Citation: Bertone, E., Purandare, J., Durand, B. (2019). “Spatiotemporal prediction of *Escherichia coli* and Enterococci for the Commonwealth Games triathlon event using Bayesian Networks” Marine Pollution Bulletin, Volume 146, Pages 11-21

Spatiotemporal prediction of *Escherichia coli* and Enterococci for the Commonwealth Games triathlon event using Bayesian Networks

E. Bertone^{a,b}, J. Purandare^{b,c}, B. Durand^c

^a *School of Engineering and Built Environment, Griffith University, Gold Coast Campus, QLD 4222, Australia*

^b *Cities Research Institute, Griffith University, Gold Coast Campus, QLD 4222, Australia*

^c *Gold Coast Water and Waste, City of Gold Coast, QLD 4211, Australia*

Email: e.bertone@griffith.edu.au

Abstract:

A number of Bayesian Networks were developed in order to nowcast and forecast, up to 4 days ahead and in different locations, the likelihood of water quality within the 2018 Commonwealth Games Triathlon swim course exceeding the critical limits for Enterococci and *Escherichia coli*. The models are data-driven, but the identification of potential inputs and optimal model structure was performed through the parallel contribution of several stakeholders and experts, consulted through workshops. The models, whose main nodes were discretised with a customised discretisation algorithm, were validated over a test set of data and deployed in real-time during the Commonwealth Games in support to a traditional water quality monitoring program. The proposed modelling framework proved to be cost-effective and less time-consuming than process-based models while still achieving high accuracy; in addition, the added value of a continuous stakeholder engagement guarantees a shared understanding of the model outputs and its future deployment.

Keywords: Bayesian Network; Enterococci; *Escherichia Coli*; Prediction Modelling; Triathlon; Water Resources Management

1 Introduction

Gastrointestinal diseases can be contracted when swimming in faecal contaminated waters having high *Escherichia coli* (*E. Coli*) or Enterococci counts (Wade et al. 2006, Wiedenmann et al. 2006), which are the most typically monitored faecal indicator bacteria (FIB). Thus, understanding, monitoring and potentially predicting FIB levels in swimming waters is critical, especially if a competitive sport event is planned, and organisers want to ensure the athletes' health will not be compromised. Since culture-based test methods or the most probable number (MPN) microplate method can take up to 2 days to estimate *E. coli* and Enterococci numbers, there is a need for a faster, near real-time estimation (Lušić et al. 2016) of FIB to achieve a more proactive management of recreational waters. Currently, it is not possible to monitor FIB in real-time, and as such the most achievable option is to predict FIB concentrations based on the state of potential input parameters.

E. coli and Enterococci levels in natural waters are particularly impacted by rainfall and storms (Olyphant et al. 2003, Solo-Gabriele et al. 2000), tide (Boehm et al. 2003, Solo-Gabriele et al. 2000), and turbidity (Money et al. 2009, Whitman et al. 2004). Other potential predictors include temperature and oxidation reduction potential of the water (Praveena et al. 2013). There are also a potential number of anthropogenic inputs, such as stormwater, sewage discharge or farm animals (Anastasi et al. 2012, Ishii and Sadowsky 2008, ten Veldhuis et al. 2010).

Predictive *E. coli* and Enterococci models have been previously developed. Nevers and Whitman (2005), using a regression modelling approach, tried to predict *E. coli* levels at a number of effluent-dominated Lake Michigan beaches. The model was very sensitive to wind direction and could predict threshold (i.e. Environmental Protection Agency closure limit, being 235 CFU/100 ml) exceedances on only 6 out of 11 occasions. A regression approach was also used by Hou et al. (2006) to predict Enterococci in a Californian beach. Money et al. (2009) applied a simple log-regression modelling approach to predict *E. coli* based on turbidity values, and used a Bayesian Maximum Entropy method to extrapolate the predictions to unmeasured spatial locations. Herrig et al. (2015) developed

predictive regression models for, among others, *E. coli* and Enterococci and an accuracy of around 70% was achieved for both models, although slightly lower for the simplified models having only three inputs, i.e. solar radiation, ammonium nitrogen and turbidity. A hybrid (expert- and data-driven) Bayesian Network (BN) model was developed by Goulding et al. (2012) to estimate the adverse health risks related to sewage overflow and certain raw sewage *E. coli* and Enterococci concentrations, based on different waterways uses; a number of scenarios were assessed and discussed, but the model was not used for a specific real-time estimation of *E. coli* concentrations and health risks. BNs have been also developed for prediction of the removal efficiency of a number of pathogens and microbial indicators (including *E. coli*) of an activated sludge system, based on historical data (Carvajal et al. 2015). Different data-driven predictive models were developed to “nowcast” FIB at Santa Monica beach in the United States for recreational water quality management purposes by Thoe et al. (2014). It was found that the most “complex” models (artificial neural network, classification tree) outperformed other more basic models such as multiple linear regressions. Despite these previous attempts, most of the described models did not aim at forecasting the output, but simply at “nowcasting”, i.e. avoiding the 18-24 hours lag between the sample and the laboratory result (Brooks et al. 2016). The ability to accurately forecast FIB few days in the future would allow much more flexible and proactive management of certain water resources.

Triathlon is an endurance sport that consists of swim, swim-to-cycle transition, cycle, cycle-to-run transition, and run over a variety of distances (Bentley et al. 2002). Research has shown that swimming performance is the one that can mostly affect the final result of a triathlete (Olbrecht 2011, Peeling and Landers 2009). A number of triathlon events were included in the programme of the 2018 Commonwealth Games held in the City of Gold Coast, in Queensland, Australia (GC2018), over a three-day period (5-7 April 2018). The event was governed by the International Triathlon Union (ITU), which stipulates specific monitoring criteria in the lead up to and during a triathlon event. This includes specific water quality minimum and maximum limits for Enterococci, *E. coli*, and pH. In order to ensure FIB-related parameters met ITU standards and ensured the safety of athletes during

the event, a data-driven BN model was developed to nowcast and forecast, up to 4 days ahead, the likelihood of water quality exceedances within the GC2018 Triathlon swim course.

