

A physical descriptive model for predicting bacteria level variation at a dynamic beach

Cheng He, Yvonne Post, John Dony, Tom Edge, Mahesh Patel and Quintin Rochfort

ABSTRACT

A rational-based physical descriptive model (PDM) has been developed to predict the levels of *Escherichia coli* in water at a beach with dynamic conditions in the Greater Toronto Area (GTA), Ontario, Canada. Bacteria loadings in the water were affected not only by multiple physical factors (precipitation, discharge, wind, etc.), but also by cumulative effects, intensity, duration and timing of storm events. These may not be linearly related to the observed variations in bacteria levels, and are unlikely to be properly represented by a widely used multiple linear regression model. In order to account for these complex relationships, the amounts of precipitation and nearby creek discharge, the impact of various time-related factors, lag time between events and sample collection, and threshold for different parameters were used in determining bacteria levels. This new comprehensive PDM approach improved the accuracy of the *E. coli* level predictions in the studied beach water compared to the previously developed statistical predictive and presently used geometric mean models. In spite of the complexity and dynamic conditions at the studied beach, the PDM achieved 75% accuracy overall for the five case years examined.

Key words | beach closing, *Escherichia coli*, forecasting

Cheng He (corresponding author)

Tom Edge

Quintin Rochfort

National Water Research Institute, Environment
Canada,

867 Lakeshore Road, Burlington, Ontario,
Canada L7R 4A6

E-mail: cheng.he@canada.ca

Yvonne Post

John Dony

Department of Environmental Engineering,
University of Guelph,
50 Stone Road East, Guelph, Ontario,
Canada N1G 2W1

Mahesh Patel

Toronto Public Health,
399 The West Mall, Toronto, Ontario,
Canada M9C 2Y2

INTRODUCTION

Beach closures due to fecal pollution are a global public health issue. In the Great Lakes, most recreational water is monitored for *Escherichia coli* as an indicator organism for fecal contamination. There are 11 recreational beaches in the Greater Toronto Area (GTA) located on the north shore of Lake Ontario. The City of Toronto collects daily water samples to monitor the *E. coli* level, and posts different flags accordingly at each beach during the swimming season. *E. coli* levels over 100 colony forming units (CFU)/100 mL in surface water collected at chest depth (wading in from the beach) result in a beach closure. However, because *E. coli* requires approximately 24 hours to incubate prior to reading, the *E. coli* levels of the beach water are only known the day after. Therefore, there is a great need to develop a reliable predictive model which could warn beachgoers of conditions hazardous to human

health. In recent years, daily beach water quality predictions have been developed to provide more guidance to beach management. Traditionally, beaches have been posted once the previous day's bacterial testing was completed and results analyzed, resulting in an element of risk to swimmers. Recently, computer models have been used to try to manage this risk. As summarized by Francy (2009) and assessed by Thoe *et al.* (2015), the models which have been used at recreational beaches include: rainfall-based alerts, deterministic models, and multivariable statistic and best fit-based models. In the rainfall-based model, the relationship between rainfall and fecal-indicator bacteria concentrations is based on historical data and can be established either qualitatively or statistically. This kind of model is relatively simple, and only used for a beach where the observed bacteria levels depend largely on the amount of

doi: 10.2166/wh.2016.206

rain (USEPA 1999; Ackerman & Weisberg 2003; Kuntz 2006). If multiple influence variables are involved, it is often termed a classification tree model.

Deterministic models use mathematical representations of the processes that affect bacteria concentrations. They calculate bacterial transport from discharges, sediment and other sources to the areas of concern (usually swimming beaches) using a two or three dimensional hydrodynamic model (Hydro Qual Inc. 1998). The quantity of bacteria available at each source needs to be known for these types of models. Depending on the size and complexity of the study area, model run times may still be a concern even with the power of modern computer systems, because it needs to simulate multi-dimensional flow conditions. There can still be a large uncertainty in hydrodynamic modelling results, as well as it being difficult to quantitatively simulate bacterial contributions from non-point sources along the shoreline due to wet weather impacts and wave action. Deterministic models are thus rarely used in practice as a daily predicting tool; however, they can be useful to assess the relative importance of processes influencing *E. coli* levels, and to understand the variation of beach bacteria levels resulting from point contamination sources and local flow hydraulic conditions (Liu *et al.* 2006; Thupaki *et al.* 2010).

In most cases, beach bacteria levels are controlled by multiple physical factors, so the majority of operational predictive models are developed based on statistics with multiple linear regression (MLR) or other regression methods with various names (Francy & Darner 2002; Eleria & Vogel 2005; Nevers & Whitman 2005; Olyphant 2005; Wymer *et al.* 2005; Francy *et al.* 2006; Mas & Ahlfield 2007; Thoe & Lee 2014). These models are usually based on readily available weather, environmental, and hydrodynamic data (Kay *et al.* 2005), and are generally found to outperform the traditional beach monitoring methods in issuing correct beach advisories. However, there are several assumptions associated with the MLR method. One of the most basic assumptions is that the response variable is linearly related to the explanatory variable, which is not likely to be satisfied in many cases. For instance, data analysis by Fuss & O'Neill Inc. (2010) showed that precipitation may be the most influential factor on beach water bacteria level variation, yet the measured bacteria level does not

necessarily have a linear relationship with the total amount of rain. This is because the rain duration, frequency, intensity and water sampling time (whether it is in the first flush period or later) are all important factors to determine the measured bacteria level. The linearity may be improved to a certain degree with various data transformation methods such as logarithmic transforms (Olyphant 2005). In addition, for the MLR or other similar methods, the step-wise and best subsets selection procedures often need to be carried out before MLR modeling (Helsel & Hirsch 1992), which can induce some artificial factors and inconsistencies.

Statistical models depend solely on the correlation between the input and affected parameters, regardless of real causes, thus their performance can vary greatly. In general they performed better for cleaner or less dynamic beaches, but had difficulty in predicting exceedance events because low bacteria conditions tend to over-influence the mathematical relationships (Thoe & Lee 2014; Thoe *et al.* 2015). These mathematical-based models can be susceptible to predominant factors. Nevers & Whitman (2011) recently reported that predictive models have to be tailored to a specific beach, and do not always result in reduced management error.

