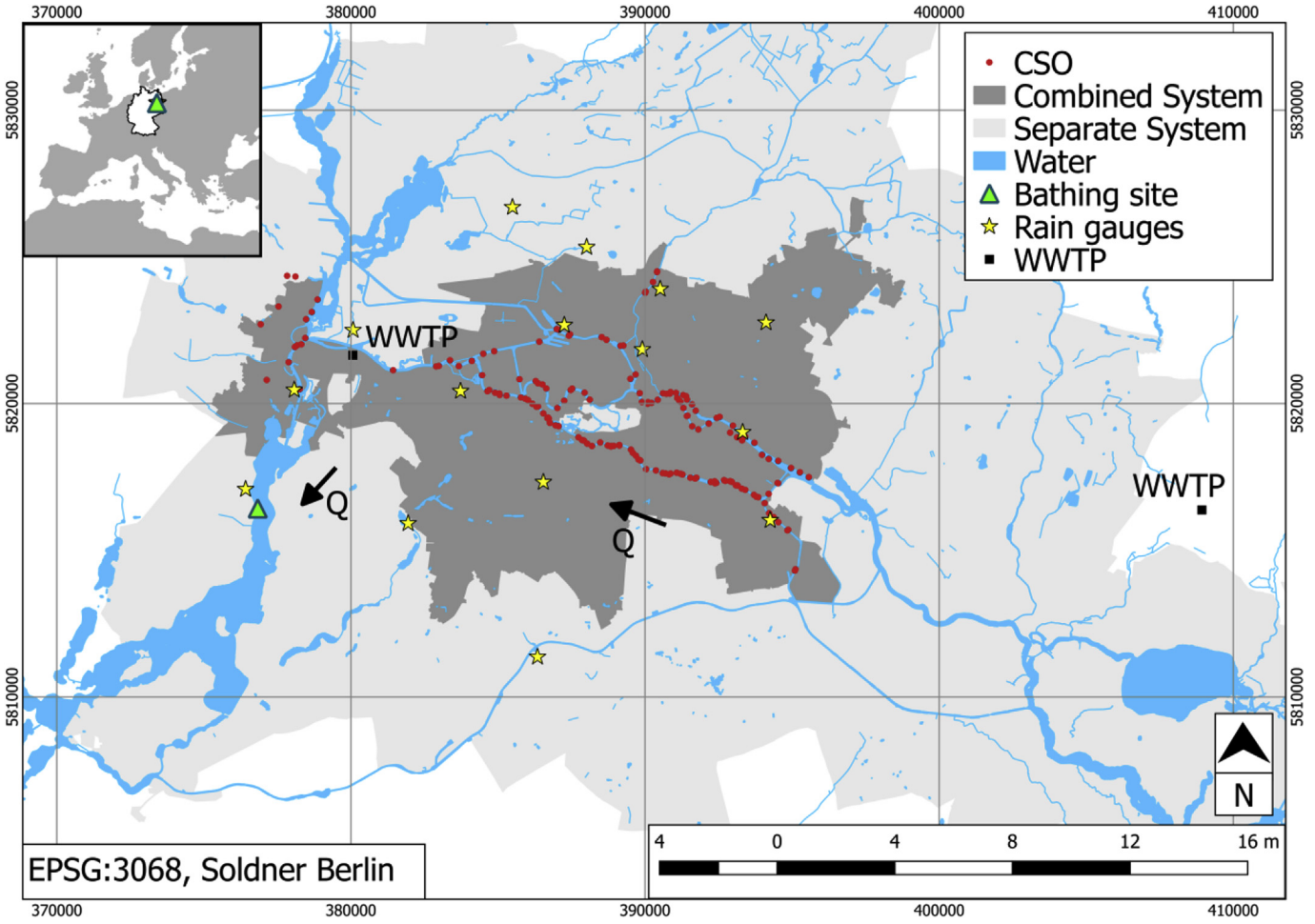

Fig. 1. Overview of study area: WWTP: wastewater treatment plant, Q + arrows indicate flow direction.

management. Samples were taken during the five days following rain events. Since impacts from urban wastewater were expected, samples were diluted 4-fold in order to ensure that the upper detection limit would suffice. Samplers contained 24 1L PE bottles (HACH PS241000). Each bottle was filled within 1 h taking 200 mL every 12 min. Samplers were cooled at 4 °C. Samples were taken about 5 m from the riverbank at a depth of approximately 1.2 m. Prior to the actual sampling tubes were rinsed three times to avoid carry over. For the transport to the laboratory, bottles were stored in cooling boxes with cold packs. Samples were unified to 12 h composite samples. The sampling bottles were washed in a laboratory washing machine at 55 °C with de-ionized water for 60 min and dried in a drying cabinet at 50 °C. The latter was considered acceptable against the monitoring objective of detecting periods of high fecal pollution from municipal wastewater. Due to logistic reasons (laboratory working hours) samples had to be collected around noon. Thus, given that the initiating rain event occurred in the afternoon the first 24 h sampling period was stopped at noon of the following day. This way the first sampling period was shortened. Starting times of the individual composite samples are presented in the supplementary material. Samples were analyzed for FIB, the afternoon samples were collected using MPN methods (ISO 9308-3). Therefore, the first aliquots of the first 12 h composite sample of the 24 h sampling period may have exceeded the recommended time-period for analysis of 24 h by up to 4 h and might be under-represented. Since the intention was to sample different rain weather conditions no specific threshold level for the minimum precipitation was defined. Due to the low flow velocities in Berlin and the short driving distances to the bathing site, automated samplers could be started manually.

By implementing event-based sampling as an additional sampling protocol, two distinct data sets are available. The first one consists of official surveillance data based on grab samples from 2010 to 2017 (N = 114) with a lower detection limit (LOD) of 15 MPN/100 mL (ISO 9308-3). The second one consists of 12 h composite samples collected during 2016–2017 with a LOD of 40 MPN/100 mL (N = 118). As the data sets are distinct regarding the information they contain, the two types of data were used differently for model checking and validation (cf. section 2.2.9.1 and section 2.2.9.2).

2.2.4. *Variable selection for regression modelling*

For the implementation of an early warning system for event-scale variability due to heavy rainfall, explanatory variables for regression modelling were selected considering the following criteria: a) explanatory variables should be measured in high timely resolution b) data should be readily available to minimize efforts for implementation, and c) there should be a plausible explanation for the effect of the explanatory variable on fecal contamination at the bathing site. Following these criteria, precipitation (P), river flow (Q) and the volume of the non-disinfected discharge of the WWTP (WWTP) were selected as key explanatory variables. The Berlin Water Utilities (BWB) (WWTP, P) and the Berlin Senate Department for the Environment, Transport and Climate Protection (SenUVK)

(Q) provided the necessary data. A more detailed justification for choosing these variables is provided in the supplementary material.