2 Materials and Methods

2.1 Study location

The Broadwater, located in the northern Gold Coast (Figure 1), is a semi-enclosed tidal estuary that forms the primary connection of four of the City's major river systems (the Nerang, the Coomera, the Pimpama and the Logan-Albert) to the Coral Sea. The Broadwater is connected to the Coral Sea through two entrances – the original Nerang River entrance (locally known as the Seaway) located at the southern tip of South Stradbroke Island, which was trained and fixed in 1986, and Jumpinpin, which is a natural break in the sandy South and North Stradbroke Islands that was formed towards the end of the 19th century.

Figure 1 – Map of the Broadwater with sampling point locations. Top-right corner: Gold Coast map with relative location of the Broadwater

The Broadwater is one of the primary recreational water bodies on the Gold Coast and is utilised throughout the year for numerous water-based events, including the annual Luke Harrop Memorial Triathlon, held in April each year. The Broadwater is a macro tidal, diurnal inlet, which is relatively

hydrodynamically efficient. The key inputs of sediment and water to the Broadwater are from fluvial inputs, with tidal inputs being secondary.

2.2 *Methodological framework and stakeholders' engagement*

Figure 2 summarises the methodological approach that led to the development and deployment of the proposed BN; this generally follows the framework from Jakeman et al. (2006), but also highlights the extensive stakeholder engagement process adopted from start to end of the project.

Participatory modelling consists of involving stakeholders and experts in one or multiple steps of the modelling process (Hare 2011), substantially enhancing the likelihood of deployment of the final product. Participatory modelling implies that the model's underlying assumptions, limitations and intended uses are clear to the stakeholders, who at the same time have to go through the same thinking process as the modellers (Castelletti and Soncini-Sessa 2007). As such, an early involvement of relevant internal, i.e. within the Council of the City of Gold Coast (CoGC), and external (through the provision of a specially appointed panel) experts and stakeholders was essential to ensure such sharing understanding is established, and clarity and transparency are guaranteed throughout the model development process.

A number of workshops were arranged and experts provided (1) advice on the interpretation of the data analysis results, with potential indications on inclusion/exclusion of some predictor parameters in the BN; (2) feedback on original BN structure; and (3) comments and suggestions on the BN scenario analysis. Six workshops were conducted throughout the development process. Participants were selected based on their years of experience and field of expertise, and were divided into two groups: (1) an internal expert working group comprised of stakeholders from within the CoGC's various environment and catchment management, environmental health, and scientific services units providing local, operational and historical knowledge of the Broadwater; and (2) an external expert working group, facilitated by the International Water Centre (IWC), that was comprised of known local experts in the field of water quality, water management, and bacterial and microbial science. Thus stakeholder engagement was effectively used for participatory modelling, with experts' inputs

used alongside numerical data analysis outputs to construct and validate the BN. Through an iterative process, experts helped conceptualising the initial model (i.e. identifying the target variables, predictors, and scope), providing feedback on the data analysis outputs (i.e. unexpectedly high/low correlations), and refining/simplifying the conceptual model through the integration of their expert knowledge and what extracted from the numerical data.

Figure 2 – Conceptual methodological framework highlighting stakeholders' engagement

2.3 *Data analysis and statistical modelling*

To inform the GC2018 Triathlon events, a water quality monitoring program (GC2018 Water Quality Project) was designed and established by the CoGC, which represented the main source of data alongside the Australian Bureau of Meteorology (BoM). From a spatial point of view, three sites where the swimming portion of the Triathlon event was to take place, identified as monitoring sites CWG/GWCW T2, T3 and T4, were analysed and modelled (Figure 1). Data for *E. coli* and

Enterococci for such locations was available for variable time periods between 2016 and 2018. Rainfall data were made available for four different locations; however after preliminary data analysis, it was decided to focus on the two locations in closest proximity to the swimming course, i.e. the Gold Coast Seaway, located downstream the event location, and Loders Creek, which discharges in close vicinity to the event course. Rainfall data were deemed important as directly linked to river flow, which as mentioned before is a main source of sediments for the Broadwater. In addition, unlike other collected data, rainfall data for these locations was available both in daily and hourly frequency formats, from May 2005 to 2018. Although the inclusion of Gold Coast Seaway rainfall data for analysis seemed to be counterintuitive, as this is measured downstream the sampling locations, it was deemed important to include it because:

- It is the main rainfall measurement point for the Gold Coast and thus had the largest amount of historical data available;
- Due to the high variability of rainfall events in the region, relying on two stations allows to capture more rain events, leading to a more reliable and robust model development;
- Since tidal inputs are also an important source of sediments for the area, although the station is located downstream it can still provide valuable predictive power.

Continuous physical-chemical data from the GC2018 Triathlon Water Quality Project were also available for a 6-month period starting in August 2017. These data was collected by two continuous monitoring systems: (1) a solar powered metocean buoy, which included sensors measuring chlorophyll and phycoerythrin, temperature, salinity, pH, dissolved oxygen and current, as well as atmospheric conditions (including air temperature, pressure, wind, rainfall and relative humidity); (2) a sample filtration and archiving instrument (SAFA) which collected in-situ samples at set intervals, storing up to 24 samples within the device, which were collected periodically for laboratory testing using quantitative polymerase chain reaction (qPCR) techniques.

Data was analysed in order to understand: (1) the correlation of *E. coli* and Enterococci peak events with rainfall amount, duration and location; (2) importance of tide; (3) importance and correlation with other data continuously monitored, including turbidity. These activities were designed based on

knowledge acquired from the literature and through experts' advice provided during the internal and external stakeholder workshops.

Data was firstly visually inspected and pre-processed. FIB data were typically provided in triplicates; the highest result only for each triplicate was retained, in order to develop a conservative estimate of each scenario. The thresholds of concern as defined by ITU are 250 CFU/100mL for *E. coli* and 100 CFU/100mL for Enterococci, where CFU stands for colony forming units. Different cumulative rainfall variables were created and the optimal inputs were retained (i.e. those yielding highest correlations with *E. coli* and Enterococci). Linear and nonlinear regression and scatter plots were subsequently used to statistically identify which variables to retain as model inputs and which ones to discard.