This study focused on the Marie Curtis Park East Beach (MCPEB). The water quality and bacteria levels in the beach water can be very dynamic, because a local creek (Etobicoke Creek) discharges into Lake Ontario through the middle of this 500 m long beach, and the creek drains a large populated urban area with extensive runoff from commercial, residential and highway zones. At the present time, a geometric mean, calculated from the previous 2 days' sampled data, is used to determine whether the beach should be posted open or closed by the City of Toronto. This method relies entirely on grab samples and does not consider any of the conditions at the beach for that day. Therefore, the results are usually very predictable, following the trend of the past 2 days regardless of present weather and other conditions. In addition, the geometric mean requires 2 days' worth of data, so if the water samples are missing for 1 day, *E. coli* results cannot be posted for the following 2 days. A predictive model based on an MLR method and optimized using 2007–2008 data was previously developed for the MCPEB by Fuss & O'Neill Inc. (2010) with limited success. The tested accuracy for this model with

2009 data was around 54%, only 4% better than the monitoring-based method used by the City of Toronto.

The objective of this study was to understand the key quantitative factors causing bacteria level variations in beach water at the MCPEB, and to develop a more reliable and accurate physical descriptive model (PDM) for daily beach status prediction and thus improve risk management. To the authors' best knowledge, no model was available for predicting bacteria level variation in beach water that considered the impact of amount, intensity and duration of the events and sampling lag time. These combined effects are site dependent, nonlinear, complex and cannot be adequately represented by linear regression or statistical-based models; their relationships to *E. coli* levels have to be examined individually.

MATERIALS AND METHODS

Initial data analysis

To reduce the uncertainty and for accurate representation, *E. coli* data were collected daily from surface water by wading into chest deep water at five locations (Figure 1) in the morning between 9 a.m. and 12 p.m. The arithmetic means of *E. coli* levels in beach water were given 1 day later. Hourly precipitation data were collected from the

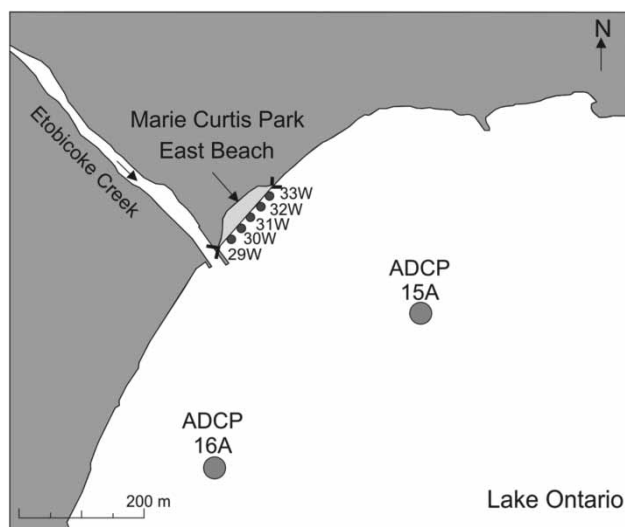


Figure 1 | Locations of water sampling and instrument deployment in the study area.

Environment Canada Toronto City auto station (<http://climate.weather.gc.ca/>) about 15 km away from the study site. By comparing the precipitation and *E. coli* variations (using 2012 data as an example) in Figure 2, it can be seen that, in general, *E. coli* values tend to peak above the provincial standard of 100 CFU/100 mL around the same time that precipitation events occur. This implies, as expected, that at the MCPEB, precipitation will be one of the important influential factors in controlling the bacteria level in beach water. However, it was found that: (1) bacteria levels were not always proportional to the amount of precipitation (the relationship was obviously nonlinear and could be very complex); and (2) the correlation coefficients calculated from daily bacteria level and precipitation measured at 10 a.m. (or daily average precipitation using hourly measurements) were not high. Thus, they were not representative enough to describe their close relationship due to delay, cumulative and continuous effects that precipitation can have on *E. coli* levels as well as a sampling time which was not fixed, resulted in some misaligned peaks. The relationship between precipitation and *E. coli* levels appears to be closely correlated, but a qualitative relationship could not be easily verified. Similar relationships were also observed among the other physical parameters with *E. coli* level variations. The popular multi-statistic or regression-based models may not be the most appropriate solution under these circumstances, as explained earlier, and a different approach needs to be explored.

Model development

It would be difficult to develop a simple and reliable explicit relationship between the variation in bacteria levels and the amount of precipitation and creek discharge, various time-related effects and other factors due to the complexity of their relationships. In this study, the combination of visual examination, correlation coefficient assessment and statistical methods were adopted to describe the relationships. Although the visual method may sound unsophisticated, it can provide an initial insight into a complex relationship and can be effectively used by a wide range of scientific studies (intentionally or otherwise). In this study the quantitative relationships between the controlling factors and bacteria level variations were established based on logical

