Water Research 85 (2015) 148–157

Contents lists available at ScienceDirect

Water Research

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

Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany

Ilona M. Herrig [a, b, *], Simone I. Böer [a, 1], Nicole Brennholt [a], Werner Manz [b]

[a] Federal Institute of Hydrology, Department G3 – Bio-Chemistry, Ecotoxicology, Am Mainzer Tor 1, 56068 Koblenz, Germany
[b] University of Koblenz-Landau, Department of Biology, Institute of Integrated Natural Sciences, Universitätsstraße 1, 56070 Koblenz, Germany

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form
31 July 2015
Accepted 2 August 2015
Available online 5 August 2015

Keywords:
Escherichia coli
Intestinal enterococci
Somatic coliphages
Bathing water quality
Monitoring
Management tool

ABSTRACT

Since rivers are typically subject to rapid changes in microbiological water quality, tools are needed to allow timely water quality assessment.

A promising approach is the application of predictive models. In our study, we developed multiple linear regression (MLR) models in order to predict the abundance of the fecal indicator organisms *Escherichia coli* (EC), intestinal enterococci (IE) and somatic coliphages (SC) in the Lahn River, Germany. The models were developed on the basis of an extensive set of environmental parameters collected during a 12-months monitoring period. Two models were developed for each type of indicator:

1) an extended model including the maximum number of variables significantly explaining variations in indicator abundance and 2) a simplified model reduced to the three most influential explanatory variables, thus obtaining a model which is less resource-intensive with regard to required data.

Both approaches have the ability to model multiple sites within one river stretch. The three most important predictive variables in the optimized models for the bacterial indicators were NH_4–N, turbidity and global solar irradiance, whereas chlorophyll *a* content, discharge and NH_4–N were reliable model variables for somatic coliphages.

Depending on indicator type, the extended mode models also included the additional variables rainfall, O_2 content, pH and chlorophyll *a*.

The extended mode models could explain 69% (EC), 74% (IE) and 72% (SC) of the observed variance in fecal indicator concentrations. The optimized models explained the observed variance in fecal indicator concentrations to 65% (EC), 70% (IE) and 68% (SC). Site-specific efficiencies ranged up to 82% (EC) and 81% (IE, SC). Our results suggest that MLR models are a promising tool for a timely water quality assessment in the Lahn area.

1. Introduction

Rivers serve as water sources, transportation routes and recreation areas. High levels of fecal indicator bacteria in surface waters constitute potential health risks and thus impair several water uses. The main source of allochthonous pathogenic microorganisms in rivers is agricultural and urban waste water originating from either diffuse sources or point sources such as sewage treatment plants and combined sewer overflows. The Lahn River, in particular, is famous for water sports like water skiing or canoeing and thus is of great touristic value for the region. Therefore, it is vital to identify and monitor potential health risks posed to those who engage in water-associated activities.

Abbreviations: AIC, Akaike information criterion; AR(1), autoregressive process of the first order; BfG, Federal Institute of Hydrology; EBWD, European Bathing Water Directive; EC, *Escherichia coli*; EU, European Union; IE, intestinal enterococci; MLR, multiple linear regression; OLS, ordinary least squares; REML, restricted maximum likelihood; SC, somatic coliphages; VIF, variance inflation factor; xd-sum, sum of x days including day of sampling (x either 2,3,4,5 or 6); RAIN, rainfall [mm].

* Corresponding author. Federal Institute of Hydrology, Department G3 – Bio-Chemistry, Ecotoxicology, Am Mainzer Tor 1, 56068 Koblenz, Germany.
E-mail address: herrig@bafg.de (I.M. Herrig).
[1] Present address: LUFA Nord-West, Institute for Food Quality, Ammerländer Heerstraße 115–117, 26129 Oldenburg, Germany.

http://dx.doi.org/10.1016/j.watres.2015.08.006
0043-1354/

Nomenclature

MUG	4-methylumbelliferyl-beta-D-glucuronide
COND	conductivity [μS/m]
CFU	colony forming units [CFU/100 mL]
FS	filterable solids [mg/L]
GSI	global solar irradiance [Wh/m^2]
h	hours
MPN	most probable number [MPN/100 mL]
Q	daily means of water discharge [m^3/s]
TN_b	total nitrogen bound [mg/L]
PFU	plaque forming units [PFU/100 mL]
PO_4-P	phosphate phosphorus [mg/L]

In Germany, bathing water quality assessments are specified under the European Bathing Water Directive 2006/7/EC (EBWD) (EU, 2006), which demands regular testing for the bacterial indicators *Escherichia coli* and intestinal enterococci. At present, the EBWD has not incorporated any viral indicators on the list of parameters to be tested, even though somatic coliphages have been proposed as indicators for viruses (Grabow, 2001).

For indicator detection, the EBWD recommends culture-dependent methods, which are quite time consuming and the analysis requires up to 48 h after sampling. Results obtained this way may not reflect the current hygienic situation since flowing waters underlie strong and rapid changes in water quality. The timeliness of water quality data can be improved by the application of predictive models, which may serve as a more appropriate tool to evaluate microbial quality and issue timely health warnings to the public. Therefore the development of different modelling approaches to predict fecal indicator concentrations is a current field of research.

Besides others, a common modelling approach is regression based modelling such as multiple linear regression (MLR). Given the accessibility, ease of implementation, their wide acceptance in the water resources, larger scientific and engineering communities (Mas and Ahlfeld, 2007), regression models proved to be a promising tool in water quality assessments.

Several studies have utilized linear regression techniques either to explore the importance of various explanatory variables or to predict bacteria concentrations in estuaries (Ferguson et al., 1996), coastal bathing beaches (Crowther et al., 2001), lakes (Francy et al., 2006; Nevers and Whitman, 2005; Olyphant and Whitman, 2004) and rivers (Christensen, 2001; Motamarri and Boccelli, 2012; David and Haggard, 2011; Brady et al., 2009; Eleria and Vogel, 2005; Rasmussen and Ziegler, 2003). Approaches targeting rivers and streams may be based solely on meteorologic (Vermeulen and Hofstra, 2014), hydrologic, physicochemical (Christensen, 2001) or land use variables (Hampson et al., 2010; Schoonover and Lockaby, 2006) or on combinations of these (Mas and Ahlfeld, 2007; David and Haggard, 2011; Eleria and Vogel, 2005), resulting in models with different predictive capacities.

Previous approaches have mainly focused on modelling sampling sites independently, which resulted in separate models for each site and indicator, as well as differing explanatory variables (David and Haggard, 2011; Brady et al., 2009; Christensen et al., 2002). If an application at a larger scale is desired, modelling each site independently will increase modelling and monitoring efforts considerably, thus complicating model development and application. Real-time prediction of bathing water quality is suggested as an appropriate regulatory approach in recent WHO Guidelines (WHO, 2003) and new European Union (EU) standards for bathing water quality (EU, 2006; Stidson et al., 2012).

Predictive models are increasingly implemented as management tools in the USA (Francy, 2009). Also in Scandinavia pilot projects have shown that early warning systems based on models are not only appropriate management tools, but can even open up new recreational attractions (Erichsen et al., 2003).