2.4 Bayesian Network development and nodes discretisation

Bayesian Networks (BN) are probabilistic graphical models, which can handle highly uncertain systems, missing data, as well as the integration of empirical data and expert opinion (Chen and Pollino 2012, Uusitalo 2007). BN are, more specifically, directed acyclic graphs, with each variable presented as a node. Each node in a BN is represented as a discrete parameter that can only take on one of a number of predefined states (e.g. "high" or "low"). Nodes are connected through "arcs", with the direction of the arc going from a "parent" node(s) to a "child" node(s). Such connections, representing conditional dependence, are quantified through probability distributions. These probabilities are defined in the Conditional Probability Tables (CPTs) for each child node. BN have been previously applied by the authors in similarly, highly participatory, water-quality related modelling problems (Bertone et al. 2016) and in general they provide a suitable modelling candidate to achieve the research goals, since: (1) the available data have different frequencies and timeframes; (2) there is high uncertainty in some results (including triplicates); and (3) there was a need of involving all the relevant stakeholders to both get input for model development and to guarantee future model deployment.

As explained in Figure 2, the BN structure was initially qualitatively conceptualised based on background knowledge of the system; however, this was iteratively modified through data analysis outputs and feedback from consultation with stakeholders and experts. Different BNs were developed to model *E. coli* and Enterococci exceedances, as well as for the three different sampling locations. Although the conceptualisation of the models for the three different locations was the same (i.e. same relevant predictors were identified), the numerical quantification of the correlations between parameters yielding the highest accuracy was expectedly different between the three sites. Hence, as explained later, three BNs having the same structure but different CPTs were developed. Also, although time is not easily integrated in BN, as well as for instance feedback loops (Uusitalo 2007), for this application the models were effectively dynamic, designed to predict, sequentially, the state of the target node for the same day (“now-casting”) and for each following day up to four days ahead. The models’ final structure is displayed later as it was modified based on data analysis results and experts’ consultation. The relationships between parent-child nodes are numerically quantified through the CPTs; the inherent probabilities of different states of the child node given the state of the parent nodes were extracted from the historical data. The CPTs of parentless nodes were not defined for the default BN, as these were mainly weather forecasts input variables and would be populated based on day-specific BoM forecasts (which are probability-based). The values for the main CPTs are reported in Appendix A. The BN models were developed with the software Netica 5.18 64bit (Norsys Software Corp.).

One of the most common challenges with BN development is the discretisation of continuous variables, since this can only capture some of the features of the original continuous distribution (Friedman and Goldszmidt 1996). As discussed in Uusitalo (2007), a number of discretisation algorithms have been proposed in the past, as well as more recently (e.g. Chen et al. (2017), Mayfield et al. (2017)), but there is no consensus on a generic satisfactory algorithm for all BNs (Myllymäki et al. 2002), and thus a model- and problem-specific discretisation still seems to be the most reasonable approach (Uusitalo 2007).

In this case, since the target nodes had user-defined states based on the research objectives (i.e. provide the probability of exceedance of 250 CFU/100mL for *E. coli* and 100 CFU/100mL for Enterococci), the main discretisation problem was related to the parent nodes, which based on the outputs of data analysis and expert inputs, resulted to be only two per target child node. Firstly, it was decided to keep two states per node only, given the limited dataset, in order to be able to have evidence for all the resulting CPTs combinations. Based on this, a discretisation algorithm was required to find the optimal discriminatory threshold between states maximising the performance of the BN model. A code was developed in Matlab R2017a (The Mathworks Inc.) for this purpose. The code, based on historical input data, runs through all the potential threshold values a and b for the parent nodes X_1 and X_2 , and for each scenario assessed, it evaluates the four resulting inputs' state combinations for the target node's CPTs (i.e., $X_1 > a \ \& \ X_2 > b$, $X_1 > a \ \& \ X_2 < b$, $X_1 < a \ \& \ X_2 > b$, $X_1 < a \ \& \ X_2 < b$). For each scenario, a score is calculated. Since the purpose is to find the threshold values a and b minimising uncertainty by clearly separating high and low *E. Coli* and Enterococci counts, for each inputs' state combination of the CPT the two probabilities must be as far apart as possible. That is, a 50%/50% chance of high/low FIB does not provide insightful information, while a 90%/10% provides a quite clear separation, thus less uncertainty. Therefore, the more apart the numbers are, the higher the score would be. However, it may as well be that for a particular inputs' state combinations of the CPT, there are no or few counts (i.e. limited number of days in the historical dataset where the parent nodes had values within those combinations of states – e.g. high rainfall amount and low cloud cover). Hence, the score also penalises scenarios where the historical FIB data are not evenly distributed among the 4 inputs' state combinations for the target node's CPTs. After all the scenarios of different thresholds are run, the one leading to the highest score was saved and used, since it represents the discretisation of parent nodes providing the lowest uncertainty and optimal historical data distribution.

2.5 *Bayesian Network validation and deployment*

The model was validated with an independent *E. coli* and Enterococci test set of data collected between 1st February 2017 and 24th April 2017 ($n=33$). Accuracy was checked for the “nowcasting” nodes of the models (i.e. predicting for the same day of the collected sample). The forecasting models

could not be checked due to the datasets not being continuous (i.e. often *Enterococci* and *E. coli* were not sampled for a number of consecutive days). The models were nevertheless deployed and tested during the actual GC2018 Triathlon events period.

A method to test the accuracy of a BN classifier is to estimate the receiver operating characteristic curve (ROC) and related Area Under the Curve, or AUC (Hanley and McNeil 1982, Ling et al. 2003). The ROC illustrates the diagnostic ability of the model when the discriminatory threshold is varied. The x axis represents the false positive (FP) rate (i.e. $1 - \text{specificity}$) of the model, and the y axis the true positive (TP) rate (i.e. sensitivity). The line $y=x$ would represent a worthless model (i.e. random). As a rule of thumb, AUC values over 0.80 are associated to good models, while over 0.90, are associated to excellent models. AUC of 0.50 would represent the accuracy of a random model.