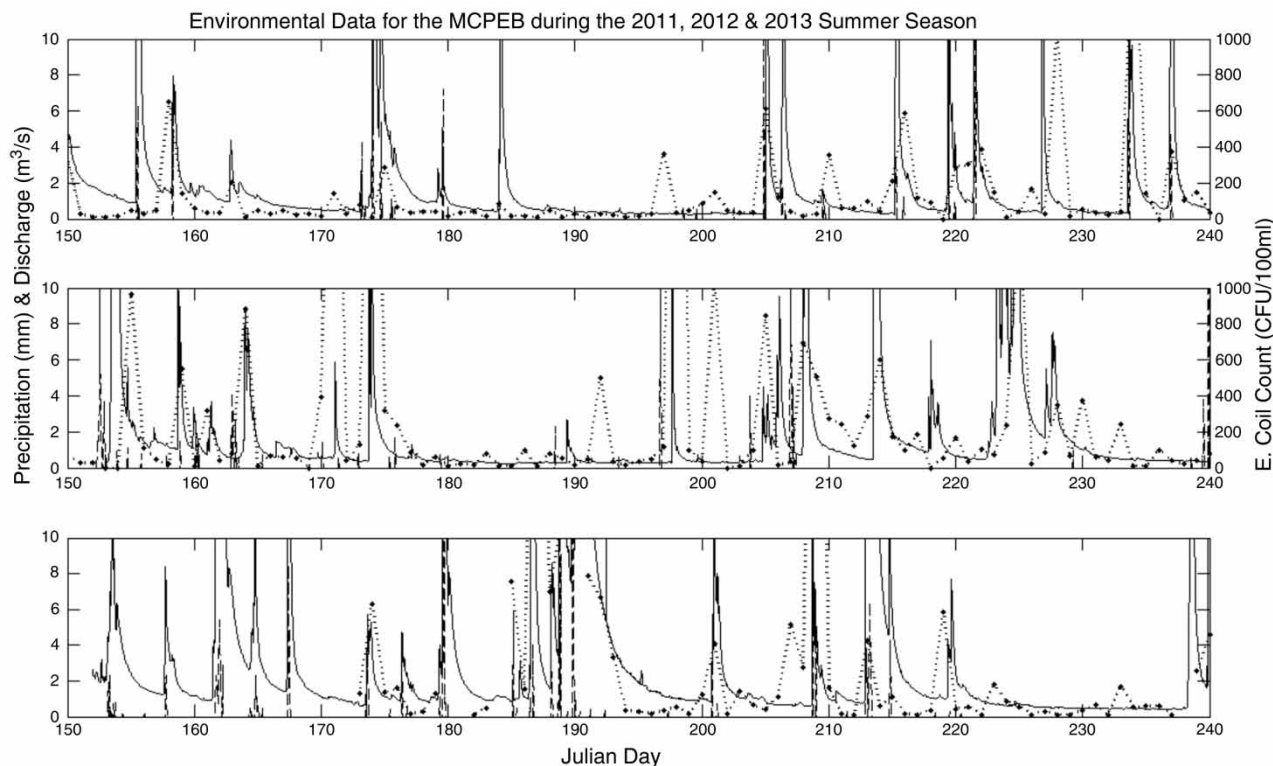


Figure 2 | Comparison of the measured hourly precipitation, 15-minute creek discharge and daily *E. coli* counts for 2011 (top panel), 2012 (middle panel) and 2013 (bottom panel). Dotted curve, precipitation; dashed curve, creek discharge; solid curve, *E. coli* counts.

linkages, rather than on purely mathematical numbers. Each controlling factor was examined individually and sequentially instead of together as a group, as often occurs with regression models.

After visually examining the amplitude and peak positions for all measured data, it was concluded that the time-related effects (such as start and end time of each storm event), as well as the time window length during which the cumulative or average amount of an event was assessed, played a major role in the overall influence on bacteria level variations and would be the main focus in the model development. After quantitative examination of cumulative precipitation amounts occurring within different lengths of time windows for the years 2011–2013, it was found that events >5 mm of cumulative precipitation within 12 hours prior to sampling time had the greatest influence (with the highest associated accuracy of prediction) on the bacteria level variation in the water. Therefore, the threshold of 5 mm cumulative precipitation (within 12 hours) was selected as a key parameter in the determination of bacteria levels in the developed model. Choosing

appropriate time windows to process these kind of time series data with delayed and continuous effects was critical, and often missed in other predictive models. When precipitation events occurred long before the sampling time, they had much weaker influences on the beach water bacteria variations, but if the examination window was too short, some of the important effects from earlier events could be missed. From 2011 to 2013, 65%, 75% and 100% (respectively) of the precipitation events with a 12-hour cumulative precipitation over 5 mm had *E. coli* levels over 100 CFU/100 mL. As expected for a dynamic beach, there were also a few elevated *E. coli* days with <5 mm cumulative precipitation within the 12-hour window. Due to many other controlling factors and complex relationships between precipitation and bacteria level variations, using the 5 mm criteria alone would obviously not be suitable for all situations.

The Etobicoke Creek discharge records for 2011–2013 (15-minute intervals) were downloaded from the Water Office of Environment Canada's website (<http://www.wateroffice.ec.gc.ca/>), and the 2012 data were plotted in Figure 2 as an

example. The MCPEB daily bacteria levels mirrored the creek discharge peaks in most instances. In order to identify the discharge events, the background discharge level (baseflow) needed to be determined as a threshold. For all discharges on the days without precipitation, over the 3 years, an average value of around $1.5 \text{ m}^3/\text{s}$ was calculated. Using a similar method to the precipitation data analysis, a daily average discharge over a 24-hour period (after testing a range of time windows) was found to have the strongest link to predict changes in bacteria levels. If a daily average discharge was above $1.5 \text{ m}^3/\text{s}$, it was most likely (and had the fewest incorrect predictions) for *E. coli* levels to be above 100 CFU/100 ml. The need to consider a longer window for the creek discharge was mainly due to the fact that in general, the amount of water discharged from the creek was much larger than that flushed from the surrounding beach itself. The creek water quality could also be much worse, because the runoff could come from polluted sources including commercial and highway areas. It is important to note, therefore, that the creek discharge could have a longer-reaching influence on the water quality at the beach site.

This discharge information alone (on the days with 12-h cumulative precipitation $<5 \text{ mm}$), could correctly predict the bacteria level variations 50%, 64% and 83% of the time in 2011, 2012 and 2013, respectively. The correct prediction percentage for each year was lower than when using 12-h cumulative precipitation $>5 \text{ mm}$ as an influence indicator, because large enough precipitation always induces the creek discharge. This indicates that precipitation should be considered before the creek discharge in determining bacteria level variations. It should be pointed out that: (1) the above-mentioned criterion only served as an initial step in processing the influence from the creek discharge on the bacteria level variation, the more complex relationships will be further developed in the results section; (2) not only the volume, but also the quality of the creek discharge will have a large influence on beach water bacteria level. Due to the difficulty in monitoring bacteria information from the creek discharge continuously, the developed PDM will therefore miss an important input parameter, and as a result it will bring some uncertainties into the prediction results.

Turbidity was measured daily by the City of Toronto while collecting water samples to test for *E. coli*. It can be a very useful parameter for indicating elevated bacteria

possibly caused by sediment re-suspension (associated with strong winds and high waves), or from strong creek discharges. The comparisons between the available turbidity data and the *E. coli* levels measured from 1 June to 31 August in 2011 to 2013 were plotted in Figure 3. The results showed moderate influences from measured turbidity on associated bacteria levels. To quantitatively define their relationship, the various turbidity threshold values were examined and compared with bacterial loading data from 2011 to 2013. Testing of various scenarios showed that using turbidity information only (on the days with precipitation $<5 \text{ mm}$ and discharge $<1.5 \text{ m}^3/\text{s}$), and a threshold of $>6 \text{ NTU}$ (nephelometric turbidity units) could give the most accurate prediction for observed bacteria level variations. The 3-year average correct prediction rate (within these restrictions) was around 92%. However, there were relatively few days affected only by turbidity, so its overall influence in the predictive model is likely to be small.

Hourly wind data measured from the City Centre Airport on Toronto Island (about 15 km away) were downloaded (http://climate.weather.gc.ca/climateData/dailydata_e.html) and examined visually with the *E. coli* variations. It was found that only when the discharge from the creek was fairly small, the wind speed and direction together could push the plume onto the beach and thus play a minor role in affecting *E. coli* levels. When the discharge events were larger, it was more likely that there would be some portion of the creek plume traveling towards the beach because of the location of the creek discharge. Data analysis showed that in 2011 an elevated *E. coli* level occurred over 3 days following a small discharge event. Furthermore, in 2012, elevated *E. coli* levels were present for 5 days after small discharge events; however in 2013, elevated *E. coli* levels were recorded for only 1 day following such an event. In all of these cases, there was similar wind behavior, as the wind direction was blowing along the shore in the northwards direction (toward the beach). However, the relationships described above did not occur for every case, and a few exceptions were found in the 2011 and 2012 data. It is therefore necessary to consider the influences of wind, together with other factors such as turbidity, in order to best describe the bacteria level changes in the developed PDM. The actual implementation will be discussed in the results section in detail.