2.2.5. Variable construction and transformations

Due to changing travel-times of CSO discharges to the bathing site water quality at the bathing site might be impaired concurrently by both rain events that occurred just recently and rain events that happened several days ago. The arrival time of the contamination depends on the river flow Q. In order to account for these kinds of variations, different explanatory variables were constructed. From WWTP discharges as well as from flow data Q, daily sums (WWTP) and averages (Q) were calculated up to five days prior to sampling for each individual day (Q_1, Q_2, … Q_5 and $WWTP_1$, $WWTP_2$ … $WWTP_5$). Moreover, variables were constructed which summed/averaged over multiple days prior to sampling, e.g. $Q_{1\text{-}3}$, $WWTP_{1\text{-}3}$ represent the average/sum over three days prior to sampling. A similar approach has already been successfully applied by Cyterski et al. (2012) and Herrig et al. (2015). Rainfall variables were created analogously, with two exceptions. First, spatial averages over all 15 rain gauges were calculated before averaging over time. Secondly, the created rainfall variables were log-transformed. A value of 1 was added before log-transformation. The rationale for log-transformation is that while discharges of CSO and stormwater may increase FIB concentrations by orders of magnitude in the first instance, a further increase in rainfall and consequently discharge volume will not increase the concentration further linearly on a $\log_{10}$ scale. By log-transformation, the effect of higher rainfall levels is weakened. The sampling day was not included in the averaging, since in the case of historical data precipitation might have started after sampling creating artefacts of wet weather conditions, when the sample might actually have been taken during dry weather conditions. Due to the lognormality assumption given by the BWD (cf. section 2.1.1), *E. coli* data were $\log_{10}$-transformed. Values at the lower and upper detection limit were kept at these values.

2.2.6. Model formulations

The constructed explanatory variables were used to construct nine different models. Table 2 gives an overview of the constructed models and explanatory variables using the notation outlined in section 2.2.5. Models 1–4 focus on different times scales prior to sampling regarding WWTP and P. Models 5–7 include the WWTP with two individual days prior to sampling and differ regarding the rain variables. Model 8 uses only Q and rainfall data P. Model 9 was created by stepwise forward selection, allowing for pairwise interactions of explanatory variables and limiting the maximum number of explanatory variables to five.

Interactions: Due to the low flow conditions and long travel times many discharges located upstream of the bathing site are only relevant if flow conditions are high enough to allow the fecal contamination to reach the bathing site within a certain time. This dependency is considered by including interaction effects between P and Q as well as between WWTP and Q. In statistical modelling interaction effects are included as the product of two or more predictors (see Table 2). Thereby, the effect of Q depends on P and vice versa.

2.2.7. Model fitting

All models were fitted using Hamiltonian Monte Carlo (HMC) for Markov Chain Monte Carlo (MCMC) using the programming languages R (R Development Core Team, 2008) and Stan (StanDevelopmentTeam, 2017b) accounting for full parameter uncertainty. In R, the package for applied Bayesian regression modelling *rstanarm* and the package function *stan_lm()* (StanDevelopmentTeam, 2017a) was used. Priors were based on the R^2 statistic ($R^2 = 0.8$).

2.2.8. Data separation for model fitting and validation

For model fitting and determination of regression parameters θ the historical data (grab samples) from regular bathing water surveillance from 2010 to 2015 ($N = 74$) were used. For model validation both the collected event-based samples ($N = 118$) and the surveillance data of the regular surveillance monitoring of the years 2016 and 2017 ($N = 40$) were used. (cf. section 2.2.9.1 and section 2.2.9.2). The two validation years were very distinct. The year 2016 was dryer than average with 89% of the annual average precipitation and low flow conditions in June, July and August. In contrast, 2017 was one of the rainiest years ever recorded, with precipitation of more than 200% of the annual average for the same months, including a 120-year-rain-event. Therefore, models could be validated against a broad range of different conditions.

2.2.9. Model checking and comparison

Normality and homoscedasticity of residuals were tested using the *Shapiro-Wilk Test* and the *Breusch Pagan Test*, respectively. R^2 was used to analyze how much of the variance in the training data is explained by each statistical model. However, since the proposed decision criteria for early warning are based on the upper percentiles of the PPD, we are less interested in R^2 but rather in whether the model is able to capture the variance in the test data. Thus, the criteria outlined in section 2.1.3 were applied to each bathing season as well as to all grab samples.

2.2.9.1. Percentage coverage. For the application of the percentage coverage criterion, only the official surveillance data (grab samples) were used to ensure, that the data for fitting and validating the

Table 2
Overview of model equations used for statistical modelling. *E. coli*: *E. coli* concentration [MPN/100 mL], Q: river flow [m^3/s], P: rain log [mm/d], WWTP: discharge of WWTP [1000m^3/d], ε: Error term.