The Scottish Environment Protection Agency (SEPA) has developed the first EU real-time bathing water quality predictions at 10 sites throughout Scotland since 2004 (Stidson et al., 2012). Here, we present various approaches to model fecal indicators in the river Lahn, which may be the basis for other streams as well. Thereby, the aim of the present study was to develop multiple linear regression models, which

(i) allow a timely assessment of fecal indicators
(ii) address bacterial and viral indicators
(iii) do not require time and resource intensive data collection
(iv) are applicable to several sites within a river stretch
(v) are simple and as uniform as possible with respect to indicator types

2. Material and methods

2.1. Study area

The Lahn River originates in the Rothaar Mountains (North Rhine-Westphalia, Germany) and flows into the Rhine River at Lahnstein, near Koblenz (Rhineland-Palatinate, Germany). The Lahn is about 245 km long, with a mean annual discharge (MQ) of 46.6 m^3/s (refers to data from 1935 to 2010; gauge Kalkofen; http://undine.bafg.de/servlet/is/18941/; last access 2/2014) and a catchment area of approximately 5900 km^2, thus belonging to the smaller tributaries of the Rhine. It is strongly impounded and characterized by a large number of locks. The river stretch from Lahn-km −11.075 (near Giessen) to the mouth (Lahn-km 137.3) is designated as a federal waterway, which is predominantly used by smaller motor yachts, as well as paddle- and rowboats and it is frequently used for other recreational purposes like fishing, canoeing or water skiing.

The study area is situated in the lower Lahn Valley (Fig. 1). The river bed is deeply (approximately 200 m) incised into the Rhenish Slate Mountains and predominantly surrounded by forested slopes, as well as narrow strips of meadows and pastures at the valley bottom. Agriculture is concentrated on the heights surrounding the valley and plays a minor role in the valley itself. Overall, the Lahn valley is considered as a pristine and natural environment.

2.2. River water samples

In the present study five sites (Lahn 1–5) along the Lahn (Fig. 1) were monitored for fecal indicator concentrations as well as 15 physical, chemical, biological and hydro-meteorological parameters (Table 1). The sites were distributed along a river stretch of approximately 11.5 km and are located at differing distances to upstream point-source outlets to ensure that locations with varying degree of impact are sampled.

Surface water samples were collected according to the standard procedure ISO 19458:2006 (ISO, 2006). Sampling was conducted between 9:00 and 12.00 am. Samples were collected in sterile glass bottles with a telescopic bar, approximately 1 m off shore and at a depth of approximately 0.3 m. All samples were transported to the laboratory in cooler boxes and processed within 4 h after arrival.

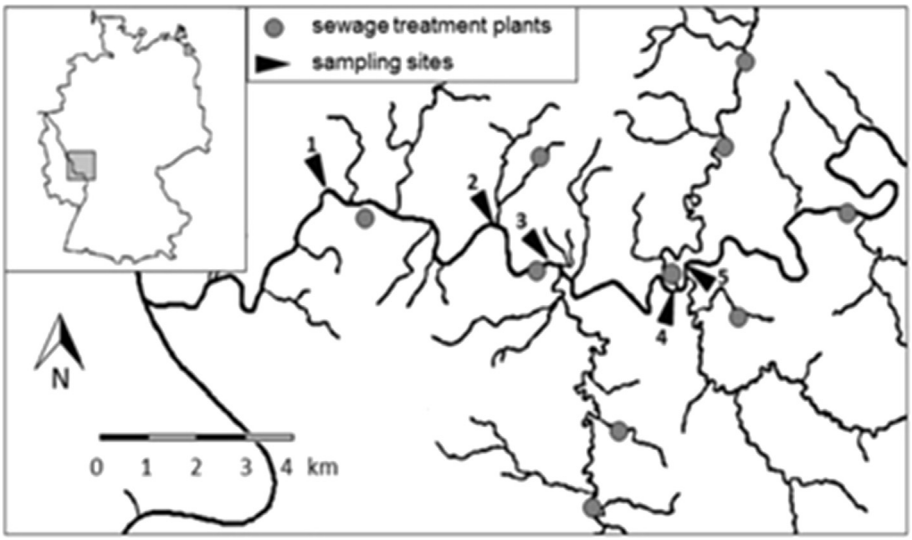

Fig. 1. Lahn River study reach with its tributaries. Location of sampling sites (Lahn 1–5) between Nievern and Obernhof, Rhineland-Palatinate, Germany are indicated as arrows, sewage treatment plants are shown as circles.

Table 1
Parameters, sites, methods and sampling intervals of the monitoring campaign.

Parameter	Abbreviation	Unit	Method	Site	Interval
E. coli	EC	MPN/100 mL	MUG[a]/EC micro plates	Lahn 1–5	weekly
Intestinal enterococci	IE	CFU/100 mL	membrane filtration	Lahn 1–5	weekly
Somatic coliphages	SC	PFU/100 mL	plaque -assay	Lahn 1–5	weekly
Filterable solids	FS	mg/L	filtration	Lahn 1–5	weekly
Water temperature	TEMP	°C	multiparameter sonde	Lahn 1–5	weekly
Conductivity	COND	µS/cm	multiparameter sonde	Lahn 1–5	weekly
pH	pH	–	multiparameter sonde	Lahn 1–5	weekly
Turbidity	TURB	NTU	multiparameter sonde	Lahn 1–5	weekly
Chlorophyll *a*	CHL	µg/L	multiparameter sonde	Lahn 1–5	weekly
Dissolved oxygen	O_2	mgL/	multiparameter sonde	Lahn 1–5	weekly
Nitrite nitrogen	NO_2–N	mg/L	photometric	Lahn 1–5	weekly
Nitrate nitrogen	NO_3–N	mg/L	photometric	Lahn 1–5	weekly
Ammonium nitrogen	NH_4–N	mg/L	photometric	Lahn 1–5	weekly
Total nitrogen bound	TN_b	mg/L	photometric	Lahn 1–5	weekly
Phosphate phosphorus	PO_4-P	mg/L	photometric	Lahn 1–5	weekly
Discharge (daily means)	Q	m^3/s	provided data	Kalkofen	continuously
Rainfall (daily totals)	RAIN	mm	provided data	Nassau	continuously
Global solar irradiance (daily totals)	GSI	Wh/m^2	provided data	Grenzau	continuously

[a] MUG: 4-methylumbelliferyl-beta-D-glucuronide.

A regular monitoring scheme, i.e., weekly sampling, was performed from December 2011 to December 2012. Additional samples were retrieved during high flow conditions or precipitation events, in order to include varying hydrologic conditions. Data obtained during this time provided the basis for the establishment and calibration of the predictive models, particularly to identify suitable independent variables. For validation purposes, additional samples were taken randomly from April to November 2013 (Table 2).

2.3. Fecal indicator analyses

Water samples were analysed for *E. coli* (EC) and intestinal enterococci (IE) according to the EBWD. In addition, concentrations of somatic coliphages (SC) were determined. EC were enumerated following ISO 9308-3:1998 (ISO, 1998). The tests were performed using standardized microtiter plates according to the instructions of the manufacturer and accompanying product-specific MPN tables (Dr. Brinkmann Floramed GmbH, Germany). Samples for IE

Table 2
Overview of sampling sites and periods of the regular monitoring (without parentheses) as well as additional validation data collection (in parentheses) with corresponding numbers of observations.