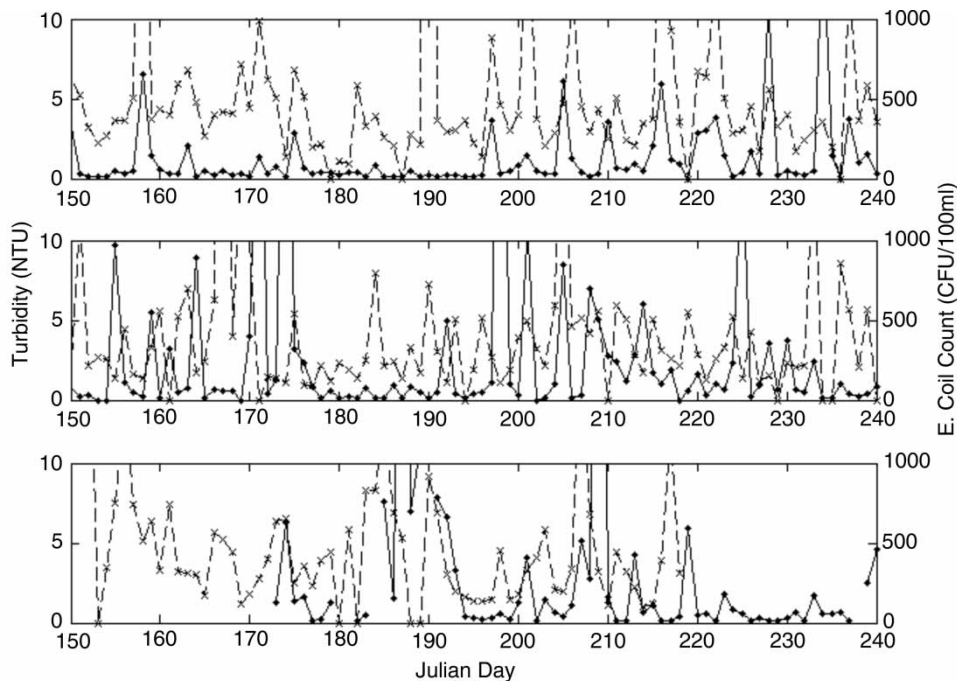


Figure 3 | Comparison of the measured daily turbidity and *E. coli* counts for 2011 (top), 2012 (center) and 2013 (bottom). Dashed curve, daily turbidity values; solid curve, daily bacteria counts.

The effects of wave action on *E. coli* levels were also examined because large waves could have enough power to re-suspend bottom sediments and increase the flux from beach sources, resulting in elevated *E. coli* concentrations in waters near shore (Beverdorf *et al.* 2007; Gao *et al.* 2011). The hourly average wave height was measured by a pressure sensor on the bottom-mounted acoustic Doppler current profiler (ADCP) 15A located in front of the beach in 2013 (see Figure 1). The comparisons between the measured wave height, bacteria level variations and turbidity were plotted in Figure 4.

It can be seen on Julian day 203 and 208 that the wave height and *E. coli* level were both very high, but since there were no indications from the other measured factors, this may imply that the elevated *E. coli* level was due to wave effects alone. However, on Julian days 183, 195 and 204, there were no signs of elevated *E. coli* levels, even though the waves were strong. This indicates that the wave height was only loosely correlated to bacteria variation at the MCPEB in 2013. There were no wave height measurements around the MCPEB area available for 2011 and 2012, so wave height data measured by Environment Canada at

two offshore moored buoy stations were used to compare with the measured bacteria data, similar to those shown for 2013. The influence of wave height on the observed bacteria level variations was inconclusive, considering that: (1) local wave information is not always available; (2) publicly accessible data measured in an open water area may not always be very representative of specific local conditions (as determined from comparisons of waves measured from near and off shore locations in 2013), possibly because wind conditions and the fetch lengths could be very different; (3) in general, the wave data do not present a consistent relationship with bacteria levels; and (4) the influence of wave height on bacteria levels could be reflected in the model through associated factors such as elevated turbidity. For these reasons, it was decided that wave height was not a critical component of the PDM.

Waterfowl populations at the beach were counted when the beach water samples were collected by the City of Toronto. There was no correlation between the waterfowl counts and bacteria levels, possibly due to the fact that the one-time snapshot waterfowl counts did not accurately represent the cumulative waterfowl population at the beach for

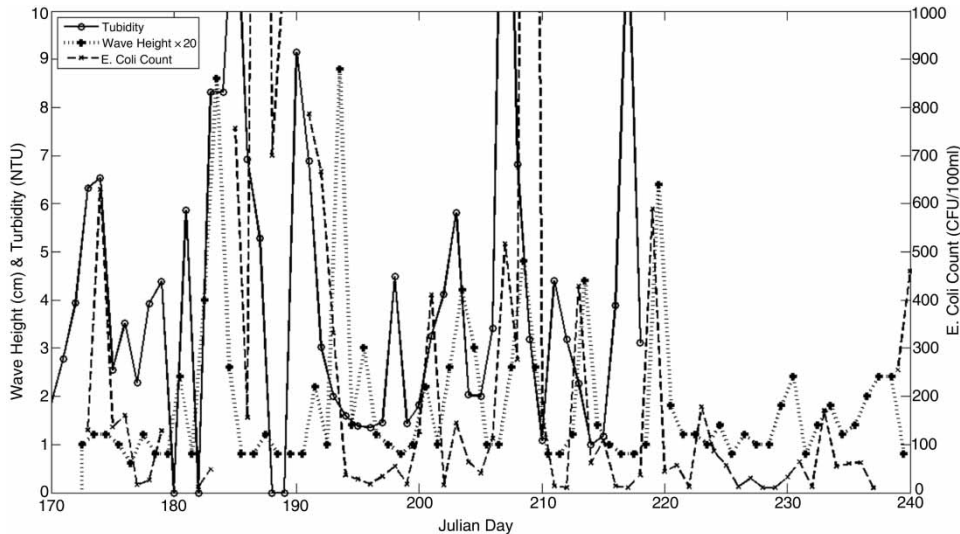


Figure 4 | Comparison of the measured daily turbidity, wave height and daily *E. coli* counts for 2013.

the whole day and was therefore not used in the developed PDM. The City of Toronto also measured air and surface water temperature at the same time as sampling the beach area, and compared them to the daily measured *E. coli* values. Neither air nor water temperatures were found to have a measureable effect on bacteria levels at the beach for all 3 years. Water level data were downloaded from the Fisheries and Oceans Canada website (<http://www.medsdmm.dfo-mpo.gc.ca/isdm-gdsi/twl-mne/index-eng.htm>), and were filtered using a 2-hour time filter in MATLAB (Mathematical Statistical and Computing software by MathWorks, Natick, MA) to reduce the influences of high frequency waves. This was compared with the average measured *E. coli* data, but no correlation was found and therefore it was concluded that the water level has little to no effect on bacteria levels at the MCPEB.