Model	Equation
Model1	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_2 + \beta_P{*}\log P_2 + \beta_{WWTP}{*}WWTP_2 + \beta_{Q,P}{*}Q_2{*}\log P_2 + \beta_{Q,WWTP}{*}Q_2{*}WWTP_2 + \varepsilon$
Model2	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_P{*}\log P_{1\text{-}3} + \beta_{WWTP}{*}WWTP_{1\text{-}3} + \beta_{Q,P}{*}Q_{1\text{-}2}{*}\log P_{1\text{-}3} + \beta_{Q,WWTP}{*}Q_{1\text{-}2}{*}WWTP_{1\text{-}3} + \varepsilon$
Model3	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_P{*}\log P_{1\text{-}4} + \beta_{WWTP}{*}WWTP_{1\text{-}4} + \beta_{Q,P}{*}Q_{1\text{-}2}{*}\log P_{1\text{-}4} + \beta_{Q,WWTP}{*}Q_{1\text{-}2}{*}WWTP_{1\text{-}4} + \varepsilon$
Model4	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_P{*}\log P_{1\text{-}5} + \beta_{WWTP}{*}WWTP_{1\text{-}5} + \beta_{Q,P}{*}Q_{1\text{-}2}{*}\log P_{1\text{-}5} + \beta_{Q,WWTP}{*}Q_{1\text{-}2}{*}WWTP_{1\text{-}5} + \varepsilon$
Model5	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_P{*}\log P_{1\text{-}2} + \beta_{WWTP1}{**}WWTP_1 + \beta_{WWTP2}{*}WWTP_2 + \beta_{Q,WWTP1}{*}Q_{1\text{-}2}{*}WWTP_1 + \beta_{Q,WWTP2}{*}Q_{1\text{-}2}{*}WWTP_2 + \beta_{Q,P}{*}Q_{1\text{-}2}\log P_{1\text{-}2} + \varepsilon$
Model6	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_P{*}\log P_{1\text{-}2} + \beta_{WWTP1}{*}WWTP_1 + \beta_{WWTP2}{*}WWTP_2 + \beta_{Q,WWTP1}{*}Q_{1\text{-}2}{*}WWTP_1 + \beta_{Q,WWTP2}{*}Q_{1\text{-}2}{*}WWTP_2 + \beta_{Q,P}{*}Q_{3\text{-}5}\log P_{3\text{-}5} + \varepsilon$
Model7	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_{WWTP1}{*}WWTP_1 + \beta_{WWTP2}{*}WWTP_2 + \beta_{Q,WWTP1}{*}Q_{1\text{-}2}WWTP_1 + \beta_{Q,WWTP2}{*}Q_{1\text{-}2}WWTP_2 + \beta_{Q,P}{*}Q_{1\text{-}5}\log P_{1\text{-}5} + \varepsilon$
Model8	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_{1\text{-}2} + \beta_P{*}\log P_{1\text{-}2} + \beta_P\ \log P_{3\text{-}5} + \beta_{Q,P}{*}Q_{1\text{-}5}\log P_{3\text{-}5} + \varepsilon$
Model9 (forward selection)	$\log_{10}E.\ coli \sim \alpha + \beta_Q{*}Q_2 + \beta_P{*}\log P_{1\text{-}3} + \beta_{WWTP}{*}WWTP'_{1\text{-}2} + \beta_{Q,WWTP}{*}Q_2{*}WWTP_{1\text{-}2} + \beta_{P,WWTP}{*}\log P_{1\text{-}3}{*}WWTP_{1\text{-}2} + \varepsilon$

model have the same quality. For the individual years of 2016 and 2017, as well as for both years combined, the minimum number of data points required to fall within the test intervals for validation of the 95% credible interval and the 95th percentile, are 19/21, 17/19 and 36/40 (90%). For validation of the 90th percentile, the criteria are 17/21, 16/19 and 34/40 (86%). Validation data are considered inside the prediction interval if the best estimate of the MPN analysis falls within the predicted interval.

2.2.9.2. Graphical model checking. The concentrations measured within the collected composite samples represent the arithmetic mean (AM) of the FIB concentrations during the sampling period. Model predictions are based on the $\log_{10}$-transformed data and consequently predict the geometric mean (GM) of FIB concentrations. In comparison to the GM, the AM is easier dragged to higher concentrations in case a pollution event occurs during the sampling period. Therefore, the AM can be considered a more conservative estimate of the average concentration during the sampling period than the GM (Reicherts and Emerson, 2010). For the assessment of model quality, we used the composite samples for graphical model checking. Since composite samples give information about the duration of contamination at a 12 h resolution, we used these samples to check whether the models are able to cover that 12 h trend (see supplementary material).

2.2.9.3. Use of information criteria. The approximate leave-one-out-cross validation information criterion (LOO-IC) was used as an additional indication for the predictive performance of each model (Vehtari et al., 2017). All calculations were conducted using the *"loo" –package* in combination with the function *loo()* (Vehtari et al., 2016). The LOO-IC was calculated for both the training set as well as after refitting the models with all grab samples. Lower LOO-IC value indicate better predictive performance.

2.3. Comparison to long-term classification with regard to health protection

In order to analyze the health implications of the suggested management approach, it was applied retrospectively to the years 2016 and 2017. For each day from May 15 to September 15 the 90th and 95th percentiles of the PPD were simulated using the *posterior_predict()* function (StanDevelopmentTeam, 2017a) and compared to the percentile thresholds as outlined in the BWD (cf. Table 1). Thereby, bathing water quality was classified for each day using the hydraulic information from the days before. The results were compared to official bathing water classifications complemented with official warnings communicated by the local health authority. The latter are communicated via press release and are mainly experience-based *ad-hoc* decisions.

3. Results

3.1. Checking for lognormality and basic models assumptions

The individual test results for normality and homoscedasticity for each model are given in the supplementary material. Two (models 6 and 7) out of nine models failed the normality test. The residuals showed no heteroscedasticity. A comparison of the normality assumption before and after regression modelling is shown in Fig. 2. Results show that while both raw data and lognormally transformed data are not normally distributed, regression modelling now leads to normally distributed residuals. The latter justifies the use of a lognormal parametric approach for estimating the 90th and 95th percentiles (Fig. 2).

3.2. Checking for percentage coverage

The results of model checking against the percentage coverage criteria are shown in Fig. 3. Fig. 4 shows an illustrative example of the best performing model (model 3). Models, which violated the normality assumption, are not shown anymore. From the remaining seven models, only three (model 3, model 4 and model 8) passed the applied criteria. The 95th percentile test interval indicates to be the strictest criterion among the applied test intervals. Fig. 5 shows the relation between the variance-explained (R^2) and the different percentage coverage criteria for both validation years. The model covering the highest proportion of validation data (model 3) has the second lowest R^2 whereas the model with the highest R^2 (model 5) value performs the worst. This behavior shows that optimizing models only in terms of R^2 may lead to overfitting and thereby yields to poor predictive accuracy. Applying the percentage coverage criterion instead reduces overfitting by ensuring

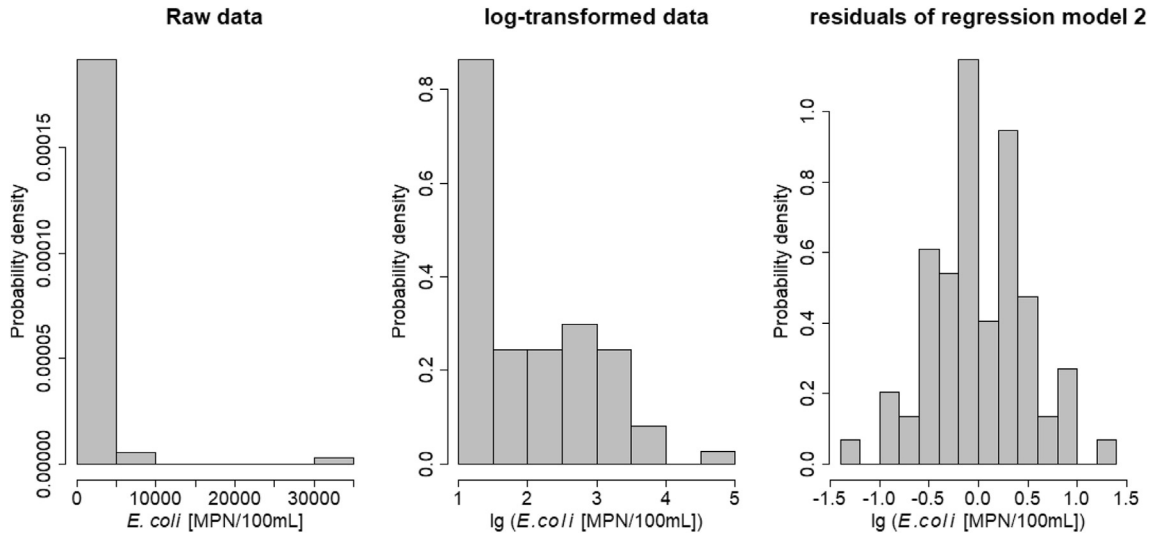

Fig. 2. Comparison of the lognormality assumption before and after regression modelling. From left to right: a) raw FIB training data, b) log-transformed training data and c) residuals of fitted regression model.

