Science of the Total Environment 630 (2018) 20–31

Contents lists available at ScienceDirect

Science of the Total Environment

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

Developing, cross-validating and applying regression models to predict the concentrations of faecal indicator organisms in coastal waters under different environmental scenarios

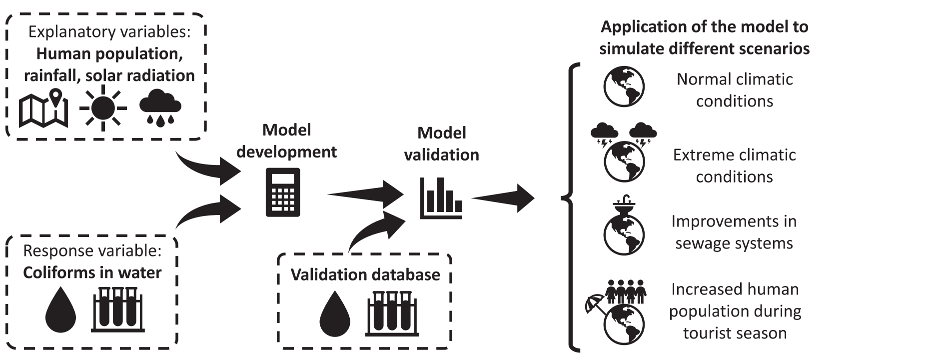
Robson V. de Souza [a,*], Carlos J.A. Campos [b], Luis H.P. Garbossa [a], Walter Q. Seiffert [c]

[a] Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (Epagri), Rodovia Admar Gonzaga, 1347, Itacorubi, Florianópolis, SC 88034-901, Brazil
[b] Centre for Environment, Fisheries & Aquaculture Science (Cefas), Weymouth Laboratory, Barrack Road, The Nothe DT4 8UB, UK
[c] Universidade Federal de Santa Catarina (UFSC), Rodovia Admar Gonzaga, 1346, Itacorubi, Florianópolis, SC 88034-001, Brazil

HIGHLIGHTS

- Enhanced models predicted coliforms in coastal waters based on environmental data.
- Cross-validation indicates adequate characterization of coliform variability.
- High rainfall and low solar radiation increased coliforms by 5 $\log_{10}$ MPN.100 mL^{-1}.
- In the summer, coliform die-off offsets higher contamination due to tourism.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 October 2017
Received in revised form 9 February 2018
Accepted 12 February 2018
Available online 19 February 2018

Editor: H.M. Solo-Gabriele

Keywords:
Coliform
Bathing water
Shellfish aquaculture
Rainfall
Solar radiation

ABSTRACT

This study developed, cross-validated and applied a regression-based model to predict concentrations of faecal indicator organisms (FIOs) under different environmental conditions in the North and South bays of Santa Catarina, South of Brazil. The model was developed using a database of FIO concentrations in seawater sampled at 50 sites and the validation was performed using a different database by comparing 288 pairs of measured and modelled results for 15 sites. The index of agreement between the model outputs and the FIO concentrations measured during the validation period was 66%; the mean average error was 0.43 $\log_{10}$ and the root mean square error was 0.58 $\log_{10}$ MPN.100 mL^{-1}. These validation results indicate that the model provides a fair representation of the FIO contamination in the bays for the meteorological conditions under which the model was trained. The simulation of different scenarios showed that under typical levels of resident human population in the catchments and median rainfall and solar radiation conditions, the median FIO concentration in the bays is 0.4 MPN.100 mL^{-1}. Under extreme meteorological conditions, the combined effect of high rainfall and low solar radiation increased FIO concentrations up to 5 $\log_{10}$ MPN.100 mL^{-1}. The simulated scenarios also show that increases in resident population during the summer tourist season and average rainfall concentrations do not increase median FIO concentrations in the bays relative to periods of time with average population, possibly because of higher bacterial die-off in the waters. The models can be an effective tool for management of human health risks in bathing and shellfish waters impacted by sewage pollution.

* Corresponding author.
E-mail address: robsonsouza@epagri.sc.gov.br (R.V. de Souza).

https://doi.org/10.1016/j.scitotenv.2018.02.139
0048-9697/

1. Introduction

Recreational use and consumption of raw or lightly cooked bivalve shellfish harvested from waters contaminated with sewage pollution can pose significant human health risks (Prüss et al., 2002). To reduce these risks, public health officials and water resource managers are increasingly using mathematical models to predict water quality and communicate to the public when pathogens in hazardous concentrations are likely to be present in the waters. Fundamentally, two different types of models have been used to assess the safety levels for swimming/shellfish harvesting: regression-based models and mechanistic models (de Brauwere et al., 2014). Regression-based models link a set of input (explanatory) variables and an output (for example, concentrations of faecal indicator organisms; FIOs) based on methods of regression analysis. The explanatory variables commonly used to predict FIO concentrations include land use or land cover characteristics of catchment(s), resident human or animal populations, and meteorological and hydrological data (Campos et al., 2013; de Brauwere et al., 2014). Regression-based models rely on relatively basic statistical concepts and are easy to implement (de Brauwere et al., 2014) while mechanistic models require more resources (e.g. meteorological model outputs, water quality database, coastal geometry and bathymetry) and a detailed understanding of the physical and biological factors governing bacterial dispersion and decay (Ge and Frick, 2007).