To investigate the impact of lake currents on the beach water *E. coli* level variation, the currents were measured in front of the beach at a water depth of 5 m with the two bottom-mounted ADCPs in 2013 at the locations indicated in Figure 1. Both of the ADCPs were set up with 0.5 m vertical resolution and 20-minute recording interval. The surface flow velocities measured by ADCP 15A were much stronger than the flow velocities near the lake bed, as expected, because the near shore current was mainly driven by wind. For 57% and 51% of the time, respectively,

the surface and bottom currents of the two horizontal velocity components were not in the same direction. The horizontal flow uniformity was also examined by comparing the flows measured from two ADCP locations. The differences in the flow velocities at these two measurement locations with the similar depths were very notable even though they were only about 500 m from each other. The characteristics of the measured flow velocities indicated that the lake currents in front of the MCPEB were spatially complex and non-uniform in both the horizontal and vertical directions. Because the creek mouth is adjacent to the MCPEB, the discharge plumes are likely to reach the beach in most instances, even under weak currents and wind pushing the surface current away from the beach, while the currents at depth are moving in a different direction. This may explain why the water quality at the MCPEB is very dynamic and difficult to predict.

In summary, the investigation of various influencing parameters showed that precipitation seemed to play the most important role in determining whether the *E. coli* levels at MCPEB would be above or below the provincial water quality standard on any given day. The event-based flow from Etobicoke Creek was also an important parameter in determining beach *E. coli* levels. Initial data analyses would suggest that using available field data, processed within properly defined timescales and selecting appropriate

threshold values, such key controlling factors could be used to predict the bacteria level variations. The influences of various factors on beach water bacteria levels could be defined by several nonlinear discrete numbers. The detailed relationships between controlling factors and bacterial level variation will be further explored in the results section. Wind speed and direction, as well as turbidity (where measurements are available), were also included in the PDM as they both proved to have some influence under certain circumstances. The influences on beach bacteria variations from wave height, temperature and other physical factors have been found to be inconclusive or minimal, and therefore they were not included as input parameters of the developed PDM.

RESULTS

Optimizing model parameters

The more complex and detailed relationships serving to improve the model predictions are investigated and described in this section. Most of the physical controlling factors were related (not independent), and each of them had a different degree of influence on bacteria level variations. The strategy for developing a new model in this study was to examine all of the controlling factors sequentially, beginning with the most influential factor. In this way, the duplicated influences from multiple controlling factors with the same cause (such as precipitation, creek discharge and turbidity) could be avoided. This reduced the chance that a particular influence factor would be weighted too heavily as well as allowing each factor to be accommodated differently. This method follows the path of logical reasoning rather than relying on weak correlation or best fit criteria.

As previously discussed, 24-hour average discharge values ($>1.5 \text{ m}^3/\text{s}$) prior to 10 a.m. had the greatest overall influence on the bacteria levels, at the site, and were used as the main assessment threshold in the PDM. By averaging the discharge values over a 24-hour period, a potential problem was that some of the smaller discharge events might get overlooked. In order to account for these recent smaller discharge events (which could still have significant impacts

on water quality), the discharge values in different time length windows prior to 10 a.m. were re-examined. It was found that the criterion of a 5-hour average where the calculated discharge was $>1.5 \text{ m}^3/\text{s}$ could capture the most additional elevated *E. coli* levels on the beach; otherwise they would be indicated as low bacteria levels because the 24-hour average discharge was $<1.5 \text{ m}^3/\text{s}$.

An extremely large precipitation event in 2013 drew attention to the fact that the creek flow during the prior 24 hours could also be important to beach water quality. It was observed that after a certain period of high discharge, *E. coli* levels began to drop at the beach site even though the discharge remained above the threshold of $1.5 \text{ m}^3/\text{s}$. This beach water quality improvement was attributed to the reduction in contaminants and the flushing effect from cleaner water at the end of a large discharge event, where the initial flows could contain high levels of contaminants washing off the impervious surfaces in the watershed.

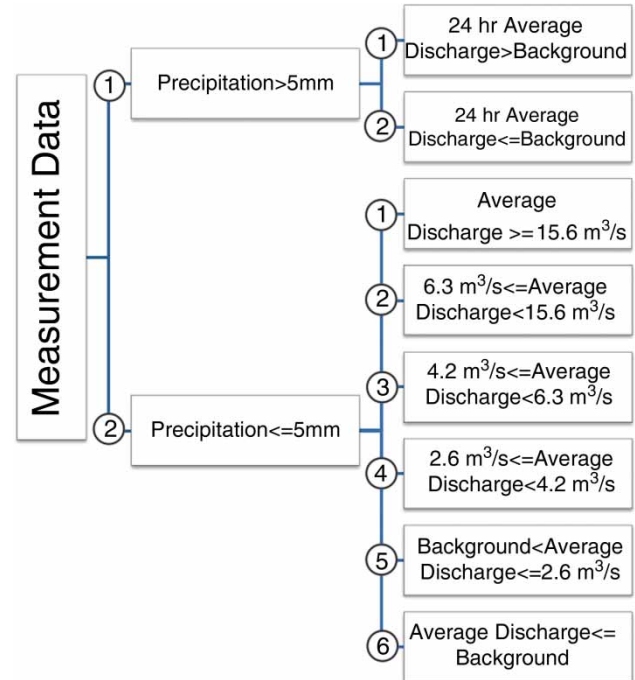
The lag time between the end of a discharge event and the sampling time was also found to be an important factor in determining whether or not the discharge would have an effect on *E. coli* levels, and was implemented in the PDM. Obviously, the lag time should be a function of the creek discharge rates. Four pre-assigned lag times were used in the PDM, corresponding to different 24-hour average discharge value ranges. Each lag time was determined (using data from all 3 years) by gradually reducing tested lag time, starting from 20 hours until the highest prediction accuracy for the bacteria levels was achieved for a specified daily average discharge range. The optimized lag times are summarized in Table 1. It can be seen that lag time increases with larger average flow rates, because it has longer effect time. It was concluded that a lag time did not need to be calculated when the average discharge was below the maximum background discharge of $1.5 \text{ m}^3/\text{s}$ or larger than $15.6 \text{ m}^3/\text{s}$, because if the discharge was too small it only had a very short residual effect, and if the discharge was very large it was expected to have an effect on the *E. coli* levels regardless of what time the discharge event ended in the 24-hour window. In the PDM, if the detected lag time was longer than the pre-assigned lag time under a particular discharge rate, the discharge was expected to have no effect on bacteria levels.