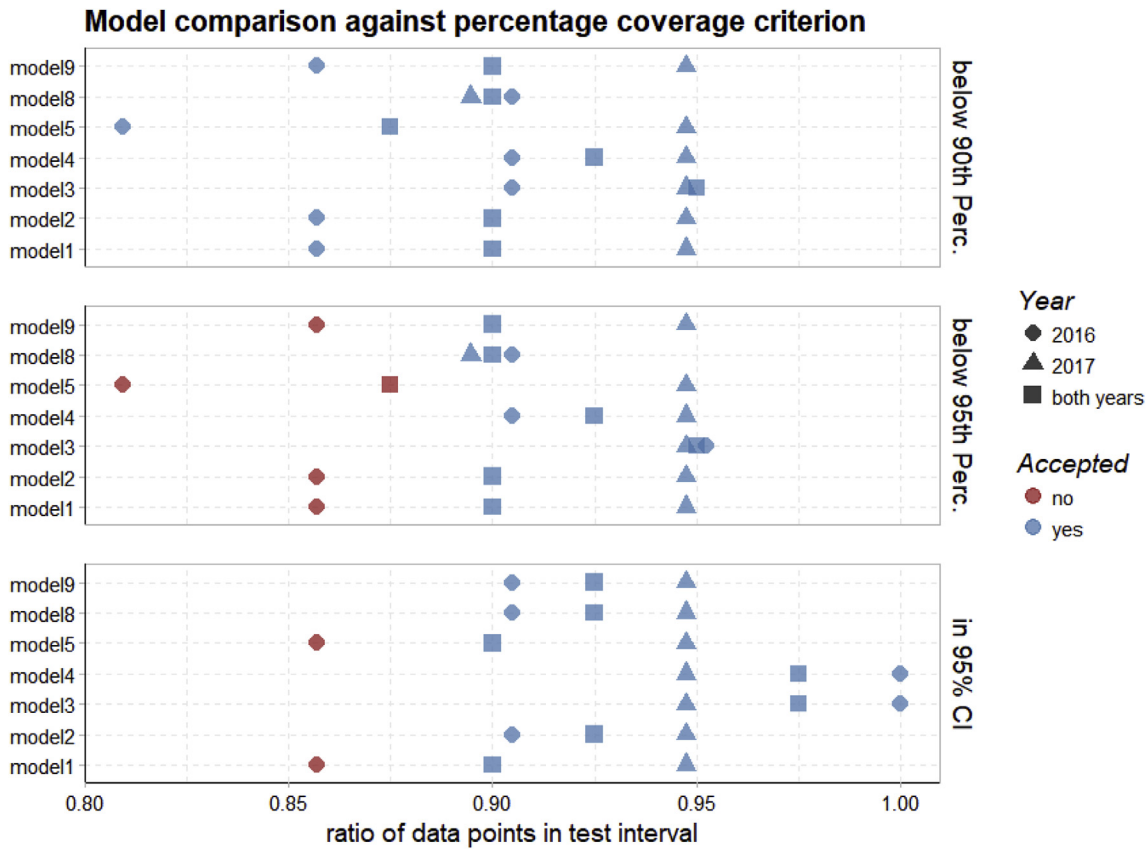

Fig. 3. Model comparison against percentage coverage criteria. Perc. = percentile, CI = credible interval of posterior predictive distribution.

that the model's residual standard deviation remains sufficiently high to reflect the variation in the test data. Fig. 4 also shows that the standard deviation of the predicted distributions has to be very high (SD = 0.6) to cover this variation. The PPD covers approximately 2–2.5 orders of magnitude.

3.3. Graphical model checking and information criteria

Times series plots of the seven models which passed the normality test are given in the supplementary material. In order to visualize the temporal trends, the confidence intervals of the geometric mean are connected by linear interpolation. The graphical comparison to the 12 h composite samples indicates that the three models, which passed the percentage coverage criterion, are able to predict periods of severe fecal pollution. However, models 3 and 4 cover both composite samples and grab samples better than model 8. Due to the lower LOO-IC value (supplementary material) of model 3 in comparison to model 4, model 3 is preferred.

3.4. Comparison to long-term classification with regard to health protection

Fig. 6 shows the comparison between long-term classification based on data from the previous four years (official classification) and daily classifications based on model 3. The latter is based on the information and evidence contained in the training data from 2010 to 2015. Fig. 7 shows the related time series plots, monitoring data (grab and composite samples), and periods of "*ad-hoc*" precautionary warnings. Due to the low concentrations measured between 2013 and 2016, in 2017 bathing water quality was classified as "excellent" by long-term classification. However, due to the rainy weather in 2017, the results from composite sampling after heavy rainfall reveal severe fecal contamination with measured *E. coli* concentrations of up to 10^5 MPN/100 mL. It is worth to mention that even weekly grab samples did not capture this contamination. The results prove that long-term classification only based on the surveillance data from the previous four years and not considering any other environmental conditions may lead to substantial misjudgments of the actual bathing water quality and may pose a hazard to human health. In contrast, by updating percentile estimates on a daily basis, bathing water quality classifications would alternately indicate periods of better and "poor" quality. Predictions of "poor" comply with elevated concentrations measured in composite samples, which represent 12 h averages. The model generally shows high consistency regarding the results from both sample types. However, at three occasions (marked with white arrows in Fig. 7) the model would have predicted "sufficient" water quality or better, while the results from composite samples indicate concentrations above the percentile threshold of 900 MPN/100 mL. At two of the three occasions (1 and 3), heavy rainfall (90 mm, 12.8 mm) on the sampling day itself could be identified as the cause for these high values. However, including rainfall on the sampling day as an additional explanatory variable in the regression model did not improve predictions for these days. Thus, the information was missing in the training data. With regard to observation 2 only minor rainfall occurred on the sampling day (2 mm). However, the approach presented demonstrates that health protection is improved substantially in comparison to the current way of "*ad-hoc* warning" and especially in comparison to long-term classification.