Site no.	Denotation	Location	Sampling period	Observations
1	Lahn 1	Nievern	12/2011–12/2012 (04/2013–11/2013)	49 (6)
2	Lahn 2	Dausenau	12/2011–12/2012 (04/2013–11/2013)	50 (7)
3	Lahn 3	Nassau	12/2011–12/2012 (04/2013–11/2013)	49 (7)
4	Lahn 4	Schloss Langenau	12/2011–12/2012 (04/2013–11/2013)	48 (6)
5	Lahn 5	Obernhof	12/2011–12/2012 (04/2013–11/2013)	49 (7)

determination were filtrated with sterile filters (PALL, 0.45 μm pore size) and incubated on Slanetz–Bartley agar (Oxoid) with a subsequent confirmation on bile aesculin agar (Merck) following ISO 7899-2:2000 (ISO, 2000a).

Procedures for SC enumeration were carried out using the phage plaque assay according to the ISO standard procedure ISO 10705-2:2000 (ISO, 2000b). To reduce the bacterial background in the samples, nalidixic acid was added to a final concentration of 250 mg/L. Sample volumes of 1–5 mL were applied in the test depending on the expected phage concentrations.

Counts were expressed as MPN/100 mL (EC), CFU/100 mL (IE) and PFU/100 mL (SC).

2.4. Physicochemical and biological parameters

Measurements of water temperature (TEMP), specific conductivity (COND), pH, turbidity (TURB), dissolved oxygen (O_2) and chlorophyll *a* (CHL) were conducted *in situ* with a YSI 6600 V2 multiparameter sonde (YSI, USA). In addition, the amount of filterable solids (FS) was determined in mg/L dry weight. For this purpose 1 L of Lahn water was filtered with previously rinsed and dried glass fibre filters and subsequently dried for 2 h at 105 °C. Nutrient concentrations were measured spectrophotometrically (Xion 500, Hach-Lange, Germany) using cuvette tests (Hach-Lange, Germany) LCK 341/342 (NO_2–N), LCK 339 (NO_3–N), LCK 304 (NH_4–N), LCK 349 (PO_4-P) and LCK 138 (total nitrogen bound, TN_b) according to product instructions.

2.5. Hydro-meteorological data

Global solar irradiance (weather station Grenzau, Germany) and precipitation data (weather station Nassau, Germany) were obtained from the service centres for the rural area of Rheinland-Palatinate (www.am.rlp.de, last access 02/2014) and reported as daily total values. Daily mean values of water discharge (gauge Kalkofen) were provided by the German Federal Institute of Hydrology (BfG), Koblenz, Germany. An overview of all determined parameters with corresponding methods and sampling intervals is shown in Table 1.

2.6. Data preparation and correlation analyses

Correlation and multiple linear regression analyses as well as model validation were computed with the open-source software R (version 3.1.0, R Core Team, 2014).

In few cases, irregularities such as uneven or excessive growth of bacterial indicators, sensor failures or ice-coverage resulted in data gaps. Observations with missing values were omitted from further analyses, resulting in different numbers of observations per site (Table 2). At least 48 single observations were recorded for each site, and the comprehensive dataset from all 5 sites comprised a total of 248 observations (Table 2).

The lower and upper detection limit of EC concentrations were 15 MPN/100 mL and 34660 MPN/100 mL, respectively. Measurements outside these limits were assigned a value of 7.5 MPN/100 mL (lower detection limit) and 35000 MPN/100 mL (upper detection limit) for Spearman's rank correlation analyses. Conditions prevailing prior to sampling were accounted for by calculating sums of individual data for continuously measured data (rainfall, global solar irradiance, discharge). For this purpose, the daily measurements were summed up in data sets encompassing up to 6 consecutive days, starting at the day of each sampling, resulting in 6 data sets for each variable. This temporal synchronization has been shown to improve the predictive power of explanatory variables in MLR models significantly (Cyterski et al., 2012).

Finally, a global Spearman's rank correlation analysis was conducted in order to identify appropriate correlations between the dependent and the independent variables as well as correlations among the independent variables themselves. Correlation analyses were performed i) on the entire dataset including data from all sites and ii) for each site separately. Correlation analyses were applied on the z-standardized data using the *cor*() function in *R*. Significances were corrected using the Bonferroni algorithm to avoid multiple comparisons (e.g. Ramette, 2007). Prior to modelling, data were transformed according to Box and Cox (1964) in order to normalize the data and to stabilize variance. Lambda values, as a measure for best power of the Box-Cox-transformations, were calculated using the *boxcox*() function implemented in R in the *MASS* package (Venables and Ripley, 2002, Table 3).

2.7. Variable selection and fitting the model

Variables, which did not correlate with fecal indicator concentrations significantly (Bonferroni-corrected $p > 0.00033$) were omitted. The temporal synchronized data (1 to 6-days-sums) of the continuously measured variables (rainfall, global solar irradiance, discharge) correlating best with fecal indicator concentrations were chosen for further analyses.

In order to avoid collinearity, independent Box-Cox-transformed variables correlating with each other ($r > 0.69$ and $r < -0.69$, respectively) were also excluded, whereby the variables correlating the best with fecal indicator concentrations were retained.

Collinear variables were removed by a stepwise iterative variance inflation factor (VIF) analysis as described elsewhere (http://www.r-bloggers.com/collinearity-and-stepwise-vif-selection/, last access 01-06-2014). VIF values were calculated for all explanatory variables and the variable with the highest value was removed. This process was repeated until all VIF values are below the threshold of 5. The remaining variables (VIF < 5) were included in the subsequent analyses. Preliminary linear models for each type of indicator were generated by ordinary least squares (OLS) estimation.

The set of explanatory variables for the extended mode models was identified by automatic forward selection applying the *step*() command (*stats* package in R). Variables were retained in the model on a significance level of $p < 0.05$.

Optimized models were created to comply with the aim of generating models offering highest simplicity, containing fewer but consistent variables. The most suitable subsets for each indicator were chosen by an exhaustive iterative procedure using the *regsubset*() command (*leaps* package (Lumley, 2009)). Autocorrelation was accounted for by refitting the preliminary OLS models using the *gls*() function in the *nlme* package (Pinheiro et al., 2007) in R, which fits a linear model using generalized least squares. The adequate structure was selected using AIC (Akaike information criterion, Akaike, 1973). Models were finally fitted for an AR(1) process (autoregressive process of the first order) by restricted maximum likelihood (REML) that allows the fitting of an autoregressive term.

2.8. Model selection and evaluation

The best extended mode model was chosen on basis of the lowest AIC, whereas the best optimized models were selected based on a combination of the lowest AIC, the best model R^2 and equal explanatory variables incorporated in the different models.

Models were validated by running them on data collected separately in 2013 (Table 2). The performance assessment was based on R^2 values of observed vs. predicted indicator concentrations. R^2 values were calculated for the entire calibration data (12/

Table 3
Overview of applied Box–Cox transformations.

Best lambda	1.5 to 2.5	0.75 to 1.5	0.25 to 0.75	−0.25 to 0.25	−0.75 to −0.25	−1.5 to −0.75
Transformation	Square	None	Square root	log_{10}	Inverse square root	Reciprocal
Variables transformed accordingly	COND	O_2	TEMP	EC	Q	NO_3–N
	pH		RAIN	IE	TURB	
			GSI	SC	CHL	
				NO_2–N	TN_b	
				NH_4–N		
				FS		

2011–12/2012) set including all sampling sites, for each site individually and for the validation data (04–11/2013).

Further, the ability of the models to predict violations of assigned reference values correctly, i.e. false negative and false positive rates, was assessed for the calibration data as well as for the validation data, respectively.