The following step consisted in finding the optimal discriminatory threshold (i.e., the minimum BN posterior probability value to be linked to a FIB exceedance), in order to maximise the BN models' accuracy. Based on the test set of data, both the general BN model accuracy (i.e. sum of true positive and true negatives, divided by n) and the Youden's J statistic (Youden 1950), which sums sensitivity and specificity-1, were estimated for different discriminatory thresholds ranging between 0% and 100%. The Youden's J statistic was preferred to the general model accuracy, as the latter would be affected by the fact that the dataset had only few exceedances, meaning that it would not be affected by very low specificity as long as there are several true negatives; the Youden's J statistic instead evenly weighs sensitivity and specificity, thus it gives the same importance to how the BN model predicts both exceedances and low values.

3 Results and Discussion

3.1 Identifying peak predictors and spatiotemporal variations

Through comprehensive data analysis, the highest correlation with *E. coli* and Enterococci data was found with the cumulative rainfall amount fallen in the previous seven days (Figure 3).

Figure 3 – Time series for Gold Coast Seaway rainfall, Loders Creek rainfall, and *E. coli* at T3.

Additionally, a strong correlation was found when, as second input, the maximum hourly rainfall amount of the past day was added. Thus two critical lagged rainfall-related predictors were identified, facilitating the development of a more accurate forecasting model. Similar patterns and accuracy levels were achieved for the three different locations, however T3 was selected as the main modelling site due to a higher availability of historical data. Another spatial consideration was related to the location of the input rainfall data. Looking at the coefficients of determination (R^2) of the linear regression models developed for T3 using Gold Coast Seaway historical rainfall data versus Loders Creek rainfall data, it can be noticed that higher accuracy was achieved using the Seaway data (Table 1). This was somewhat counterintuitive, since this is located downstream T3, and thus a stronger correlation with Loders Creek's rainfall, located upstream, seemed more logical. This became one of the main discussion points during the earlier experts' workshops.

Table 1. R^2 values of FIB regression models, T3

Model	With Gold Coast Seaway rainfall inputs	With Loders Creek rainfall inputs
<i>E. coli</i>	0.75	0.68
Enterococci	0.74	0.67

Interestingly however, when the regression models were used over historical data for predictions of exceedances only (i.e. not for prediction and comparison with the exact measured values), then such accuracy discrepancy vanished, with the model relying on Loders' Creek rainfall data even slightly outperforming the other one (Table 2). This could be due that R^2 based on highly variable data might not be the optimal accuracy metric as it does not explicitly account for the variance/covariance in the regression matrix, and thus it could be misleading when highlighting the best input parameters; looking instead at problem-specific thresholds of concern, it seems that the most logical inputs are identified as the best ones too.

Table 2. Exceedances prediction accuracy of FIB regression models, T3

Model	With Gold Coast Seaway rainfall inputs	With Loders Creek rainfall inputs
<i>E. coli</i>	78.4%	79.4%
Enterococci	76.5%	83.3%

In order to rely on data but also on experts and on a logical representation of the system, it was agreed to use the maximum rainfall amount recorded at either of the two locations as input (both for the cumulative past week rainfall and for the maximum hourly rainfall of the past day). This allowed to account for the high rainfall variability typical of the area, provided a more conservative estimate, and yielded a regression model with high accuracy especially in terms of the ability of predicting thresholds' exceedances (Table 3).

Table 3. Accuracy of final rainfall regression model

Model	R^2	Correct predictions
<i>E. coli</i>	0.68	78.4%
Enterococci	0.68	82.6%

Following the first external expert workshop, it was suggested to check the effect of the length of the dry spell before the analysed day, i.e. the number of days since the last rainfall event. This was agreed due to the ephemeral nature of the feed river systems (upper catchment zones), which can often result in the accumulation of sediment and nutrients in the upper catchment where no rainfall is available to ‘flush’ the system. A new regression model for T3 was created for *E. coli* and Enterococci, with 3 inputs (max hourly rain, rain past week, and days since last rain). This was calculated using data for the location which had the largest amount of rainfall. The correlation did not improve significantly from the 2-input model, with adjusted- R^2 improvements even lower due to the extra input. Hence this third variable was disregarded for modelling purposes.

Another focus of the workshops was the effect of tide. The Gold Coast is subject to a semi-diurnal tidal regime with neap and spring tidal ranges of 0.3m to 2m respectively (Sedigh et al., 2016). As such, it was expected that tides could have a significant impact on the residence time of FIB in the Broadwater, specifically in the GC2018 triathlon swim course; particularly during slack tides should the event coincide with such. However, no statistically significant relationship was found, that could have improved the accuracy of the rainfall-based models.

Despite only a little temporal overlap with the FIB data, the continuous physical-chemical data from the GC2018 Triathlon Water Quality Project were also checked for correlations. Also in this case, no significant correlations were found. Most of the parameters were relatively constant over time also during FIB peaks, with the only parameter with noticeable changes being turbidity.

Turbidity is widely understood to positively correlate with FIB, particularly *E. Coli*. (Malin et al. 2000), and as such, closer analyses were performed to determine the significance of the correlation between turbidity and FIB on the input flows from the main river catchments and at T3. Surprisingly, the correlation between turbidity and FIB was very weak, indicating that FIB did not have a significant relationship with turbidity, doing little to improve the model accuracy. This was discussed in depth within the expert panel, and conclusions drawn that resulted in the agreed exclusion of turbidity from the model.

Other expected parameters that showed some variations and were expected to be correlated with FIB such as dissolved oxygen and conductivity were also found to be instead uncorrelated. Given the high accuracy achieved by the rainfall-based models, and the propensity of BNs to work better with a limited number of parent nodes (to keep CPTs relatively small and manageable especially when only a small dataset is available), the lack of any other relevant predictor did not compromise the development of a reliable BN model.