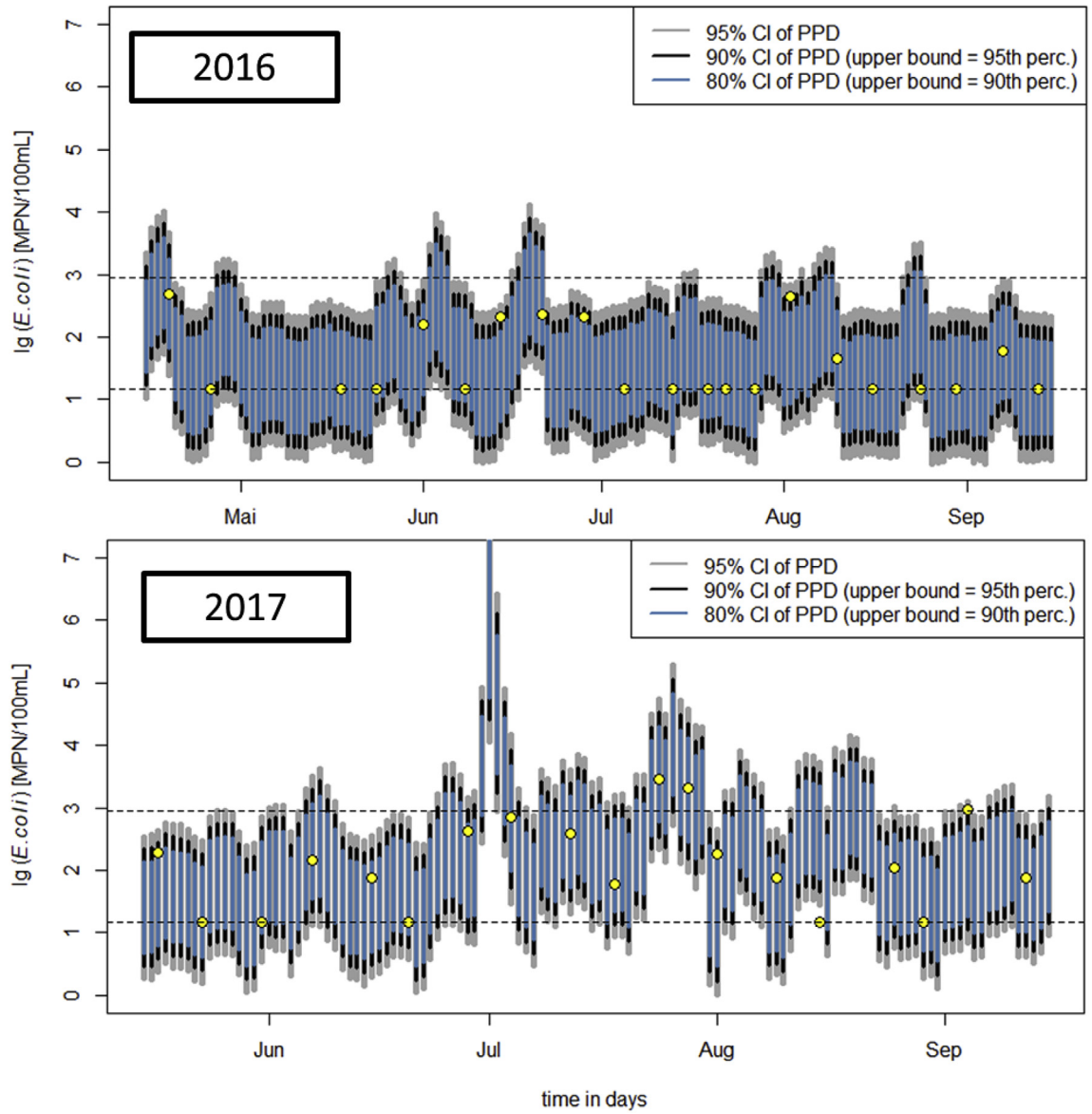

Fig. 4. Estimated distributions (model3) for each day in bathing seasons 2016 and 2017 based on evidence contained in the training data. Black horizontal lines indicate the LOD (lower line) and the percentile threshold for sufficient bathing water quality of 900 MPN/100 mL (upper line). Yellow dots indicate the measured surveillance data, which are used for percentage coverage validation. CI of PPD: credible interval of the posterior predictive distribution. Perc.: percentile. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

4.1. Comparison of proposed alert values to similar studies

Previous studies which focused on implementing early warning systems for bathing water management in Europe used the 90th percentile threshold for "sufficient quality" either as threshold for classification models (Stidson et al., 2012 for marine waters) or compared the maximum likelihood estimate (MLE) of regression models to this threshold (Herrig et al., 2015; Bedri et al., 2016). Mälzer et al. (2016) used an even higher concentration of 1800 MPN/100 mL as a classification threshold. Each of these approaches tolerates that the bulk of the probability mass of the measured data is located very close or even above the percentile threshold. When using the MLE as applied by (Herrig et al., 2015) the "probability of exceeding" is tolerated to be up to 50%. When using the approach proposed by Mälzer et al. (2016) up to 100% of the data points are tolerated to fall between 900 MPN and 1800 MPN, without warning. If 900 MPN is used as a classification threshold, 100% of the data should fall below the threshold. On the other hand, since the lognormal condition is not considered, measured data would be allowed to be located very close to the threshold. In contrast, if the lognormality condition is considered, the bulk of the probability mass has to be located much lower. As to our case, to determine "sufficient" water quality the evidence in the training data has to indicate that 50% of the data falls below 152 MPN/100 mL 80% below 488 MPN/100 mL, only 10% between 500 and 900 MPN/100 mL, and an additional 10% is tolerated above 900 MPN/100 mL. Thus, the approach proposed is much stricter regarding the expected frequency of data between 500 and 900 MPN/100 mL, while it will, like the BWD, tolerate occasional outliers. Fig. 8 illustrates the tolerable ranges and frequencies the different approaches allow the data to fall in without warning.