Reference values were defined as 900 MPN EC per 100 mL and 330 CFU IE per 100 mL. According to the EBWD, a site is classified as 'sufficient' if the 90th percentiles are below these values. At present, no regulatory reference values or thresholds are defined for SC, thus these calculations were restricted to the EC- and IE models solely.

3. Results

3.1. Fecal indicator concentrations

Overall, highest mean concentrations were recorded for EC, followed by SC and IE with the lowest mean concentrations. The highest mean EC, IE and SC concentrations were detected at site Lahn 4 (6268 MPN/100 mL, 1081 CFU/100 mL, 1922 PFU/100 mL). The lowest mean EC and IE concentrations were observed at site Lahn 2 (2866 MPN/100 mL, 654 CFU/100 mL), and the lowest mean SC concentrations at site Lahn 1 (1530 PFU/100 mL). Overall, indicator concentrations varied considerably (Table 4).

3.2. Correlation analyses

Correlation analyses performed on the complete dataset revealed that the environmental parameters, except PO_4-P, correlated significantly with the three fecal indicators investigated (Table 5). All fecal indicators correlated significantly with each other, whilst bacterial indicators correlated stronger with each other ($r = 0.92$, $n = 245$, $p < 0.00033$), compared to correlations of the bacterial indicators with somatic coliphages (EC: $r = 0.80$, $n = 245$, $p < 0.00033$; IE: $r = 0.76$, $n = 245$, $p < 0.00033$). Negative correlations with EC, IE and SC were determined for the parameters temperature, conductivity, pH, chlorophyll *a* and global solar irradiance. In contrast, discharge, filterable solids, turbidity, O_2, rainfall, NO_2–N, NO_3–N, NH_4–N and TN_b were positively correlated with fecal indicators. Fecal indicators showed the strongest correlations with the 3-days-sums of solar irradiance ($GSI_{(3d\text{-}sum)}$), the 4-days-sums of rainfall ($RAIN_{(4d\text{-}sum)}$) and with the discharge on the day of sampling. Overall, strongest correlations with EC, IE and SC were found for NH_4–N and global solar irradiance (3-days-sums) with correlation coefficients of 0.73 (EC), 0.74 (IE), 0.72 (SB) and −0.66 (EC), −0.69 (IE) and −0.66 (SC).

The sampling sites are mostly distinguished by the strengths of correlations and fecal indicator abundances (Table 5).

3.3. Data selection

The 3-days-sums of global solar irradiance, the 4-days-sums of rainfall and discharge data from the day of sampling were chosen for further analyses.

Filterable solids correlated strongly with turbidity. Temperature correlated with both, O_2 and global solar irradiance. Conductivity showed a strong correlation with discharge. Further correlations were found between NO_2–N and NH_4–N. In contrast, NO_3–N correlated with total bound nitrogen as well as global solar irradiance. Variables showing the weakest correlations with fecal indicator concentrations were excluded.

Consequently, the parameters discharge, pH, turbidity, chlorophyll *a* content, O_2 content, 4-days-sums of rainfall, NH_4–N, total bound nitrogen and the 3-days-sums of global solar irradiance were considered as explanatory variables (Table 6). A subsequent

Table 4
Overview of fecal indicator concentrations at the five sampling sites at the river Lahn throughout the study period (12/2011–12/2012).

	Site	Unit	Mean	Median	Range	SD	n
E. coli	Lahn 1	MPN/100 mL	3149	1406	212–27730	4600	49
	Lahn 2	MPN/100 mL	2866	573	30–27730	5813	50
	Lahn 3	MPN/100 mL	<2975	<640	<15 – >34660	6315	49
	Lahn 4	MPN/100 mL	<6268	<1985	61 – >34660	9734	48
	Lahn 5	MPN/100 mL	2537	647	15–16740	4424	49
	Total	MPN/100 mL	<3545	<943	<15 – >34660	6538	245
Intestinal enterococci	Lahn 1	CFU/100 mL	730	220	75–10150	1560	49
	Lahn 2	CFU/100 mL	654	160	16–7400	1409	50
	Lahn 3	CFU/100 mL	739	180	2–7900	1583	49
	Lahn 4	CFU/100 mL	1081	292	27–11125	2075	48
	Lahn 5	CFU/100 mL	679	135	3–11450	1761	49
	Total	CFU/100 mL	775	190	2–11450	1683	245
Somatic coliphages	Lahn 1	PFU/100 mL	1530	1180	120–6760	1561	49
	Lahn 2	PFU/100 mL	1644	870	30–10100	2061	50
	Lahn 3	PFU/100 mL	1538	870	40–8300	1908	49
	Lahn 4	PFU/100 mL	1922	1005	20–21500	3181	48
	Lahn 5	PFU/100 mL	1601	640	60–8550	2083	49
	Total	PFU/100 mL	1646	960	20–21500	2207	245

Table 5
Overview of Spearman–Rank correlations between environmental variables and fecal indicator concentrations. Correlations significant at the Bonferoni-corrected level ($p < 0.00033$) are highlighted (bold).