Table 1 | Selected lag times based on 24-hour average discharge rates

24-hour average discharge rate (m ³ /s)	Corresponding lag times (h)
0–background value 1.5 m ³ /s	N/A
Background–2.6 m ³ /s	3
2.6–4.2 m ³ /s	7
4.2–6.3 m ³ /s	12
6.3–15.6 m ³ /s	15
Over 15.6 m ³ /s	N/A

The influence of turbidity levels was addressed only after precipitation and creek discharge information was processed for the days with elevated *E. coli*. If turbidity measurement was equal to or >6.00 NTU, the beach would be predicted to be closed. The wind information was the last parameter to be included in the PDM. Alone, it could not provide enough input to determine *E. coli* levels in the beach water, possibly due to its weak impacts on bacteria level variations and the complexity of the flow patterns generated. The turbidity information thus needed to be considered together with the wind to determine the extent of its influence.

In this study, a PDM was developed in MATLAB (other software packages/tools could be used with equal success) to predict bacteria level variations based on the basic relationships of the various controlling parameters outlined above. Figure 5 shows the main flow chart of the model. Each large branch has been broken into smaller branches for easier viewing. Figure 6 (branches 1–1 and 1–2) shows the decision tree when the sum of the precipitation over the 12-hour period prior to 10 a.m. was above 5 mm. The first component checks if the average discharge was larger than the background discharge to confirm whether the precipitation event also occurred near the MCPEB or the creek upstream. This is important, because the hourly precipitation data were collected at the City of Toronto (around 15 km from the study beach), and the event could be localized (i.e. not affect the creek discharge). If this condition holds true, the *E. coli* levels were expected to be high and the beach would have to be closed. If the average discharge was found to be less than the background discharge, the average discharge in the 5 hours prior to 10 a.m. was considered to avoid overlooking smaller events that may have

**Figure 5** | Main flow chart of the newly developed physical descriptive forecasting model.

occurred closer to the time when the water samples were collected. If the average discharge over the 5-hour period was considered to be less than the background value, then the ratio of the 12-hour and 24-hour average flow was used to determine if there was an undetected small discharge occurring during a dry period in which the background discharge was much lower than the pre-set average minimum background of 1.5 m³/s. If the case condition was met, the wind was assessed because it could push the discharge plume toward the beach, causing elevated bacteria levels in the beach water. If no signs of a small increased discharge were evident, turbidity measurements (if available) were taken into consideration, to factor in days that *E. coli* levels were elevated, caused by increased turbidity.

When the sum of the 12-hour precipitation was <5 mm, the program would flow into one of six discharge branches, depending on discharge rates as shown in Figure 5. To illustrate the program logic of this model, the most complex sub-branch (2–5) is shown in Figure 7 and discussed in detail. The program flows into sub-branch 2–5 when the 24-hour average discharge is <2.6 m³/s. Lag time is checked first in the branch; if it is shorter than 3 hours,

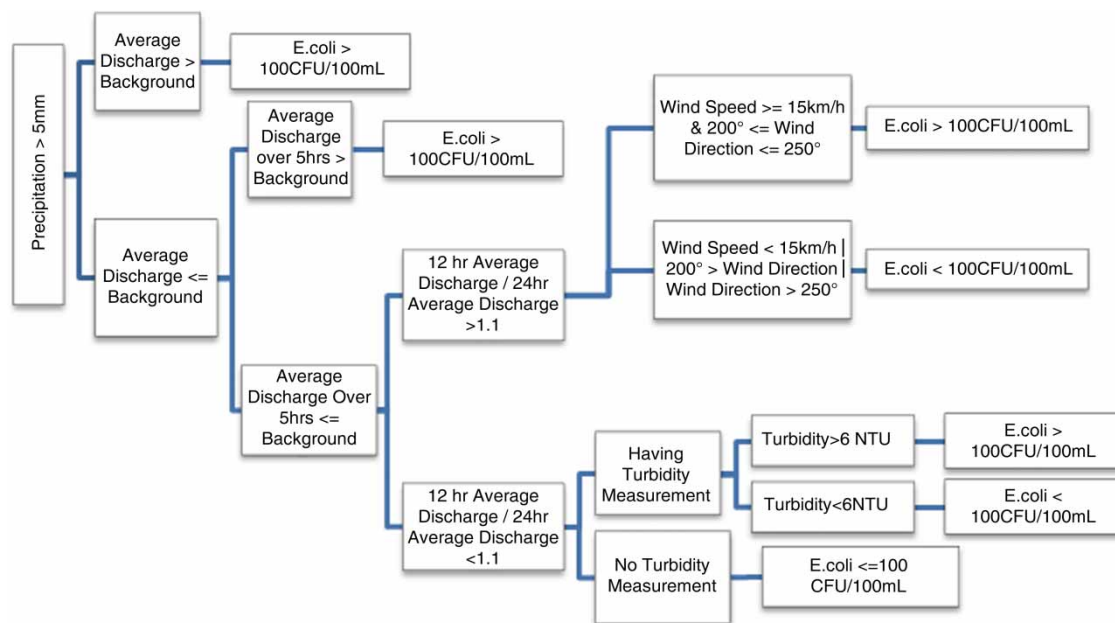


Figure 6 | Flow chart of branch 1 in the developed model.

the secondary cumulative discharge amount and the discharge value at the 10 a.m. sampling time would be assessed next to see if a large event occurred prior to the first 24-hour discharge window. The cumulative discharge threshold of 900,000 m³ was found to be the optimal value to predict large effects on bacteria level variations based on 2013 observations; this could be refined as more data are gained over time. The turbidity data would then be sequentially evaluated at various stages, depending on the secondary cumulative discharge amount. The turbidity parameter was assessed to improve the prediction because in this sub-branch of the model, the average discharge was weaker and thus had less influence compared to other sub-branches.

The wind influence was also introduced under this 'average discharge' scenario, because of the creek discharge location and non-uniform lake current in front of the beach area. Even under small elevated discharge events, it was found after many assessments that the plume from the creek would likely have an influence on beach bacteria levels unless the wind speed was larger than 30 km/h and the direction was in the range >305 and <125 degrees, which would likely push the entire plume away from the beach. The wind was examined during the 5-hour interval up to and including

10 a.m. of the low discharge period, and was considered to have a significant impact if at least 60% (e.g. >3/5) of the valid hourly readings met the criteria of the wind speed and direction. The purpose of using a 5-hour assessment window was to reduce the uncertainties of the represented wind. If the wind conditions were not met, turbidity during the event was re-assessed to see if it exceeded a secondary lower limit of 2.0 NTU; if so, the beach water could still be affected by the creek discharge and the PDM would predict an elevated *E. coli* count for the day. However, if the lag time was longer than the pre-set window, the creek discharge would only influence beach water quality when a strong wind pushed the plume towards the beach.