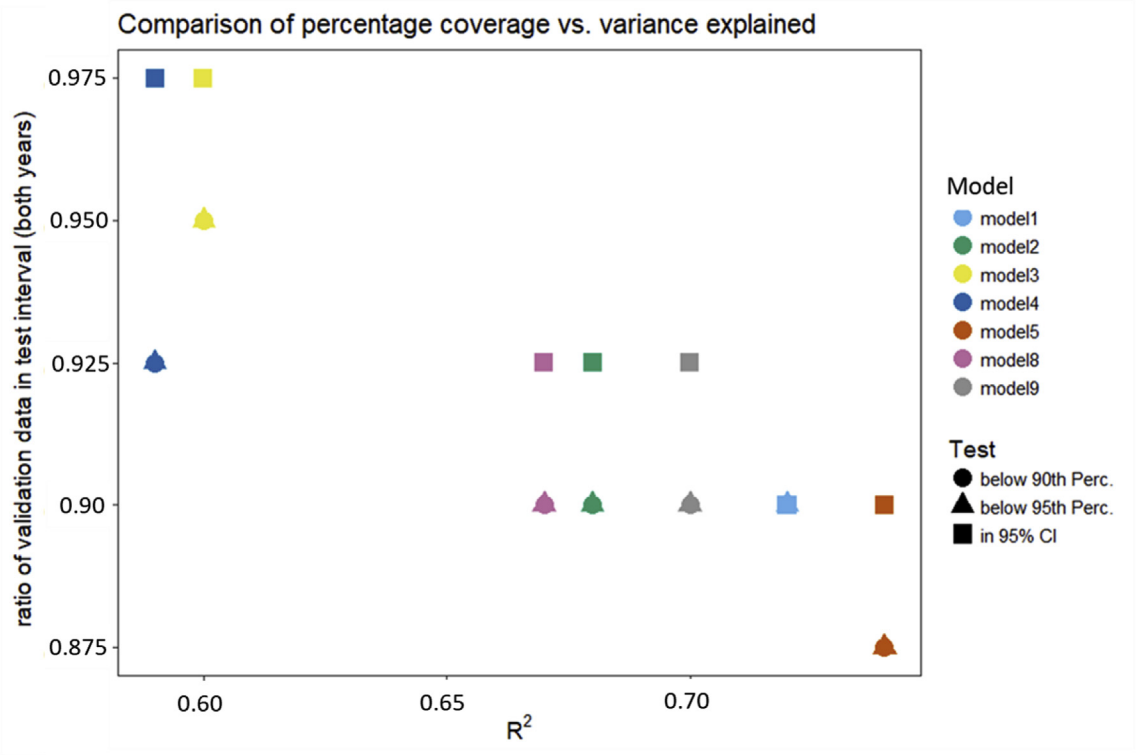

Fig. 5. Comparison of explained variance (mean R^2) vs. percentage coverage of validation data. Perc.: percentile, CI: credible interval of posterior predictive distribution.

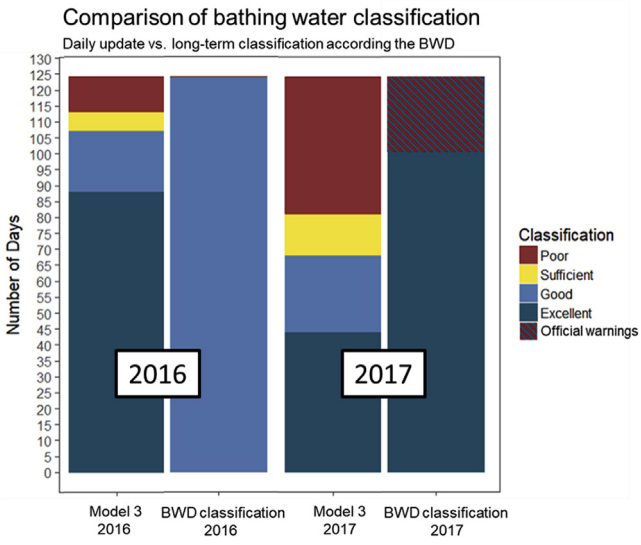

Fig. 6. Comparison of daily classifications made by model 3 in comparison to long term bathing water classification according to BWD and official warnings by the responsible health authorities.

4.2. Model validation by percentage coverage

The introduced model validation criteria based on percentage coverage provide a suitable way to apply probabilistic model checking when only periodic surveillance data are available. Moreover, it prevents overfitting and ensures that the residual standard deviation remains wide enough when models are intended to be used for prediction. The latter is particularly important for low sample sizes. The criterion has to be considered as an additional minimum requirement for checking model consistency against the test data and not as a proof that the predicted distribution represents the true variability during each individual day. To validate the latter, many grab samples in high timely resolution would be necessary. However, heavy rainfall, the major source of temporal variation (US-EPA, 2010), and accordingly of health concern, shows rather stochastic patterns. Therefore, grab samples collected periodically over long time-periods are likely to capture event-scale variations in the end. Overall, the approach showed very promising results at a large riverine bathing water site in Berlin and should therefore be tested and verified at other locations and for other types of bathing waters.

4.3. Sources of uncertainty

Assessment approaches which are based on statistical inference, i.e. which make decisions upon estimates of unobservable quantities like parametric 90th or 95th percentiles are always conditional on a) the chosen statistical model b) the quality and quantity of available data. This accounts for both the currently used method for long-term classification as well as for the proposed method for regression modelling.

4.3.1. Uncertainties resulting from (random) sampling error

All statistical inference is conditional on the observed data. In particular, the estimation of high percentiles is subject to statistical error due to low sample sizes (Berthouex and Hau, 1991). Moreover, if by random sampling no short-term pollution is detected and no other prior information is used to build the model, a statistical model might not be able to make correct predictions. Due to both reasons, inferences might be wrong. However, in comparison to the current BWD, which only uses the data from the previous four years, the presented approach certainly improves the inferences on bathing water quality since all available data as well as the information about their correlation to relevant predictors are used. By