		EC	IE	SC	Q	FS	TEMP	COND	pH	TURB	CHL	O_2	$RAIN_{4d\text{-}sum}$	$GSI_{3d\text{-}sum}$	NO2N	NO3N	NH4N	TNb
Lahn 1	**EC**	–	**0.90**	**0.75**	0.40	0.01	−0.42	−0.13	−0.29	0.24	**−0.52**	0.14	0.48	**−0.68**	0.33	**0.56**	**0.70**	0.45
Lahn 2	**EC**	–	**0.89**	**0.81**	**0.65**	0.27	**−0.49**	−0.30	−0.40	**0.51**	−0.31	0.31	**0.65**	**−0.65**	**0.52**	**0.54**	**0.73**	0.40
Lahn 3	**EC**	–	**0.89**	**0.81**	**0.72**	0.29	**−0.56**	−0.37	−0.48	0.48	−0.20	0.34	**0.64**	**−0.71**	**0.56**	0.44	**0.76**	0.48
Lahn 4	**EC**	–	**0.94**	**0.86**	**0.66**	**0.58**	−0.56	−0.43	−0.38	**0.71**	−0.10	0.43	**0.56**	**−0.67**	**0.68**	0.22	**0.80**	0.42
Lahn 5	**EC**	–	**0.90**	**0.81**	**0.65**	0.05	**−0.52**	−0.34	−0.46	0.26	−0.43	0.35	**0.58**	**−0.71**	0.48	0.42	**0.72**	**0.52**
overall	**EC**	–	**0.92**	**0.80**	**0.61**	**0.30**	**−0.50**	**−0.33**	**−0.38**	**0.49**	**−0.27**	**0.33**	**0.56**	**−0.66**	**0.51**	**0.38**	**0.73**	**0.40**
Lahn 1	**IE**	**0.90**	–	**0.75**	0.42	−0.03	**−0.52**	−0.16	−0.35	0.24	**−0.50**	0.24	0.48	**−0.79**	0.35	**0.64**	**0.72**	0.56
Lahn 2	**IE**	**0.89**	–	**0.74**	**0.51**	0.18	−0.46	−0.17	−0.46	0.44	−0.35	0.27	**0.59**	**−0.70**	0.46	**0.55**	**0.77**	0.48
Lahn 3	**IE**	**0.89**	–	**0.79**	**0.61**	0.28	−0.45	−0.24	**−0.49**	0.47	−0.27	0.23	**0.57**	**−0.68**	**0.49**	0.49	**0.71**	**0.50**
Lahn 4	**IE**	**0.94**	–	**0.82**	**0.66**	**0.58**	**−0.53**	−0.43	−0.44	**0.71**	−0.16	0.36	**0.56**	**−0.71**	**0.66**	0.23	**0.83**	0.46
Lahn 5	**IE**	**0.90**	–	**0.75**	**0.59**	0.04	−0.44	−0.28	−0.49	0.26	−0.41	0.26	**0.51**	**−0.68**	0.41	0.44	**0.73**	**0.55**
overall	**IE**	**0.92**	–	**0.76**	**0.56**	**0.26**	**−0.48**	**−0.27**	**−0.41**	**0.46**	**−0.30**	**0.29**	**0.54**	**−0.69**	**0.48**	**0.43**	**0.74**	**0.45**
Lahn 1	**SC**	**0.75**	**0.75**	–	**0.62**	0.22	−0.46	−0.36	−0.33	**0.50**	−0.41	0.24	0.46	**−0.65**	**0.56**	0.47	**0.73**	0.39
Lahn 2	**SC**	**0.81**	**0.74**	–	**0.61**	0.17	**−0.52**	−0.29	−0.41	0.45	−0.36	0.35	0.48	**−0.68**	**0.51**	**0.58**	**0.74**	**0.50**
Lahn 3	**SC**	**0.81**	**0.79**	–	**0.69**	0.30	**−0.55**	−0.39	**−0.56**	0.48	−0.35	0.36	**0.50**	**−0.69**	**0.60**	0.45	**0.76**	**0.58**
Lahn 4	**SC**	**0.86**	**0.82**	–	**0.75**	**0.66**	**−0.62**	**−0.51**	−0.41	**0.74**	−0.08	**0.54**	0.46	**−0.62**	**0.69**	0.23	**0.72**	0.37
Lahn 5	**SC**	**0.81**	**0.75**	–	**0.73**	0.05	**−0.58**	−0.48	**−0.53**	0.32	−0.45	0.41	**0.49**	**−0.69**	**0.59**	0.44	**0.68**	**0.50**
overall	**SC**	**0.80**	**0.76**	–	**0.68**	**0.29**	**−0.55**	**−0.41**	**−0.43**	**0.51**	**−0.33**	**0.39**	**0.47**	**−0.66**	**0.58**	**0.42**	**0.72**	**0.45**

Table 6
Overview of explanatory variables in the models.

Explanatory variable	Excluded	Considered	Included	
			Extended mode models	Optimized models
Filterable solids	•			
Water temperature	•			
Conductivity	•			
pH		•	•	
Turbidity		•	•	•
Chlorophyll *a*		•	•	•
Dissolved oxygen		•	•	
NO_2–N	•			
NO_3–N	•			
NH_4–N		•	•	•
TN_b		•		
PO_4-P	•			
Discharge		•	•	•
$Rainfall_{(4d\text{-}sum)}$		•	•	
Global solar $irradiance_{(3d\text{-}sum)}$		•	•	•

stepwise VIF selection confirmed that all collinear terms were successfully removed (VIF < 5).

3.4. MLR models

The sets of predictors included ($p < 0.05$) in the extended mode models differ partially between the three indicator types. The extended mode model for IE contains 6 predictor variables; the models for EC and SC contain 7 predictor variables each. Common predictors for all indicator types were NH_4–N, discharge, the 3-days-sums of global solar irradiance and turbidity. The inclusion of chlorophyll *a* content, the 4-days-sums of rainfall, pH and O_2 depended on indicator type (Table 7).

The optimized models of EC and IE were based on turbidity, NH_4–N and the 3-days-sums of global solar irradiance, whereas SC concentrations were predicted by discharge, chlorophyll *a* and NH_4–N content (Table 7, Table 8, Fig. 2).

Overall, R^2 values ranged from 0.69 to 0.74 (extended mode models) and from 0.64 to 0.70 (optimized models).

Model accuracy varied depending on the type of indicator and sampling location. Site-specific R^2 values ranged from 0.58 to 0.82 (extended mode models) and from 0.54 to 0.78 (optimized models). Regarding the different indicator types, best overall model performance was achieved for IE (extended mode model: $R^2 = 0.74$;

Table 7
Model equations.

Model type	Response variable	Model equation
EC_{ext}	$\log_{10}$ EC	$-2.62 \times Q^{-0.5} - 0.56 \times TURB^{-0.5} + 0.62 \times \log_{10} NH4N - 0.005 \times GSI_{(3d\text{-}sum)}^{0.5} + 0.52 \times CHL^{-0.5} + 0.097 \times RAIN_{(4d\text{-}sum)}^{0.5} + 4.53$
EC_{opt}	$\log_{10}$ EC	$-1.02 \times TURB^{-0.5} + 0.77 \times \log_{10} NH4N - 0.01 \times GSI_{(3d\text{-}sum)}^{0.5} + 5.21$
IE_{ext}	$\log_{10}$ IE	$-3.12 \times Q^{-0.5} - 0.57 \times TURB^{-0.5} + 0.63 \times \log_{10} NH4N - 0.01 \times GSI_{(3d\text{-}sum)}^{-0.5} - 0.11 \times O_2 + 0.01 \times pH^2 + 5.24$
IE_{opt}	$\log_{10}$ IE	$-1.00 \times TURB^{-0.5} + 0.70 \times \log_{10} NH4N - 0.01 \times GSI_{(3d\text{-}sum)}^{0.5} + 4.47$
SC_{ext}	$\log_{10}$ SC	$-2.69 \times Q^{-0.5} - 0.48 \times TURB^{-0.5} + 0.28 \times \log_{10} NH4N - 0.003 \times GSI_{(3d\text{-}sum)}^{0.5} + 0.71 \times CHL^{-0.5} + 0.01 \times pH^2 + 0.04 \times RAIN_{(4d\text{-}sum)}^{0.5} + 3.26$
SC_{opt}	$\log_{10}$ SC	$-4.08 \times Q^{-0.5} + 0.40 \times \log_{10} NH4N + 0.63 \times CHL^{-0.5} + 4.06$

Abbreviations as follows: *E. coli* (EC). intestinal enterococci (IE). somatic coliphages (SC). optimized model (opt). extended mode model (ext). global solar irradiance (GSI). turbidity (TURB).3-days-sum of global solar irradiance ($GSI_{(3d\text{-}sum)}$). water discharge (Q). chlorophyll *a* (CHL). 4-days-sum of rainfall ($RAIN_{(4d\text{-}sum)}$). oxygen content (O_2).

Table 8
Comparison of R^2 values of extended and optimized model types. R^2 values were calculated for the entire data set obtained from 12/2011–12/2012 for model calibration (total), each site separately (Lahn 1–5) and for the validation data (04/2013–11/2013).

Model type	Lahn 1	Lahn 2	Lahn 3	Lahn 4	Lahn 5	Total	Validation data
EC_{ext}	0.58	0.76	0.82	0.77	0.71	0.69	0.73
EC_{opt}	0.54	0.69	0.76	0.74	0.64	0.65	0.62
IE_{ext}	0.71	0.77	0.78	0.81	0.74	0.74	0.64
IE_{opt}	0.68	0.75	0.71	0.78	0.68	0.70	0.67
SC_{ext}	0.73	0.68	0.76	0.67	0.81	0.72	0.71
SC_{opt}	0.68	0.64	0.71	0.63	0.78	0.68	0.76

Abbreviations: *E. coli* (*EC*). intestinal enterococci (*IE*). somatic coliphages (*SC*). optimized model (opt). extended mode model (ext).