Models can be used to predict the FIO concentrations in coastal areas under different environmental conditions (scenarios). To be used as predictive tools, they should have their predictive power assessed (EPA, 2016; Francy and Darner, 2006). Assessing the predictive capability of a model is different from evaluating its goodness of fit in that the predictive capability relates to how well the model can be used for future observations (Ge and Frick, 2007). The statistics used for testing the goodness of fit, such as the R^2 and the model standard error, are ineffective for this purpose (Willmott, 1982). It is also important to assess how the model predictions generalise to an independent data set (cross-validation). Studies that include cross-validation of regression-based models for prediction of FIO concentrations in coastal waters are rare (Crowther et al., 2011; Ge and Frick, 2007). More studies of this nature are needed to overcome the difficulties associated with delayed notifications following testing of FIOs in water/shellfish using conventional culture-based methods and support human health risk estimates for shellfish and bathing waters.

The approach reported in this paper addresses this knowledge gap by developing, cross-validating and applying regression models to predict FIO concentrations in coastal waters under different environmental scenarios. The models were developed for the North and South Bays of Santa Catarina, which include the largest number of shellfish farming sites and some of the highest profile bathing waters in Brazil. The adjacent catchments experience large variations in human population due to tourism and therefore are the ideal sentinel site for testing the effectiveness of these models. Only ~39% of the catchment population is served by centralised sewerage collection and treatment systems (SNSA/MCIDADES, 2014) and a recent study showed that the current systems in place are not effective in reducing the FIO levels reaching the marine environment (Garbossa et al., 2017). Therefore, prediction of FIO concentrations in coastal areas can assist state authorities in managing health risks associated with bathing and shellfish consumption. This is highlighted by reports of human illness from bathing following sewage spills in parts of the bays (Globo Notícias, 2016).

2. Material and methods

2.1. Site description

A detailed description of the study site is given by Garbossa et al. (2014a). Briefly, the North and South Bays (48°33′57″W; 27°35′46″S – Datum Sirgas 2000) are two adjacent bodies of water (total area = 430 km^2) located between the mainland and Santa Catarina Island (Fig. 1). The bays connect with the Atlantic Ocean through the northern end of the North Bay (north mouth) and the southern end of the South Bay (south mouth) and are linked by a strait located midway on the western coast of the island. The bays are shallow (average depth = 3.4 m) and microtidal. The adjacent catchment (total area = 1875 km^2) is drained by several rivers and streams. The highest population density occurs around the strait between the North and South bays. The city of Florianópolis and metropolitan area has a combined population of over 800,000. Human population fluctuates considerably during the year because of water-based tourism and recreation. During the summer, this geographical area accommodates an extra 1.5 million tourists (SANTUR, 2012).

2.2. Model development

A database of concentrations of FIOs monitored monthly at 50 sampling points located in the North and South Bays of Santa Catarina (Fig. 1) from August 2012 to October 2013 was used to develop the predictive models. The period of 2012–2013 is hereafter referred to as the model training period. A detailed description of the sampling and microbiological analytical methods is given by Garbossa et al. (2014a). In summary, seawater samples were collected at a depth of 1.5 m and the sampling schedule was random with respect to the state of the tide and meteorological conditions. These samples were tested for faecal coliforms using the multiple tube (most probable number) method (ISO 9308-2: 1990).

Three models were developed: a model to predict the spatial variation of FIO concentrations in the bays (spatial model); a model to predict the temporal variations of FIO concentrations in the bays (temporal model); a model that integrates the explanatory variables of the spatial and temporal models to predict the FIO concentrations for the whole area of the bays under different environmental scenarios (integrated model).

2.2.1. Spatial model

To develop the spatial model, a geographic information system (GIS) was used to delineate the areas of the catchments and sub-catchments draining the North and South bays. The sub-catchment boundaries were integrated with a population census database (IBGE 2010, available on: http://censo2010.ibge.gov.br) to determine the human population living in each sub-catchment. The human population was used as the single predictor for FIO variations in the spatial model since Souza et al. (under review) showed that this is the most important factor associated with the variation of FIOs in this area. They also demonstrated that additional geographical parameters (i.e. total area of the catchment, urban area or percentage of impermeable cover) do not significantly increase the levels of explained variance of the models. A simple linear regression model was developed correlating the geometric mean (GM) (data from all monitoring campaigns combined) of FIO concentrations at each sampling point (response variable) with the combined human population in catchments whose outlets are located within 3.1 km from the sampling points (explanatory variable). The method for identifying the radiuses around the sampling points has been described by Souza et al. (under review). Basically, the method consists of developing a series of linear regression models considering as explanatory variables the geographical parameters of catchments with outlets located at varying distances from monitoring points and selection of the model with the highest explained variance. Six of the sampling points are located in central parts of the bays (W03, W05, W08, W11, W15, W20), away from the coastline. No catchment outlets were identified within 3.1 km from sampling points W05, W08, W11 and W15, thus these were not included in the analysis. Of the remaining sampling points, W03 had high leverage (Cook's distance = 0.7) and was considered an outlier. Therefore, it was decided to exclude all the points located in the central part of the bay from the model.