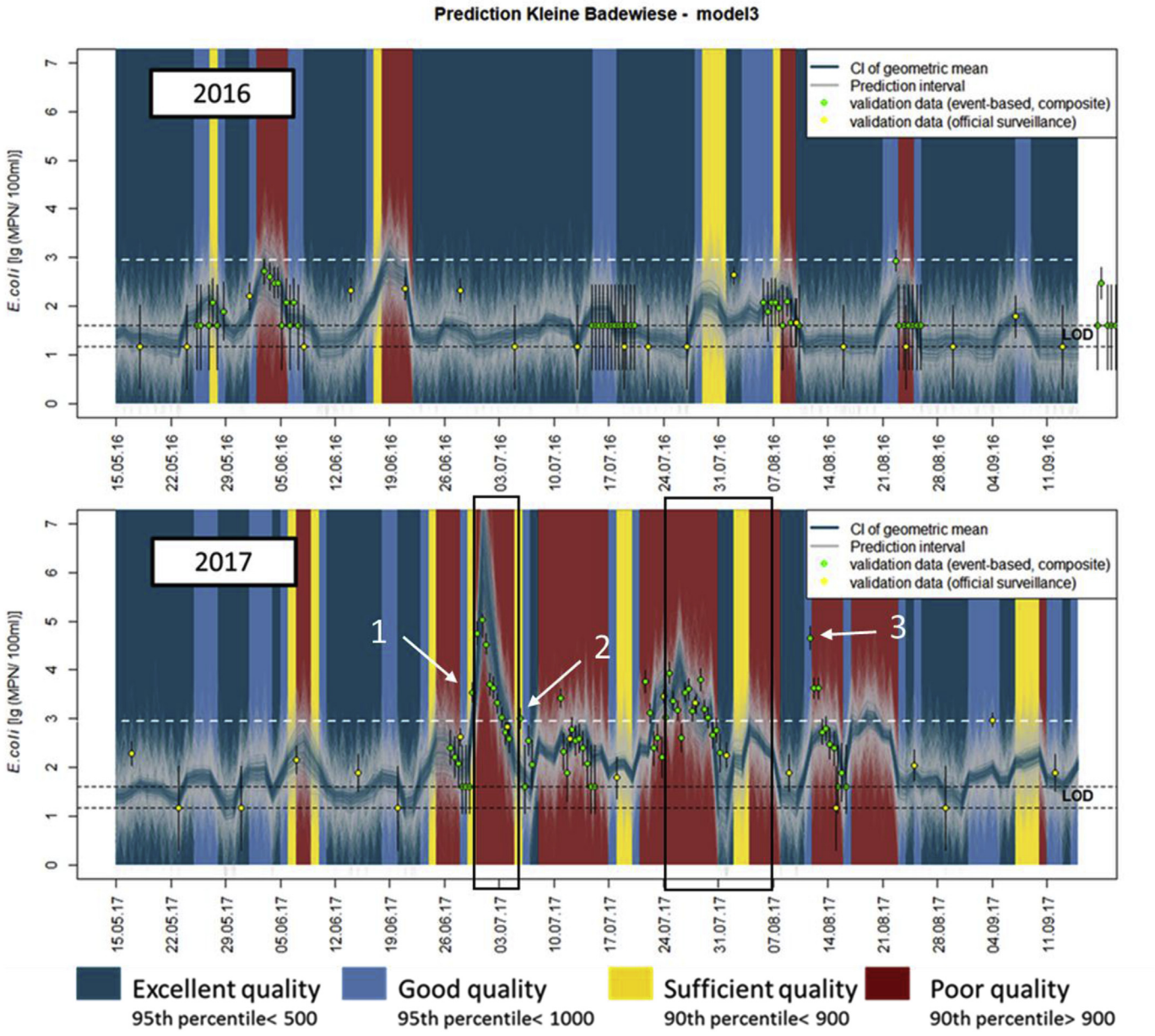

Fig. 7. Prediction, validation and classification of bathing water quality in 2016 and 2017. Figures show bathing water classification in 2016 and 2017 using model 3. The figures show the PPD as grey area, the credible interval of the geometric mean as blue line, the validation data (not used for model fitting) as green and yellow dots. Yellow dots: official surveillance data (weekly grab samples). Green dots: event-based composite samples. The dotted black horizontal lines indicate the different limits of detection (LOD) of 15 and 40 MPN/100 mL for the different dilutions of the event-based and surveillance data. The white line indicates the 90th percentile standard for "sufficient" quality. Background colors refer to the predicted classification. Black frames indicate periods of official warning (*ad-hoc*) by health authorities. White arrows and numbers: Elevated concentration events with model-predicted classification "sufficient" or better. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

updating models regularly as more data becomes available, uncertainty will be further reduced over time.

4.3.2. Uncertainties related to the statistical model

The estimation of high parametric percentiles of a $\log_{10}$-normal statistical model is sensitive to the choice of the lower detection limit (LOD). The reason for this sensitivity is that e.g. the difference between 10 MPN/100 mL and 1 MPN/100 mL, i.e. an absolute difference of 9 MPN/100 mL at the lower end of the scale represents a whole order of magnitude. Thus, it will have a larger effect on the estimate of the geometric standard deviation as e.g. an absolute difference of 500 MPN between 500 MPN and 1000 MPN (difference ~ 0.3 orders of magnitude). The BWD does not explicitly define requirements for the LOD. In the present study, we used data from official bathing water surveillance with a LOD of 15 MPN. This LOD follows recommendations according to the reference method outlined in the BWD (ISO 9308-3). Therefore, we consider this LOD as appropriately low or at least in line with current European legislation. However, due to the sensitivity of estimates of high percentiles to the lower LOD clear standards would be desirable.

5. Conclusions

- Statistical regression modelling offers a promising solution to translate the existing percentile thresholds for long-term classification to daily bathing water management.