Another suggestion from the first external experts' workshop, was to more explicitly incorporate the recovery rate information on the BN model, i.e. the time required for the concentration to return to pre-storm conditions. Based on the events analysed for recovery rate, a regression model was developed for T3 based on the predictor "rainfall previous day". Figure 4 shows the results for the *E. coli* model. Given the relevance of the correlation both based on data and experts' opinion, this was integrated into the BN model structure through a second FIB occurrence probability prediction calculation. The regression was transformed in a conditional probability table (CPT) based on the historical recovery rate data, with an incorporated level of uncertainty due to the limited historical data available. Although the regression coefficient itself is high mainly due to the extreme rainfall event outlier, by using historical rainfall data to populate CPTs, it is possible to explicitly determine the uncertainty and variability for each range of rainfall amount as defined by the CPTs states (e.g. there seems to be a lot of variability in recovery rate for rainfall amounts of around 20mm, but much more certainty for lower or higher amount of rainfall), and incorporate this in the predictions.

Figure 4 – Regression model, rain previous day vs *E. coli* recovery rate, T3

3.2 Bayesian Network models final structure

The final model structure integrates the rainfall-based and recovery rate-based models. Figures 5 shows, for *E. coli* T3, the BN model structure for $t=0$ (nowcasting), and for $t=1$ (one day ahead forecast). The full model, not shown here, replicates the components of Figure 5 to predict up to four days ahead. The values of the main CPTs with parent nodes, extracted from the data, are provided in Appendix A; the CPTs of the nodes not reported in the appendix are either a replicate of the ones provided, or they depend on day-specific weather forecasts. The Enterococci BN models, as well the models for the other locations (T2 and T4), have exactly the same structure, with different nodes, discretisation thresholds and different numbers in the CPTs, based on specific Enterococci data analysis (18-months) and discretisation algorithm outputs. It can be seen from Figure 4 that, for this particular model, the two optimal thresholds for the rainfall-based *E.coli* model CPT were 32mm for total rainfall in previous week, and 1mm for maximum hourly rainfall for the past day. The colour code is defined as follows:

- Light blue = BoM forecasts inputs
- Dark blue = Rainfall inputs
- Yellow = Calculations, recovery rate model

- Orange = Calculations, rainfall-based model
- Green = Final *E. coli* prediction

Figure 4 – BN model *E. coli* T3 subcomponent, $t=0$ and $t=1$

Looking at the nodes included in Figure 4 for the $t=0$ module, these can be described as follows:

- **Rain past 7 days MAX**: the cumulative amount of rainfall over the past 7 days. In order to calculate this, firstly, for each day, the location having the maximum rain amount is selected, in order to have the maximum rainfall amount recorded at any location for any given past day. Subsequently, a sum is made out of the 7 daily maximum rain amounts.
- **Max hourly rain last day (MAX)**: the maximum rainfall amount in any given hour of the past day, at any given location (either Gold Coast Seaway or Loders Creek).

- *E. coli* T3 $t=0$: the predicted *E. coli* probability of exceedance at $t=0$ according to the rainfall-based model. The letter “ t ” stands for time and for this model represent days.
- Rain last day: total rainfall amount in the past day at the maximum location.
- Recovery rate $t=0$: calculation of expected recovery rate according to *E. coli* levels recorded at $t=0$ and rainfall.
- *E. coli* T3 $t=1$ v2: probability of exceedance of *E. coli* for the next time step ($t=1$), based on recovery rate modelling. This will be compared with what was calculated for $t=1$ based on the rainfall model (*E. coli* T3 $t=1$ v1), and the maximum will be selected.

For rainfall-related parent nodes, when forecasting FIB up to 4 days ahead, also BoM forecasts were used, as previously mentioned. As these are usually provided in probabilistic terms (e.g. 70% chance of 10-20 mm of rainfall), the probabilistic, discrete nature of BN facilitates the integration of these information.

The BN model consists of a static component, which does not rely on older predictions, but only on lagged measured input data (i.e. rainfall-based model), and a dynamic BN component, which instead relies both on lagged input data and target node predictions at previous time steps (i.e. recovery-rate component). The final product is a hybrid static-dynamic BN which makes the most of different behaviours and correlations emerged from the data analysis, and is able to nowcast, as well as forecast, the probability of FIB exceeding ITU thresholds.

3.3 Model validation and deployment

The BN was validated using two separate methodologies. Firstly, statistical assessments of model accuracy were conducted using ROC and Youden’s J. Secondly, the BN was physically tested using historical rainfall data compared with actual FIB data collected from water quality sampling.

Figure 5 shows the ROC for both the *E. Coli* and Enterococci models, for location T3 predicting at time $t=0$ (nowcast). The resulting AUC was 0.85 for the Enterococci model and 0.86 for the *E. coli* model, representing accuracy values of very good models.

Figure 5 – ROC, BN *E. coli* and Enterococci T3, $t=0$

Figure 6 illustrates the calculated Youden's J statistics for the same two nowcasting models, estimated at a varying discriminatory probability threshold between 0% and 100%. The optimal discriminatory probability thresholds were found to be 37% for *E. coli* ($J=0.68$) and 35% for Enterococci ($J=0.59$); the resulting confusion matrix and performance indexes of the models at such optimal thresholds are displayed in Tables 4 and 5.

Figure 6 – Youden's J statistic at different discriminatory probability thresholds, BN *E. coli* and Enterococci T3 $t=0$.

