Water Resources Research

RESEARCH ARTICLE

10.1029/2019WR025736

Key Points:

- *E. coli* blooms in Australian water reservoirs were analyzed
- Potential predictors were identified but more data are required

Supporting Information:

- Supporting Information S1

Correspondence to:

E. Bertone,
e.bertone@griffith.edu.au

Citation:

Bertone, E., Kozak, S., & Roiko, A. (2019). Understanding and modeling the occurrence of *e. coli* blooms in drinking water reservoirs. *Water Resources Research*, 55. <https://doi.org/10.1029/2019WR025736>

Received 6 JUN 2019

Accepted 25 NOV 2019

Understanding and Modeling the Occurrence of *E. coli* Blooms in Drinking Water Reservoirs

Edoardo Bertone^{1,2}, Sonya Kozak³, and Anne Roiko^{1,3}

¹Cities Research Institute, Griffith University, Gold Coast, Queensland, Australia, ²School of Engineering and Built Environment, Griffith University, Gold Coast, Queensland, Australia, ³Menzies Health Institute, Griffith University, Gold Coast, Queensland, Australia

Abstract Certain strains of *Escherichia coli* have been reported to bloom in the environment, resulting in high concentrations in waters in the absence of apparent fecal input or human pathogens, and in turn, undermining its reliability as an indicator of recent fecal contamination. Given the capacity of environmental strains of *E. coli* to replicate in the environment, the objective of this research work was to investigate whether any of the routinely measured parameters could predict the onset of an *E. coli* bloom in drinking water reservoirs. Information from historical catchment, weather, and water quality data were extracted for a number of Australian reservoirs that experienced *E. coli* blooms. Data were preprocessed and analyzed with time series analysis, linear and nonlinear regression, and self-organizing maps. Findings suggest that warm water, dry catchments, algal blooms, and nutrient availability were important factors in increasing the propensity for a bloom. Nutrient availability can be affected by many extrinsic factors that are often not well characterized, such as bushfires and back burning, decomposition of aquatic species, and dust storms. Based on data analysis outputs, a data-driven Bayesian Network model was developed, which, considering the paucity of data for some key input parameters, should only be used to trigger more intensive monitoring programs whenever the predicted risk of a bloom exceeds predetermined key thresholds. Such new data could be fed into the model to continuously improve its accuracy, and to eventually predict and proactively manage future blooms.

1. Introduction

Total coliforms are defined as gram-negative, facultative anaerobic, rod-shaped bacteria. As some strains can grow in the environment outside of the gastric tract, coliforms are considered to be unreliable indicators of sewage contamination (Tyagi et al., 2006). On the other hand, *Escherichia coli* has been commonly used as an indicator since it is more likely to be associated with fecal contamination (Cumberland et al., 2012), and a number of studies have been conducted to model its concentrations and transport mechanisms in water environments (e.g., Hondzo and Al-Homoud (2007) or Dwivedi et al. (2013)). However, there is increasing evidence against it being an accurate indicator of fecal input (Blyton & Gordon, 2017); for instance, there seems to be a number of different phenotypes of *E. coli*, with some appearing to have adapted to a planktonic lifestyle in aquatic ecosystems and able to persist and replicate under favorable conditions, causing blooms. Others can survive outside of animal hosts by entering a viable but nonculturable state. Under this hypothesis, bacterial cells would enter a dormant state and could “resuscitate” given certain environmental conditions such as favorable temperature, pH, carbon, and nutrients (Pinto et al., 2011; Wei & Zhao, 2018; Winfield & Groisman, 2003). Considering that in Australia, where this research was conducted, the Australian Drinking Water Guidelines (NHRMC, 2011) presently do not cover environmental or blooming *E. coli*, it is important to better understand it and explore the potential for predicting blooms. Based on this, the main objective of this study, which was part of a larger project (Sinclair, 2019), was to collect and analyze historical catchment data for a number of Australian reservoirs and investigate whether there were any parameter which were significantly correlated to the occurrence of environmental *E. coli* blooms. A second objective was, based on the results of the data analysis, to develop a predictive model for environmental *E. coli* blooms. A probabilistic approach was used, and a Bayesian Network (BN) model was developed for this task. BNs have been extensively applied in environmental modeling (Aguilera et al., 2011) and in the prediction of health-related water quality events (Bertone, Sahin, et al., 2016) and represent an ideal modeling approach for uncertain, data-scarce systems (Barton et al., 2012; Uusitalo, 2007).

1.1. Background

An increasing number of studies seem to prove that measurements of fecal indicator organisms might not be suitable for characterizing human health risks in water (Byappanahalli et al., 2006; Edge et al., 2013; Ishii & Sadowsky, 2008; Luo et al., 2011; Percec-Merien & Lewis, 2013; Whitman et al., 2003). Nevertheless, *E. coli* is a water quality indicator that is extensively used in both fresh and marine waters worldwide. This bacterium is both genetically and phenotypically diverse (Ishii & Sadowsky, 2008), coming from a range of sources including humans (Odonkor & Ampofo, 2013), other warm-blooded mammals such as cattle (Friesema et al., 2011) or poultry (Johnson et al., 2007), and environmental sources. A number of research studies have isolated environmental *E. coli* strains from both soil and water, demonstrating through DNA fingerprinting that they are distinct from animal-borne isolates (Byappanahalli et al., 2006; Ishii et al., 2006). Importantly, environmental strains of *E. coli* have been reported to bloom in the environment, resulting in high concentrations in waters in the absence of fecal input or human pathogens (Anderson et al., 2005).