Model performance

The developed PDM was first tested with the daily *E. coli* data measured in 2011, 2012, and 2013. The accuracy of the new model predictions were compared with those from the presently used model by the City of Toronto for the MCPEB and the results are listed in Table 2. It can be seen that the performance of the newly developed PDM was more consistent for the tested years, as well as more accurate than the current City of Toronto model. It is important to note that the data

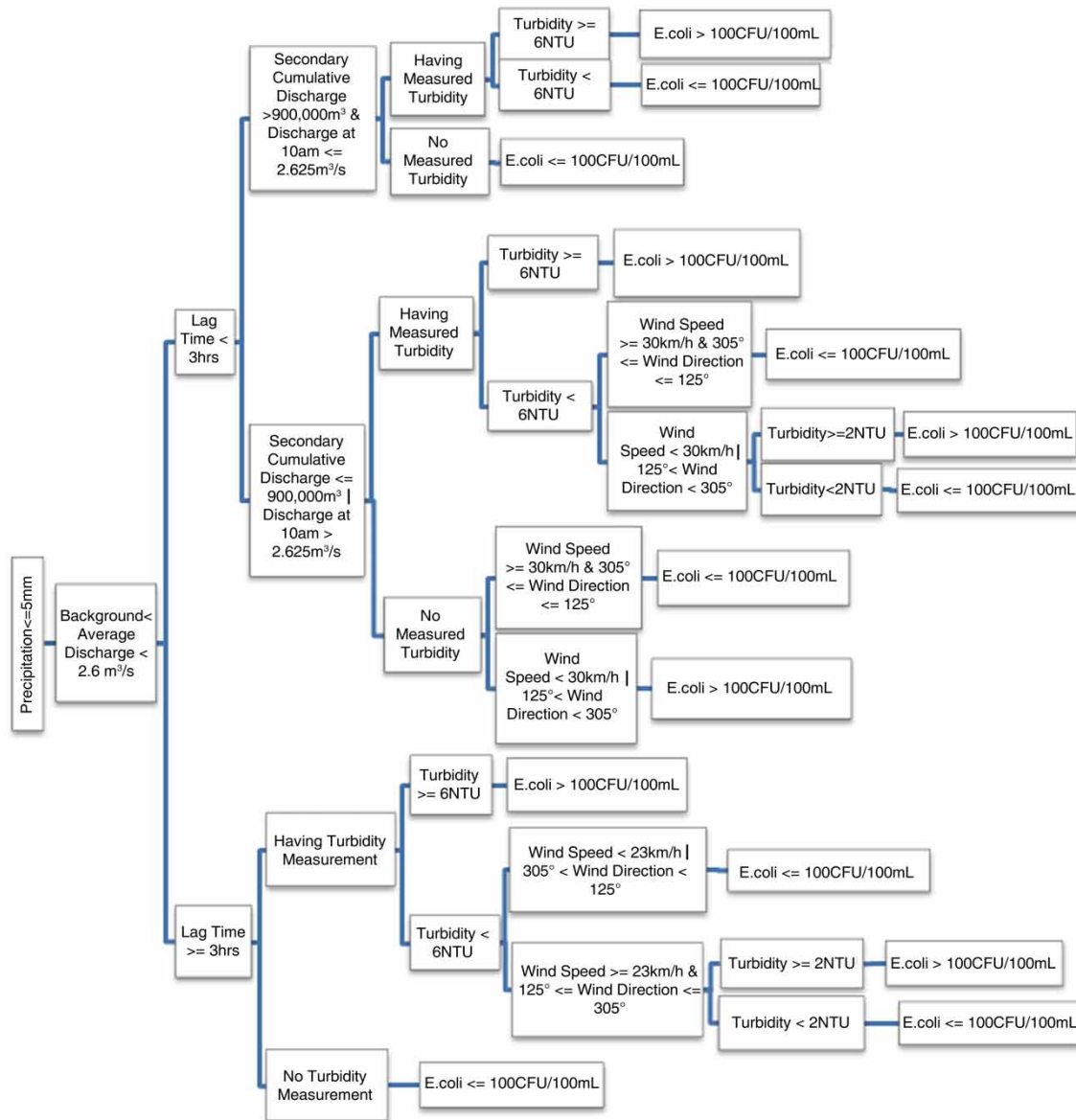


Figure 7 | Flow chart of branch 2–5 in the developed model.

from 2011 to 2013 were used in the development of the PDM, therefore the above test was not independent, but this is not an uncommon practice when developing predictive beach models (Fuss & O'Neill Inc. 2010) in initial verification. In terms of model accuracy, the overall performance was tested instead of examining the over- or under-estimations separately. Forcing the beach to close unnecessarily due to a model's over-estimation might cause as much strong public concern as keeping an over-polluted beach open due to underestimation. To reliably verify the newly developed PDM, the 2014 and

2015 measured daily *E. coli* data were used to compare the performance of each model (Table 2). The accuracy of prediction is similar to that shown in other years; it is consistently above 75%. Because of unavailability, the newly developed model could not be directly compared with the previously developed MLR model with the same year measurement data. But, the notable difference of prediction accuracy in different years should still provide some indications of their performance differences. In addition, a full analysis of the PDM results showed that the days it underestimated beach

bacteria levels tended to be days where the levels were only slightly higher than the provincial standard. The PDM outputs used logical numbers (above or below the limit), so the widely used root mean square error could not be applied to assess the model's potential error.

Factors affecting model predictions

The newly developed PDM was still considered to be inaccurate for up to 25% of the swimming season each year. The majority of the underestimated days were those days when there were no clear signs from the measured data to indicate a possibly elevated *E. coli* level. This was mainly attributed to the fact that there were no water quality monitoring data for the creek discharge which were available for the PDM input. In addition, due to limited data available at the MCPEB, the hourly wind data used in this study were accessed from the Toronto City Centre Airport and therefore were not local to the MCPEB. Local conditions, while similar, could also experience subtle differences; predictions based on these wind data could therefore generate inaccuracies.

The hourly precipitation data were also not local to the beach, even though they were checked to a certain degree by comparison with the discharge from the Etobicoke Creek. Isolated precipitation events local to the beach (e.g. thunderstorms) that were not measured at the Toronto City auto station and did not notably increase the discharge of the creek, may have been missed and the PDM could make an incorrect prediction.