Table 4. Confusion matrix and model performance at optimal discriminatory probability threshold,
BN *E. coli* T3 $t=0$

Confusion matrix Threshold 37%		Measured	
		Exceedance	Non-exceedance
Predicted	Exceedance	3	2
	Non-exceedance	1	27
Youden's J	Accuracy	Sensitivity	Specificity
0.68	90.9%	0.75	0.93

Table 5. Confusion matrix and model performance at optimal discriminatory probability threshold,
BN Enterococci T3 $t=0$

Confusion matrix Threshold 35%		Measured	
		Exceedance	Non-exceedance
Predicted	Exceedance	4	6
	Non-exceedance	1	22
Youden's J	Accuracy	Sensitivity	Specificity
0.59	78.8%	0.8	0.79

The validated models were run using historical rainfall scenarios for the preceding 12 years (2006-17) for the period where the triathlon event would occur (i.e. early April). This was to preventatively understand under which rainfall scenarios FIB would have exceeded ITU thresholds, as well as how often this would have occurred in the last 12 years, in order to estimate a “background” risk well

earlier than the GC2018 Triathlon events and in turn proactively plan and manage intensive monitoring of FIB and risks for athletes accordingly.

Daily rainfall data was obtained from BoM for the 12-year period and the period mid-March to mid-April was extracted. These data were then reviewed to determine the category of rainfall event (no rainfall, low rainfall, high rainfall, and event rainfall), and scenarios were developed from the data to represent each potential scenario of rainfall event. Each scenario was then run through the model and the FIB outputs recorded. This exercise demonstrated the potential exceedance levels that were presented for each scenario, and enabled more accurate communication of the risk to the management and event teams.

Further testing was conducted with live rainfall and FIB data in the weeks leading up to the event to fully test the accuracy of the BN. This involved actively running the live model with BoM forecasted and reported rainfall, recording the FIB output, and comparing the percentage likelihood of exceedance over the ITU requirements with actual water quality monitoring data collected. Due to laboratory processing times, actual FIB data took 24 hours to process and return, and as such the model was run continuously for the testing period, being updated with actual recorded and forecasted rainfall every few hours. Table 6 summarises the models' outputs from the real-time testing described previously. Both the nowcasting (i.e. 0 days ahead) and forecasting (1-3 days ahead) components of the models were run and thus tested.

Table 6 - BN model outputs against measured *E. coli* and Enterococci *immediately preceding and during event*. Bold numbers provide an example of how to read the table (i.e. measured value and predictions). Red numbers show the false positive.

Rainfall Period (24 hours)	Sample Date	Enterococci % likely to exceed				<i>E. Coli</i> % likely to exceed				Actual Results	
		0 days	+1 days	+2 days	+3 days	0 days	+1 days	+2 days	+3 days	Enterococci	<i>E. Coli</i>
25th 9:00 to 26th 9:00	26th 04:30	19.6%	47.1%	23.6%	23.2%	17.6%	34.7%	19.1%	18.8%	<10	27
26th 9:00 to 27th 9:00	27th 04:30	50.0%	47.1%	47.8%	3.9%	35.8%	32.9%	34.2%	2.9%	50	40
27th 9:00 to 28th 9:00	28th 04:30	50.0%	32.2%	40.3%	45.0%	35.8%	24.8%	31.9%	32.5%	80	78
28th 9:00 to 29th 9:00	29th 04:30	57.2%	44.4%	48.4%	55.3%	35.8%	31.9%	34.7%	41.7%	49	17
29th 9:00 to 30th 9:00	30th 04:30	40.2%	18.6%	25.2%	48.4%	31.9%	13.8%	26.5%	34.7%	65	110
30th 9:00 to 31st 9:00	31st 04:30	32.2%	47.1%	48.4%	48.4%	26.8%	34.7%	34.7%	34.7%	17	42
31st 9:00 to 1st 9:00	1st 04:30	19.6%	23.0%	23.6%	16.3%	11.3%	19.1%	19.1%	13.1%	8	21
1st 9:00 to 2nd 9:00	2nd 04:30	19.6%	23.0%	17.5%	3.9%	17.6%	19.1%	14.1%	2.9%	19	30
2nd 9:00 to 3rd 9:00	3rd 04:30	19.6%	16.9%	4.1%	3.9%	17.6%	13.9%	3.0%	2.9%	15	15
3rd 9:00 to 4th 9:00	4th 04:30	42.1%	16.5%	3.9%	3.9%	34.2%	13.6%	2.9%	2.9%	5	7
4th 4:00 to 5th 4:00	5th 04:30	10.5%	3.9%	3.9%	3.9%	11.3%	2.9%	2.9%	2.9%	17	27
5th 4:00 to 6th 4:00	6th 04:30	9.2%	11.8%	3.9%	3.9%	9.7%	9.4%	2.9%	2.9%	13	16
6th 4:00 to 7th 4:00	7th 04:30	9.2%	11.5%	3.9%	3.9%	9.7%	9.4%	2.9%	2.9%		

The information demonstrates that no exceedances were recorded through the field measurements during the event. It can be seen how all the 8 models performed very well in predicting low levels, particularly given that rainfall occurred almost every day, at moderate intensities, but only for a couple of hours at a time (the average daily rainfall recorded for the period was 6.4mm and the maximum rainfall that fell in an hour averaged 3.1mm). These rainfall conditions were historically the rarest, thus theoretically making it more difficult for the BN to predict accurately given the limited amount of training data under these conditions. There was only one false positive (42.1%, higher than discriminatory threshold) recorded for the Enterococci model (4th April), due to some rain falling on the preceding day. However, the predicted probability of exceedance (42.1%) is just above the optimal discriminatory threshold (35%), and still relatively low. Given such probability decreasing for the forecast for the following days, this did not cause any serious warning for the event running on the 5th. The measured values confirmed the levels being within ITU requirements. It can be noticed how the probabilities are slightly updated day after day (e.g. Enterococci model, 2 days ahead, 1st April:

23.6% but Enterococci model, today's prediction, 3rd April: 19.6%); this is due to discrepancies between the BoM forecast and the actual observed rainfall. However, the accuracy of the BN model was not affected.

4 Conclusions

A participatory modelling approach was applied to develop a Bayesian Network predicting probability of exceedance of critical thresholds for *E. coli* and Enterococci at different sites and up to 4 days ahead, for the location of the swim course of the 2018 Commonwealth Games triathlon event. The model relied mainly on rainfall-related predictors, as they resulted to be a good proxy for sediment input in the study area also given the higher amount of related historical data available compared to other potential predictors. The model was deployed in real-time during the games and predicted accurately.