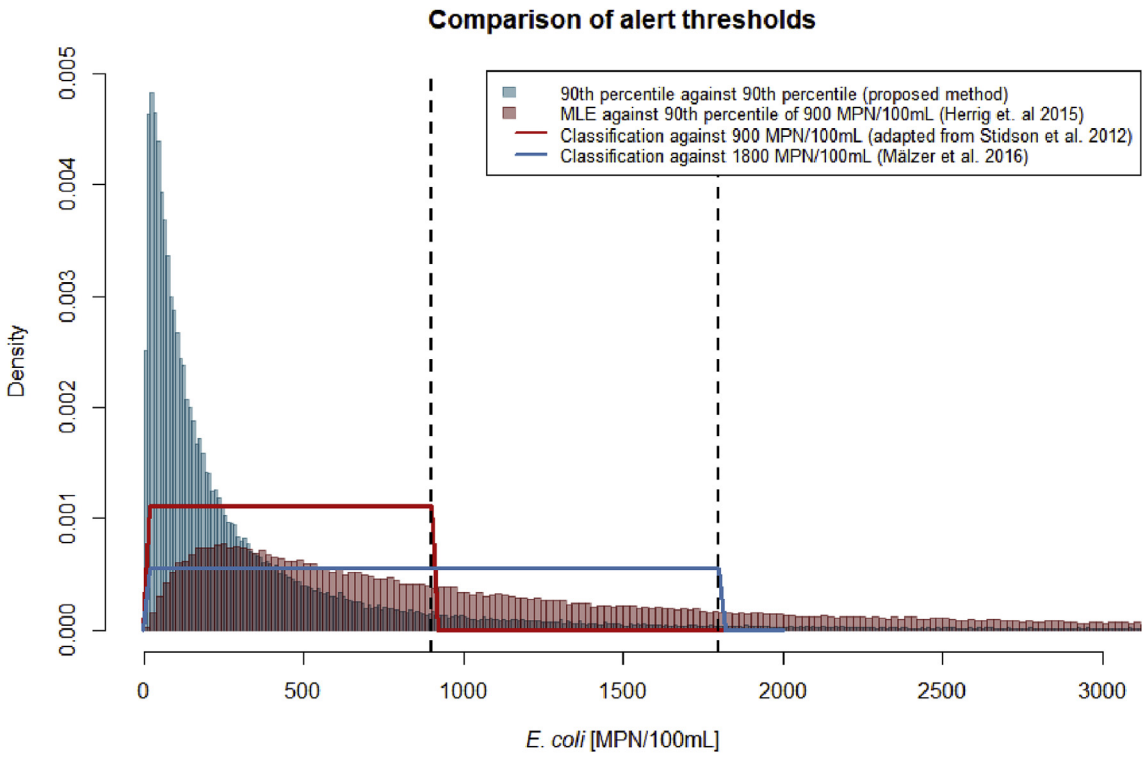

Fig. 8. Comparison of different approaches for early warning. The blue histogram indicates a $\log_{10}$-normal distribution, which fulfils exactly the minimum requirements for "sufficient" against the chosen model N ($\mu\bigcirc = 2.185$, $\sigma = 0.6$), the red histogram shows a lognormal distribution where the MLE is exactly at 900 MPN/100 mL. The SD of 0.5 is arbitrarily chosen N ($\mu\bigcirc = \log10\ (900)$, $\sigma = 0.5$). Blue and red lines indicate the regions in which the two classification approaches would allow measured data to fall in without warning. Dashed lines indicate the thresholds of 900 MPN/100 mL and 1800 MPN/100 mL. Stidson et al. (2012) conducted their study at marine waters, so the numerical values for the standard of "sufficient" are different. For conceptual comparison, we adapted the approach to the higher thresholds for inland waters. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

- The application of the derived percentage coverage criteria accounts for the probabilistic character of EU bathing legislation and reduces the risk of overfitting and thus overly optimistic prediction accuracy in comparison to optimizing models in terms of R^2.
- Event-based monitoring provided valuable additional information about periods of major fecal pollution and model checking.
- Regarding microbial safety, the application of the proposed methodology at a riverine bathing water site in Berlin demonstrated the shortcomings of current long-term classification as well as the potential for improvement by applying the proposed approach.
- Since statistical classification approaches do not account for the lognormality assumption, preference should be given to regression approaches including the corresponding uncertainty.

Acknowledgements

This research was conducted within the project FLUSSHYGIENE, "Hygienically relevant microorganism and pathogens in multifunctional surface water and water cycles: sustainable management of different river types in Germany". It was funded by the German Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung, BMBF) under sponsorship number 02WRS1278A. We would like to thank the Berlin Water Utilities (BWB), the Berlin Health Authorities (LaGeSo) as well as the Berlin Senate Department for the Environment, Transport and Climate Protection (SenUVK) for providing the necessary data. Thanks go to Julia Schmidt, Vincent Biosdeffre, Cristina Savaria Arzabe, Franziska Knoche and the Potsdam Laboratory for Water and Environment (PWU) for supporting the monitoring campaigns and ensuring rapid analysis of microbial samples. Prof. Gertjan Medema and Dr. Frederik Zietzschmann are thanked for their review and their valuable advice.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.watres.2018.06.057.

References

Bedri, Z., Corkery, A., O'Sullivan, J.J., Deering, L.A., Demeter, K., Meijer, W.G., O'Hare, G., Masterson, B., 2016. Evaluating a microbial water quality prediction model for beach management under the revised EU Bathing Water Directive. J. Environ. Manag. 167 (Suppl. C), 49–58.

Berthouex, P.M., Hau, I., 1991. Difficulties related to using extreme percentiles for water quality regulations. Res. J. Water Pollut. Contr. Fed. 63 (6), 873–879.

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.

Brady, A.M.G., 2007. Rapid Method for Escherichia coli in the Cuyahoga River.

Brooks, W.R., Fienen, M.N., Corsi, S.R., 2013. Partial least squares for efficient models of fecal indicator bacteria on Great Lakes beaches. J. Environ. Manag. 114 (Suppl. C), 470–475.

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.
76/160/EEC, 2006. In: 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. Community, E.
Heberger, Durant, J.L., Oriel, K.A., Kirshen, P.H., Minardi, L., 2008. Combining real-time bacteria models and uncertainty analysis for establishing health advisories for recreational waters. J. Water Resour. Plann. Manag. 134 (1), 73–82.
Herrig, I.M., Böer, S.I., Brennholt, N., Manz, W., 2015. Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. Water Res. 85 (Suppl. C), 148–157.
ISO 9308-3, 1998. In: Water Quality - Detection and Enumeration of Escherichia coli and Coliform Bacteria in Surface and Waste Water - Part 3: Miniaturized Method (Most Probable Number) by Inoculation in Liquid Medium (ISO 9308-3: 1998); German Version EN ISO 9308-3:1998.
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.
Mälzer, H.-J., aus 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, Part B), 671–680.
Motamarri, S., Boccelli, D.L., 2012. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Res. 46 (14), 4508–4520.
R Development Core Team, 2008. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
Reicherts, J.D., Emerson, C.W., 2010. Monitoring bathing beach water quality using composite sampling. Environ. Monit. Assess. 168 (1), 33–43.
Schumacher and Storz, 2016 (engl. by authors: Update of bathing water profile Lower Havel, Hydronumeric modelling of travel times). In: Fortschreibung Badegewässerprofil Unterhavel, Hydronumerische Modellierung der Fließzeiten, Auftraggeber: Landesamt für Gesundheit und Soziales, Berlin.
StanDevelopmentTeam, 2017a. RStanArm: Bayesian Applied Regression Modeling via Stan. R package version 2.15.3. http://mc-stan.org.
StanDevelopmentTeam, 2017b. Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. http://mc-stan.org.
Stidson, R.T., Gray, C.A., McPhail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water Environ. J. 26 (1), 7–18.
Traister, E., Anisfeld, S.C., 2006. Variability of indicator bacteria at different time scales in the upper Hoosic river watershed. Environ. Sci. Technol. 40 (16), 4990–4995.
US EPA, 2010. Sampling and Consideration of Variability (Temporal and Spatial) for Monitoring of Recreational Waters. EPA-823-R-10-1005, December 2010. U.S. Environmental Protection Agency, Office of Water.
Vehtari, A., Gelman, A., Gabry, J., 2016. Loo: Efficient Leave-one-out Cross-validation and WAIC for Bayesian Models. R package version 1.1.0. https://CRAN.R-project.org/package=loo.
Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27 (5), 1413–1432.