Finally, the default water sampling time was set to be 10 a.m. every day to simplify the model. In reality, these samples were taken any time between 9 a.m. and noon.

Therefore, in some cases, events that happened after the sampling time might be included in the PDM and could lead to overestimation errors.

CONCLUSIONS

Recognizing the need for accurate and rapid prediction of recreational water quality and the failure of the statistic-based predictive model previously developed for MCPEB, in this study, a PDM was developed in MATLAB to predict *E. coli* levels at the MCPEB. To examine the influences of the various physical factors on the bacteria levels of the beach water at chest depth, field data, such as precipitation, discharge from the nearby creek, lake current velocity, turbidity, wave height, water temperature, wind speed and direction, water level and other data, had been collected and examined. The challenge of this study was that the bacteria level in the water was not only affected by multiple physical factors, but was also the function of accumulation (from various sources), sampling time, event start and end times for discharges and so on. To account for these complex and nonlinear relationships, the main controlling factors were examined sequentially, starting with the most influential physical factor. Various lengths of examining windows and lag times prior to the sampling time, precipitation and creek discharge amounts, and different threshold values were adopted in the model to determine bacteria levels. The newly developed PDM identified causes and incorporated many detailed physical relationships. The prediction accuracy of the bacteria levels from the newly developed PDM has been greatly improved compared with the presently used geometric mean model. In spite of apparent improvements, the model was inaccurate for 20–25% of the days. The future focus will be on these exception days, in order to continue improving PDM performance in the process of model implementation.

Table 2 | Performance comparisons between the newly developed PDM and the model presently used by the City of Toronto for predicting beach closure at the MCPEB

Year	Number of correct predictions	Total number of days	Prediction accuracy of PDM (%)	Prediction accuracy of City of Toronto's model (%)
2011	73	91	80	70
2012	69	88	78	57
2013	65	87	75	63
2014	65	86	76	57
2015	65	83	78	61

ACKNOWLEDGEMENTS

The present study was financially supported by the Great Lakes Action Plan (GLAP V), with funding by Environment Canada. Thanks are also due to the City of

Toronto for the daily *E. coli* and turbidity data. Lake flow condition data were collected by the Engineering Section of the National Water Research Institute. Finally, the authors wish to express their appreciation to the two anonymous reviewers for their detailed suggestions and comments serving to improve the quality of the paper.

REFERENCES

- Ackerman, D. & Weisberg, S. B. 2003 Relationship between rainfall and beach bacterial concentrations on Santa Monica Bay beaches. *J. Water Health*. **1** (2), 85–89.
- Beverdorf, L. J., Bornstein-Forst, S. M. & McLellan, S. L. 2007 The potential for beach sand to serve as a reservoir for *Escherichia coli* and the physical influences on cell die-off. *J. Appl. Microbiol.* **102**, 1372–1381.
- Eleria, A. & Vogel, R. A. 2005 Predicting fecal coliform bacteria in the Charles River. *J. Am. Water Resour. Assoc.* **41** (5), 1195–1209.
- Francy, D. S. 2009 Use of predictive models and rapid methods to nowcast bacteria levels at coastal beaches. *Aquat. Ecosyst. Health Manage.* **12** (2), 177–182.
- Francy, D. S. & Darner, R. A. 2002 *Forecasting bacteria levels at bathing beaches in Ohio*. U.S. Geological Survey Fact Sheet FS-132-02, US Geological Survey, Columbus, OH.
- Francy, D. S., Darner, R. A. & Bertke, E. 2006 *Models for Predicting Recreational Water Quality at Lake Erie Beaches*. U.S. Geological Survey Scientific Investigations Report 2006-5192, US Geological Survey, Columbus, OH.
- Fuss & O'Neill Inc. 2010 *Document 2: Statistical Analysis and Modeling Approach*. Project No. 20091440.A10. West Springfield, MA, USA, Fuss & O'Neill Inc., Manchester, CT.
- Gao, G. H., Falconer, R. A. & Lin, B. L. 2011 Numerical modelling of sediment bacteria interaction processes in surface waters. *Water Res.* **45**, 1951–1960.
- Helsel, D. R. & Hirsch, R. M. 1992 *Statistical Methods in Water Resources*. Studies in Environmental Science Volume 49. Elsevier, Amsterdam.
- Hydro Qual Inc. 1998 *Modeling Evaluations and Users Guide*. HydroQual, Inc., Mahwah, NJ.
- 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.
- Kuntz, J. E. 2006 Predictability of swimming prohibitions by observational parameters. In: *New England Interstate Water Pollution Control Commission Beach Closure Workshop, 7 February 2006, Lowell, Massachusetts*. New England Interstate Water Pollution Control Commission.
- Liu, L., Phanikumar, M. S., Molloy, S. L., Whitman, R. L., Shively, D. A., Nevers, M. B., Schwab, D. J. & Rose, J. B. 2006 Modeling the transport and inactivation of *E. coli* and enterococci in the nearshore region of Lake Michigan. *Environ. Sci. Technol.* **40** (16), 5022–5028.
- Mas, D. M. & Ahlfield, D. P. 2007 Comparing artificial neural networks and regression models for predicting faecal coliform concentrations. *Hydrolog. Sci. J.* **52** (4), 713–731.
- 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. 2011 Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches. *Water Res.* **45** (4), 1659–1668.
- Olyphant, G. A. 2005 Statistical basis for predicting the need for bacterially induced beach closures: emergence of a paradigm? *Water Res.* **39** (20), 4953–4960.
- Thoe, W. & Lee, J. H. W. 2014 Daily forecasting of Hong Kong beach water quality by multiple linear regression models. *J. Environ. Eng.* **140** (2), 345–353.
- Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M. L. & Boehm, A. B. 2015 Sunny with a chance of gastroenteritis: predicting swimmer risk at California beaches. *Environ. Sci. Technol.* **49**, 423–431.
- Thupaki, P., Phanikumar, M. S., Beletsky, D., Schwab, D. J., Nevers, M. B. & Whitman, R. L. 2010 Budget analysis of *Escherichia coli* at a Southern Lake Michigan beach. *Environ. Sci. Technol.* **44** (3), 1010–1016.
- US Environmental Protection Agency 1999 *Review of potential modeling tools and approaches to support the BEACH program*. USEPA Office of Science and Technology, 823-R-99-002, Washington, DC.
- Wymer, L. J., Brenner, K. P., Martinson, J. W., Stutts, W. R., Schaub, S. A. & Dufour, A. P. 2005 *The EMPACT Beaches Project: Results from a Study on Microbiological Monitoring in Recreational Water*. United States Environmental Protection Agency, Technical Report 600/R-04/023, Washington, DC.

First received 10 September 2015; accepted in revised form 2 February 2016. Available online 29 February 2016