The BN modelling framework proved to be a cost-effective, less time-consuming alternative to numerical modelling (such as hydrodynamic models) or traditional statistical models. Advantages over numerical models include much faster simulations, lower amount of minimum data and information required for model development, and clear representation of the underlying uncertainty in the predictions. Advantages over more simplistic statistical or other data-driven models are also several. Unlike “black-box” approaches such as Artificial Neural Network for instance, the BN explicitly defines the input and output variables (i.e. conceptualised model, defined by both experts and data), with their relationships defined by the site-specific data. The approach to developing and testing the BN, including continuous engagement with relevant experts and stakeholders, dramatically enhances the predictive power of a simple statistical analysis, by incorporating expert opinion and site-specific expertise in the final model. Although the models predictions are site-specific and would require recalibration and adjustments (as nevertheless it would be for any other model) if another case study location was to be selected, the methodological framework, consisting of re-iterations of data analysis, expert consultation and BN model refinement, could be easily re-applied by other researchers having similar amount of data and stakeholders available.

Acknowledgements

The authors would like to thank the International Water Centre for providing access to expert water professionals who encompassed the independent expert panel. Additional thanks to both the City of Gold Coast and GOLDOC teams involved in developing and delivering this project and the 2018 Commonwealth Games.

References

- Anastasi, E., Matthews, B., Stratton, H.M. and Katouli, M. (2012) Pathogenic *Escherichia coli* found in sewage treatment plants and environmental waters. *Applied and environmental microbiology* 78(16), 5536-5541.
- Bentley, D.J., Millet, G.P., Vleck, V.E. and McNaughton, L.R. (2002) Specific Aspects of Contemporary Triathlon. *Sports Medicine* 32(6), 345-359.
- Bertone, E., Sahin, O., Richards, R. and Roiko, A. (2016) Extreme events, water quality and health: A participatory Bayesian risk assessment tool for managers of reservoirs. *Journal of Cleaner Production* 135, 657-667.
- Boehm, A.B., Fuhrman, J.A., Mrše, R.D. and Grant, S.B. (2003) Tiered approach for identification of a human fecal pollution source at a recreational beach: case study at Avalon Bay, Catalina Island, California. *Environmental science & technology* 37(4), 673-680.
- Brooks, W., Corsi, S., Fienen, M. and Carvin, R. (2016) Predicting recreational water quality advisories: A comparison of statistical methods. *Environmental Modelling & Software* 76, 81-94.
- Carvajal, G., Roser, D.J., Sisson, S.A., Keegan, A. and Khan, S.J. (2015) Modelling pathogen log₁₀ reduction values achieved by activated sludge treatment using naïve and semi naïve Bayes network models. *Water Research* 85, 304-315.
- Castelletti, A. and Soncini-Sessa, R. (2007) Bayesian Networks and participatory modelling in water resource management. *Environmental Modelling & Software* 22(8), 1075-1088.
- Chen, S. and Pollino, C. (2012) Good practice in Bayesian network modelling. *Environmental Modelling & Software* 37, 134-145.
- Chen, Y.-C., Wheeler, T.A. and Kochenderfer, M.J. (2017) Learning discrete Bayesian networks from continuous data. *Journal of Artificial Intelligence Research* 59, 103-132.
- Friedman, N. and Goldszmidt, M. (1996) Discretizing continuous attributes while learning Bayesian networks, pp. 157-165.
- Goulding, R., Jayasuriya, N. and Horan, E. (2012) A Bayesian network model to assess the public health risk associated with wet weather sewer overflows discharging into waterways. *Water Research* 46(16), 4933-4940.
- Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1), 29-36.
- Hare, M. (2011) Forms of participatory modelling and its potential for widespread adoption in the water sector. *Environmental Policy and Governance* 21(6), 386-402.

- Herrig, I.M., Böer, S.I., Brennholt, N. and Manz, W. (2015) Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. *Water Research* 85, 148-157.
- Hou, D., Rabinovici, S.J.M. and Boehm, A.B. (2006) Enterococci Predictions from Partial Least Squares Regression Models in Conjunction with a Single-Sample Standard Improve the Efficacy of Beach Management Advisories. *Environmental science & technology* 40(6), 1737-1743.
- Ishii, S. and Sadowsky, M.J. (2008) *Escherichia coli* in the environment: implications for water quality and human health. *Microbes and Environments* 23(2), 101-108.
- Jakeman, A.J., Letcher, R.A. and Norton, J.P. (2006) Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software* 21(5), 602-614.
- Ling, C.X., Huang, J. and Zhang, H. (2003) AUC: A Better Measure than Accuracy in Comparing Learning Algorithms, pp. 329-341, Springer Berlin Heidelberg, Berlin, Heidelberg.
- Lušić, D.V., Jozić, S., Cenov, A., Glad, M., Bulić, M. and Lušić, D. (2016) *Escherichia coli* in marine water: Comparison of methods for the assessment of recreational bathing water samples. *Marine pollution bulletin* 113(1), 438-443.
- Mayfield, H., Bertone, E., Sahin, O. and Smith, C. (2017) Structurally aware discretisation for Bayesian networks, 22nd International Congress on Modelling and Simulation MODSIM, Hobart, Tasmania, Australia
- Money, E.S., Carter, G.P. and Serre, M.L. (2009) Modern space/time geostatistics using river distances: data integration of turbidity and *E. coli* measurements to assess fecal contamination along the Raritan River in New Jersey. *Environmental science & technology* 43(10), 3736-3742.
- Myllymäki, P., Silander, T., Tirri, H. and Uronen, P. (2002) B-course: A web-based tool for Bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools* 11(03), 369-387.
- Nevers, M.B. and Whitman, R.L. (2005) Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan. *Water Research* 39(20), 5250-5260.
- Olbrecht, J. (2011) Triathlon: swimming for winning. *Journal of Human Sport and Exercise* 6(2).
- Olyphant, G.A., Thomas, J., Whitman, R.L. and Harper, D. (2003) Coastal Monitoring through Partnerships, pp. 289-300, Springer.
- Peeling, P. and Landers, G. (2009) Swimming intensity during triathlon: A review of current research and strategies to enhance race performance. *Journal of Sports Sciences* 27(10), 1079-1085.
- Praveena, S.M., Chen, K.S. and Ismail, S.N.S. (2013) Indicators of microbial beach water quality: preliminary findings from Teluk Kemang beach, Port Dickson (Malaysia). *Marine pollution bulletin* 76(1-2), 417-419.
- Solo-Gabriele, H.M., Wolfert, M.A., Desmarais, T.R. and Palmer, C.J. (2000) Sources of *Escherichia coli* in a Coastal Subtropical Environment. *Applied and environmental microbiology* 66(1), 230-237.
- ten Veldhuis, J.A.E., Clemens, F.H.L.R., Sterk, G. and Berends, B.R. (2010) Microbial risks associated with exposure to pathogens in contaminated urban flood water. *Water Research* 44(9), 2910-2918.
- Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L. and Boehm, A.B. (2014) Predicting water quality at Santa Monica Beach: Evaluation of five different models for public notification of unsafe swimming conditions. *Water Research* 67, 105-117.
- Uusitalo, L. (2007) Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling* 203, 312-318.
- Wade, T.J., Calderon, R.L., Sams, E., Beach, M., Brenner, K.P., Williams, A.H. and Dufour, A.P. (2006) Rapidly measured indicators of recreational water quality are predictive of swimming associated gastrointestinal illness. *Environmental health perspectives* 114(1), 24-28.