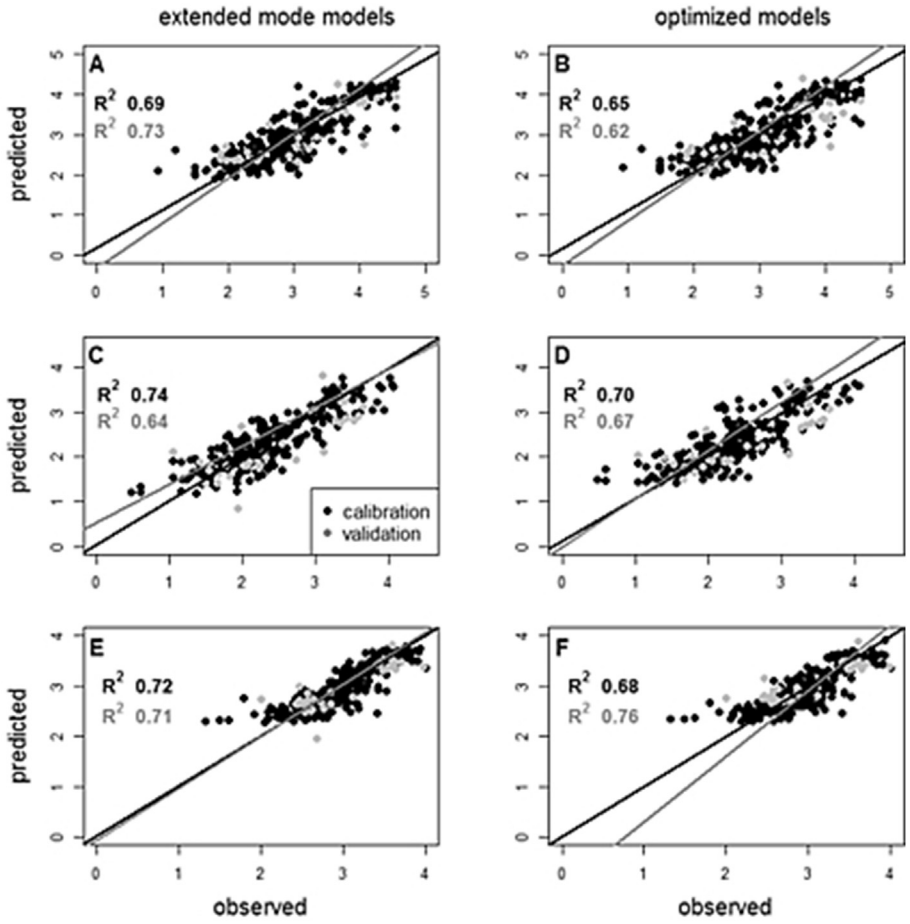

Fig. 2. Observed log indicator counts and predicted log indicator counts based on developed models for *E. coli* (A, B), intestinal enterococci (C, D) and somatic coliphages (E, F). Data used for model development and calibration (12/2011–12/2012) in black; model validation data (04/2013–11/2013) in grey.

optimized model: $R^2 = 0.70$), followed by SC (extended mode model: $R^2 = 0.72$; optimized model: $R^2 = 0.68$) and EC (extended mode model: $R^2 = 0.69$; optimized model: $R^2 = 0.65$). Concentrations of EC were modelled best at site Lahn 3 ($R^2 = 0.82$), IE concentrations at site Lahn 4 and SC concentrations at site Lahn 5 (Table 8).

3.5. Validation of model performance

Predictive models can be applied to estimate densities of fecal indicators or to determine if threshold criteria are met.

In the calibration data, 124 out of 245 observations (51%) exceeded the threshold of 900 MPN/100 mL (log 2.95) for EC, and 81 out of 245 observations (33%) exceeded the reference value of 330 MPN/100 mL (log 2.52) for IE. These thresholds are defined by 90-percentile calculations to classify a sufficient water quality in the EBWD. In the validation data, 19 (58%) observations exceeded the threshold for EC and 16 (48%) the one for IE.

The prediction of the threshold violations of the models was compared with actual violations determined in the calibration and validation data set. The extended mode model for EC successfully predicted 100 threshold violations (81%) in the calibration data and 13 (68%) in the validation data. For IE, the extended mode model successfully predicted 65 violations (80%) in the calibration data and 16 (48%) in the validation data. No differences were detected in predicted threshold violations of the validation data when comparing the extended mode- and the optimized models. Overall, no clear trend for over- or underestimation can be observed (Table 9).

Fecal indicator concentrations modelled on basis of validation

Table 9
Threshold violations as observed and modelled with corresponding false negative and false positive counts. Threshold was defined as 2.95 (log_{10} of EC concentrations) and 2.52 (log_{10} of IE concentrations), respectively and corresponded to the value applied for the percentile calculation for sufficient water quality as specified in the EBWD.

Model type	Threshold exceedings observed	Threshold exceedings modelled	True positives	False negatives	False positives
Calibration data					
EC_{ext}	124 (51%)	117	100 (81%)	24	17
EC_{opt}	124 (51%)	105	92 (74%)	32	13
IE_{ext}	81 (33%)	88	65 (80%)	16	23
IE_{opt}	81 (33%)	83	63 (78%)	18	20
Validation data					
EC_{ext}	19 (58%)	14	13 (68%)	1	6
EC_{opt}	19 (58%)	14	13 (68%)	1	6
IE_{ext}	16 (48%)	14	13 (81%)	1	3
IE_{opt}	16 (48%)	14	13 (81%)	1	3

Abbreviations: *E. coli* (*EC*). intestinal enterococci (*IE*). somatic coliphages (*SC*). optimized model (opt). extended mode model (ext).

data compared to actually measured concentrations could be also predicted with fairly high efficiency (R^2 0.62 to 0.76). The best predictions of the IE and SC concentrations of the validation data was obtained by the optimized models (R^2 0.67 and 0.76, respectively), with the most precise prediction achieved for SC (R^2 0.76). EC concentrations were more precisely predicted by the extended mode model (R^2 0.73, Table 8, Fig. 2).

4. Discussion

4.1. Fecal indicator abundance and recreational use

The monitoring results revealed high variations in water quality in the Lahn during the study period. Water quality was not acceptable according to the EBWD in more than half (EC) and one third (IE) of the tested samples, respectively. Most samples that exceeded concentrations of 900 MPN/100 mL (EC) and 330 CFU/100 mL (IE), respectively, were collected during the winter months when flow rates were high due to precipitation events. In winter, less intensive recreational use can be anticipated. However, high concentrations also appeared occasionally in spring and summer months. Since the Lahn is frequently used for canoe trips and other recreation activities during summer, days of increased fecal indicator concentrations in this season are of particular concern. To assess health risks of water recreation activities or other water uses, a method that allows a timely determination of water quality would be beneficial.

4.2. Model variables

Correlation analyses and input selection helped to identify the most important explanatory variables. Thus, results of our study contribute to the understanding of the factors affecting fecal indicator concentrations along a river stretch.