Whitman, R.L., Nevers, M.B., Korinek, G.C. and Byappanahalli, M.N. (2004) Solar and temporal effects on *Escherichia coli* concentration at a Lake Michigan swimming beach. *Appl Environ Microbiol* 70(7), 4276-4285.

Wiedenmann, A., Krüger, P., Dietz, K., López-Pila, J.M., Szewzyk, R. and Botzenhart, K. (2006) A randomised control trial assessing infectious disease risks from bathing in fresh recreational waters in relation to the concentration of *Escherichia coli*, intestinal enterococci, *Clostridium perfringens*, and somatic coliphages. *Environmental health perspectives* 114(2), 228-236.

Youden, W.J. (1950) Index for rating diagnostic tests. *Cancer* 3(1), 32-35.

Appendix A

<i>E. coli</i> T3 t=0		Below 250	Above 250
Rain past 7 days MAX	Max hourly rain last day MAX		
< 32mm	< 1mm	100%	0%
< 32mm	> 1mm	88%	12%
> 32mm	< 1mm	86%	14%
> 32mm	> 1mm	65%	35%

Recovery rate* t=0		No event	1 Day	2 Days	3 Days or over
<i>E. coli</i> T3 t=0	Rain last day				
< 250	0-15	97.115%	0.962%	0.962%	0.962%
< 250	15-30	97.115%	0.962%	0.962%	0.962%
< 250	30-60	97.115%	0.962%	0.962%	0.962%
< 250	60-120	97.115%	0.962%	0.962%	0.962%
> 250	0-15	97.115%	0.962%	0.962%	0.962%
> 250	15-30	40.385%	57.692%	0.962%	0.962%
> 250	30-60	0.962%	56.731%	41.346%	0.962%
> 250	60-120	0.962%	0.962%	14.423%	83.654%

*Generated through “Equation to Table” function based on regression equation of Figure 4

<i>E. coli</i> T3 t=1 v2	Below 250	Above 250
Recovery rate t=0		
No event	100%	0%
1 Day	50%	50%
2 Days	33%	67%
3 Days or over	20%	80%

Predicted <i>E. coli</i> T3 t=0		Below 250	Above 250
<i>E. coli</i> T3 t=1 v1	<i>E. coli</i> T3 t=1 v2		
< 250	< 250	100%	0%
< 250	> 250	0%	100%
> 250	< 250	0%	100%
> 250	> 250	0%	100%

Rain past 7 days MAX (at t+1)

1. Rain past 7 days MAX = Rain past 6 days + BoM rain forecast (t+1)
2. “Equation to table” function in Netica to generate CPT

Same process applies to other “Rain past days MAX” “Recovery rate” variables. Due to the variability in the Netica simulations to generate the CPTs, these are slightly different from each other.

Enterococci T3 t=0		Below 100	Above 100
Rain past 7 days MAX	Max hourly rain last day MAX		
< 32mm	< 1mm	100%	0%
< 32mm	> 1mm	87%	13%
> 32mm	< 1mm	81%	19%
> 32mm	> 1mm	50%	50%

Recovery rate* t=0		No event	1 Day	2 Days	3 Days or over
Enterococci T3 t=0	Rain last day				
< 100	0-15	100%	0%	0%	0%
< 100	15-30	100%	0%	0%	0%
< 100	30-60	100%	0%	0%	0%
< 100	60-120	100%	0%	0%	0%
> 100	0-15	100%	0%	0%	0%
> 100	15-30	41%	59%	0%	0%
> 100	30-60	2%	52%	46%	0%
> 100	60-120	1%	0%	16%	83%

*Generated through “Equation to Table” function based on regression equation

Enterococci T3 t=1 v2	Below 100	Above 100
Recovery rate t=0		
No event	100%	0%
1 Day	50%	50%
2 Days	33%	67%
3 Days or over	20%	80%

Predicted Enterococci T3 t=0		Below 100	Above 100
Enterococci T3 t=1 v1	Enterococci T3 t=1 v2		
< 100	< 100	99.1%	0.9%
< 100	> 100	0.9%	99.1%
> 100	< 100	0.9%	99.1%
> 100	> 100	0.9%	99.1%

Rain past 7 days MAX (at t+1)

1. Rain past 7 days MAX = Rain past 6 days + BoM rain forecast (t+1)
2. “Equation to table” function in Netica to generate CPT

Same process applies to other “Rain past days MAX” “Recovery rate” variables. Due to the variability in the Netica simulations to generate the CPTs, these are slightly different from each other.