Discharge, turbidity, NH_4–N and the 3-days-sums of global solar irradiance were identified to be the most influential parameters in this study. These variables strongly correlated with fecal indicators and are included in all extended mode models.

NH_4–N is the only predictor variable included in all model types for each tested indicator organism with the strongest correlations, demonstrating its importance. It has rarely been used in predictive models before. However, statistical significant correlations between fecal indicator bacteria and ammonia concentrations in stream waters have already been described (David and Haggard, 2011; Francy et al., 2000).

NH_4–N is a known indicator of sewage contamination. As agriculture plays only a minor role in the study area, NH_4–N is likely to mainly originate from wastewater treatment plants. High NH_4–N concentrations also coincided with increased fecal indicator loads and discharge. Thus, due to extensive rainfall, capacities of sewer systems may have been exceeded, resulting in an overflow of untreated sewage into the river and an increase in fecal indicator and NH_4–N concentrations.

Global solar irradiance could be identified as the greatest factor contributing to fecal indicator decrease in correlation analyses. This finding is supported by other studies, in which global solar irradiance has been identified as the most significant factor affecting indicator bacteria decay rates (Burkhardt III et al., 2000; Craggs et al., 2004; Sinton et al., 2002). Global solar irradiance was incorporated in all models except the optimized model for SC; hence it was less important for the prediction of somatic coliphages. In fact, viruses were shown to be less sensitive to UV radiation than bacteria (Chang et al., 1985).

Some studies characterizing and modelling riverine systems used temperature as a predictive variable (Mas and Ahlfeld, 2007; Christensen, 2001; Vermeulen and Hofstra, 2014). Since temperature is strongly correlated with global solar irradiance but showing weaker correlations with fecal indicator concentrations, we concluded that global solar irradiance was a better predictor variable in regression models than temperature in this study. Therefore, we excluded temperature as an explanatory variable from all models.

Rising indicator concentrations occurring with increased turbidity can either originate from remobilized sediments (autochthonous) or from the surrounding land surface (allochthonous). Since there are only few agricultural areas, high indicator levels associated with high turbidity are likely to originate from resuspended sediments.

Turbidity alone and in combination with temperature has been used to predict bacteria rates in other studies (Rasmussen and Ziegler, 2003; Christensen et al., 2000). Its predictive power is also reflected in our study, since turbidity is a predictor in all models for the bacterial indicators, whereas for the prediction of somatic coliphages it is integrated in the extended mode model only, indicating that turbidity is –like global solar irradiance- less influential for predicting somatic coliphages compared to fecal indicator bacteria. In contrast, for coliphages discharge seems to be more influential.

Chlorophyll has been used as a predictor for *E. coli* concentrations in lakes by Nevers and Whitman (2005). The authors found that high concentrations of chlorophyll were associated with high *E. coli* counts, which could be explained with a relation to the release of nutrients along with fecal indicator bacteria, and the subsequent increase in primary producers. In this study, however, indicator concentrations were negatively correlated with chlorophyll *a*. As annual zooplankton peaks coincide with the phytoplankton blooms in rivers (Bergfeld et al., 2009), the observed effect may be explained by grazing processes.

In general, the temporal synchronization of parameters could improve the variables' predictive power within the models, as previously shown by Cyterski et al. (2012), which is consistent with the findings of other studies (Mas and Ahlfeld, 2007; Ferguson et al., 1996; Motamarri and Boccelli, 2012; Eleria and Vogel, 2005; Heberger et al., 2008). Thus, confirming that preceding environmental conditions affect fecal indicator concentrations and should be acknowledged in the model design accordingly.

4.3. Model comparisons

The majority of studies on predictive models assessing microbial water quality focused on lakes and beaches in the USA, whereas fewer studies targeted rivers. In the USA, models are increasingly implemented as management tools (Francy, 2009), while in Germany, modelling of fecal pollution is still in its infancy.

In brief, the predictive models of comparable studies to our survey resulted in an adjusted R^2 of 0.46–0.6 for Charles River (Eleria and Vogel, 2005) and R^2 of 0.59–0.79 for Kansas River and Little Arkansas River (Rasmussen and Ziegler, 2003). For Mystic River watershed an adjusted R^2 of 0.42–0.82 was obtained by Heberger et al. (2008).

In the present study, R^2 of the overall models ranged from 0.65 to 0.74 while site-specific R^2 varied between 0.54 and 0.82. Thus, the relationships developed in our models are comparable to other studies.

Threshold violations were predicted correctly by the models in 68 and 81%. It has been described that multiple linear regression approaches may show high false negative rates (Motamarri and Boccelli, 2012). However, this effect was not observed in our study.

It has been mentioned, that models were site-dependent (David and Haggard, 2011; Brady et al., 2009), but results here indicate, that a consideration of a larger scale is possible. This was also stated by Nevers and Whitman (2008) for beach sites along a 35 km stretch of Lake Michigan, USA.

Previous modelling approaches have addressed bacterial indicators only. Though, Skraber et al. (2002) investigated relations between environmental variables and phages in comparison to bacterial indicators. The authors showed that phages were differently related to environmental parameters than bacteria. These results are supported by our study, which demonstrated that different explanatory parameters are required for the prediction of bacteria and phages. Hence, phages and bacteria should be considered separately when characterizing fecal indicator organism activity in the environment. However, Skraber et al. (2002) concluded that bacteriophages are less sensitive to environmental factors like discharge, which was not supported by our results.

Furthermore, we were able to show that routine sampling and monitoring efforts can be reduced up to half of the required variables with a loss of prediction accuracy of less than 4% of explained variance with regard to the overall model performance.

4.4. Application and future work

Our study showed that MLR models can serve as a valuable management tool to assess water quality and any deterioration in the lower Lahn area.

The models developed are easy to implement and results are provided promptly. All parameters, except NH_4–N can be readily obtained from monitoring stations or by rapid sensor measurements. However, NH_4–N measurements can also be performed quite quickly. Thus, results can be acquired within a few hours, resulting in a significant improvement in timeliness compared to culture-dependent methods. Nevertheless, to ensure reliable model performances the introduced models should be fitted with further data, as this may offer a significant improvement in model accuracy (Christensen et al., 2000).

Hence, future research should aim on the collection of prolonged time-series data to improve model performance and to minimize effects by atypical years.

The reliable application of MLR models as a tool for recreational water quality surveillance is supported by a solid model performance, which should be verified constantly and adjusted on the basis of random samples.

5. Conclusions

- Multilinear regression models developed in the present study provide a cost-effective tool for the assessment of water quality.
- The models can be implemented at a spatial scale encompassing multiple sites, thus reducing monitoring and modelling efforts.
- The timeliness of water quality assessment improved significantly using the modelling approach compared to culture-dependent methods.
- The results contribute to the understanding of the factors influencing bacterial and viral indicator concentrations along a river.

Acknowledgements

This study was conducted within the German research programme KLIWAS (Impacts of climate change on waterways and navigation – Searching for options of adaptation), financed by the Federal Ministry of Transport and Digital Infrastructure.

We further acknowledge Bianca Konrath and Michaela Theis for excellent assistance with sampling and laboratory analyses.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Proceedings of 2nd International Symposium on Information Theory, Budapest, Hungary, pp. 267–281.

Bergfeld, T., Scherwass, A., Ackermann, B., Arndt, H., Schöl, A., 2009. Comparison of the components of the planktonic food web in three large rivers (Rhine, Moselle and Saar). River Res. Appl. 25 (10), 1232–1250.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Methodol. 211–252.

Brady, A.M., Bushon, R.N., Plona, M.B., 2009. Predicting Recreational Water Quality Using Turbidity in the Cuyahoga River. Cuyahoga Valley National Park, Ohio, 2004-7, U.S. Geological Survey Scientific Investigations Report 2009–5192.

Burkhardt III, W., Calci, K.R., Watkins, W.D., Rippey, S.R., Chirtel, S.J., 2000. Inactivation of indicator microorganisms in estuarine waters. Water Res. 34 (8), 2207–2214.

Chang, J.C., Ossoff, S.F., Lobe, D.C., Dorfman, M.H., Dumais, C.M., Qualls, R.G., Johnson, J.D., 1985. UV inactivation of pathogenic and indicator microorganisms. Appl. Environ. Microbiol. 49 (6), 1361–1365.

Christensen, V.G., Jian, X., Ziegler, A.C., 2000. Regression Analysis and Real-time Water-quality Monitoring to Estimate Constituent Concentrations, Loads, and Yields in the Little Arkansas River, South-central Kansas, 1995–99, U.S. Geological Survey Water Resources Investigations Reports, 00–4126, 36 pp.

Christensen, V.G., 2001. Characterization of Surface-water Quality Based on Real-time Monitoring and Regression Analysis. Quivira National Wildlife Refuge, South-central Kansas. December 1998 through June 2001. United States Geological Survey Water-resources Investigations 01–4248:28 pp.

Christensen, V.G., Rasmussen, P.P., Ziegler, A.C., 2002. Real-time water quality monitoring and regression analysis to estimate nutrient and bacteria concentrations in Kansas streams. Water Sci. Technol. 45 (9), 205–219.

Craggs, R.J., Zwart, A., Nagels, J.W., Davies-Colley, R.J., 2004. Modelling sunlight disinfection in a high rate pond. Ecol. Eng. 22 (2), 113–122.

Crowther, J., Kay, D., Wyer, M.D., 2001. Relationships between microbial water quality and environmental conditions in coastal recreational waters: the fylde coast, UK. Water Res. 35 (17), 4029–4038.

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.

David, M., Haggard, B., 2011. Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois river watershed, Arkansas and Oklahoma, USA. Water, Air, & Soil Pollut. 215 (1–4), 525–547.

Eleria, A., Vogel, R.M., 2005. Predicting fecal coliform bacteria levels in the Charles

River, Massachusetts, USA. J. Am. Water Resour. Assoc. 41 (5), 1195–1209.
Erichsen, A.C., Burgdorf Nielsen, J., Dahl-Madsen, K.I., 2003. Copenhagen's bathing anniversary. Water 21, 42.
EU, 2006. Directive 2006/7/EC of the European Parliament and of the Council of 15 Febuary 2006 concerning the management of bathing water quality and repealing directive 76/160. Off. J. Eur. Union L 64 (L 64), 37–51.
Ferguson, C.M., Coote, B.G., Ashbolt, N.J., Stevenson, I.M., 1996. Relationships between indicators, pathogens and water quality in an estuarine system. Water Res. 30 (9), 2045–2054.
Francy, D.S., Helsel, D.R., Nally, R.A., 2000. Occurrence and distribution of microbiological indicators in groundwater and stream water. Water Environ. Res. 152–161.
Francy, D.S., Darner, R.A., Bertke, E.E., 2006. Models for Predicting Recreational Water Quality at Lake Erie Beaches. United States Geological Survey Scientific Investigations Report 2006–5192.
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.
Grabow, W., 2001. Bacteriophages: update on application as models for viruses in water. Water SA 27 (2), 251–268.
Hampson, D., Crowther, J., Bateman, I., Kay, D., Posen, P., Stapleton, C., Wyer, M., Fezzi, C., Jones, P., Tzanopoulos, J., 2010. Predicting microbial pollution concentrations in UK rivers in response to land use change. Water Res. 44 (16), 4748–4759.
Heberger, M.G., 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. Plan. Manag. 134 (1), 73–82.
ISO 9308-3, 1998. 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). International Organization for Standardization, Geneva, Switzerland.
ISO 7899-2, 2000a. Water Quality - Detection and Enumeration of intestinal enterococci – Part 2: Membrane Filtration Method (ISO 7899-2:2000). International Organization for Standardization, Geneva, Switzerland.
ISO 10705-2, 2000b. Water Quality – Detection and Enumeration of Bacteriophages – Part 2: Enumeration of somatic coliphages (ISO 10705-2:2000). International Organization for Standardization, Geneva, Switzerland.
ISO 19458, 2006. Water Quality – Sampling for Microbiological Analysis (ISO 19458:2006). International Organization for Standardization, Geneva, Switzerland.
Lumley, T., 2009. Using Fortran Code by Alan Miller - Leaps: Regression Subset Selection. R package version. 2.9. http://CRAN.R-project.org/package=leaps.
Mas, D.M.L., Ahlfeld, D.P., 2007. Comparing artificial neural networks and regression models for predicting faecal coliform concentrations. Hydrol. Sci. J. 52 (4), 713–731.
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.
Nevers, M.B., Whitman, R.L., 2005. Nowcast modeling of Escherichia coli concentrations at multiple urban beaches of southern Lake Michigan. Water Res. 39 (20), 5250–5260.
Nevers, M.B., Whitman, R.L., 2008. Coastal strategies to predict Escherichia coli concentrations for beaches along a 35 km stretch of southern Lake Michigan. Environ. Sci. Technol. 42 (12), 4454–4460.
Olyphant, G., Whitman, R., 2004. Elements of a predictive model for determining Beach closures on a real time basis: the case of 63rd street Beach Chicago. Environ. Monit. Assess. 98 (1–3), 175–190.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., 2007. NLME: Linear and Nonlinear Mixed Effects Models, p. 57. R. package version 3.
R Core Team, 2014. R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Version 3.1.0. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
Ramette, A., 2007. Multivariate analyses in microbial ecology. FEMS Microbiol. Ecol. 62 (2), 142–160.
Rasmussen, P.P., Ziegler, A.C., 2003. Comparison and Continuous Estimates of Fecal Coliform and *Escherichia coli* Bacteria in Selected Kansas Streams. May 1999 through April 2002. U.S. Geological Survey Water-Resources Investigations Report 03–4056.
Schoonover, J.E., Lockaby, B.G., 2006. Land cover impacts on stream nutrients and fecal coliform in the lower Piedmont of West Georgia. J. Hydrol. 331 (3), 371–382.
Sinton, L.W., Hall, C.H., Lynch, P.A., Davies-Colley, R.J., 2002. Sunlight inactivation of fecal indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and saline waters. Appl. Environ. Microbiol. 68 (3), 1122–1131.
Skraber, S., Gantzer, C., Maul, A., Schwartzbrod, L., 2002. Fates of bacteriophages and bacterial indicators in the Moselle river (France). Water Res. 36 (14), 3629–3637.
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.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York. ISBN 0-387-95457-0.
Vermeulen, L.C., Hofstra, N., 2014. Influence of climate variables on the concentration of *Escherichia coli* in the Rhine, Meuse, and Drentse Aa during 1985–2010. Reg. Environ. Change 14 (1), 307–319.
WHO, 2003. Guidelines for Safe Recreational Water Environments: Coastal and Fresh Waters, vol. 1. World Health Organization, Geneva, Switzerland.