In a recent review paper, Jang et al. (2017) provide, among other, a list of environmental factors which can affect survival and growth of *E. coli*. High freshwater temperatures (e.g., exceeding 18 °C) and availability of nutrients are cited most frequently as the two most important variables involved in the survival and growth of environmental *E. coli* strains (Byappanahalli et al., 2006; Davies et al., 1995; Edge et al., 2013; Faust et al., 1975; Ishii & Sadowsky, 2008; Liu et al., 2006; Winfield & Groisman, 2003). Specific nutrients include carbon, nitrogen, phosphorus, and sulfur (Anderson et al., 2005; Faust et al., 1975; Ishii & Sadowsky, 2008). Other promoting factors suggested in previous research include presence of aquatic flora (Percec-Merien & Lewis, 2013; Whitman et al., 2003), rainfall/droughts (Edge et al., 2013; Whitman et al., 2004), preceding dust storms or bushfires (Chafer, 2007; Power et al., 2005), and presence of macrophytic algae and sand (Ishii & Sadowsky, 2008; Sadowsky & Whitman, 2011). Conversely, the magnitude of, and exposure to, solar radiation is the primary mechanism for *E. coli* decay (Anderson et al., 2005; Lim et al., 2011). This is supported by experiments described in Whitman et al. (2004) who showed that the time of sampling can have a significant impact on *E. coli* counts, with samples taken at 7 a.m. being up to 70% higher than samples taken at 10 a.m. on the same day, due to solar radiation (Whitman et al., 2004).

The comprehensive analysis of the data set from five, substantially different reservoirs, provides an opportunity to identify on a larger scale which specific parameters, suggested in previous research, are correlated consistently with *E. coli* blooms.

2. Materials and Methods

2.1. Data Collection and Analysis

Historical water quality data were obtained from five reservoirs around Australia located in different climatic zones, namely (Figure 1),

1. Grahamstown Dam (GT-New South Wales-Maximum capacity: 182,305 ML)
2. Warragamba Dam (WD-New South Wales-Maximum capacity: 2,310 GL)
3. Hinze Dam (HD-Queensland-Maximum capacity: 310,730 ML)
4. Maroon Dam (MD-Queensland-Maximum capacity: 44,319 ML)
5. Millstream Dam (WA-Western Australia-Maximum capacity: 810 ML).

Weather data from the Australian Bureau of Meteorology and other types of data (e.g., land use and bushfire/dust storm occurrences) from water utilities or other governmental agencies were acquired. The data covered a wide range of frequencies (hourly to yearly), duration (from less than 5 years for Millstream Dam to over 20 for Warragamba Dam), and unit/method of measurement (e.g., colony-forming units or CFU and most probable number or MPN for *E. coli*). An important missing information was the culture temperature applied for the total coliform counts, as environmental isolates might have optimum growth at different temperature to mammalian sources.

The available data were first described by plotting each variable independently to visualize trends. Overlay plots were then created to visualize trends between *E. coli* counts and other available weather and water quality parameters, including key variables identified in the literature and listed in the Background section (where these data existed). Self-organizing maps (Kohonen, 1998) were also used in some instances to

Figure 1. Location of the five reservoirs analyzed.

visually explore correlations between multiple variables. Self-organizing maps are an effective tool for visualizing and exploring the relationships between multiple variables in order to obtain a better understanding of the system (Mannina & Viviani, 2009); this is especially important in complex problems where several variables are potentially involved, as illustrated in Bertone, Stewart, et al. (2016). Results of the data analysis for each bloom event were summarized and compared between blooms in order to understand if any common *E. coli* bloom predictor(s) exist. Bloom events were identified as having peak values of over 10,000 CFU/100 ml or 10,000 MPN/100 ml (Power et al., 2005), and with rapid growth/decline turnovers in the absence of apparent fecal input. For comparison, a number of events with no clear fecal input (e.g., without considerable catchment inflow) but with counts lower than 10,000/100 ml (though still higher than 5,000/100 ml or than the upper detection limit), or that grew/decreased relatively slowly, were included.

2.2. Bayesian Network Model Development

BNs are probabilistic graphical models. The strength of a connection between variables (called “nodes”) is quantified through probability distributions. These probabilities are defined in the Conditional Probability Tables (CPTs) for each node. The CPT of each node is a representation of the associated uncertainty (Marcot et al., 2001). A BN can handle highly uncertain systems, missing data, as well as the integration of empirical data and expert opinion (Uusitalo, 2007). Additionally, the integration of heterogeneous empirical data of different characteristics (e.g., different frequencies, such as for most of the collected data for this study) can be performed. If the model structure is well defined, they can also achieve good accuracy with a small data set; this is the case for this particular application, given the limited number of recorded *E. coli* blooms. In addition, simulations are typically fast compared to other models. BNs have been extensively applied in similar applications (Barton et al., 2012; Bertone, Stewart, et al., 2016; Paix et al., 2019), and they were selected as the most suitable modeling framework for this research project.

The structure of the developed BN (which was exclusively data-driven) was based on the findings of the data analysis; thus, the main potential predictors were used as “parent” nodes (e.g., inputs) for the target node, being the chance of an *E. coli* bloom (i.e., values of over 10,000 CFU/100 ml or 10,000 MPN/100 ml). To populate the CPTs of the BN, historical data for *E. coli*, dam volume, total algal counts (or chl-*a* concentrations if total algal counts were not available), bushfire and dust storm occurrences were reanalyzed. Given that it was established that one, or a combination of, these factors were present during *E. coli* blooms, it was important to assess also how often *E. coli* blooms did not occur, despite the occurrence of one of the trigger events or decreasing dam volume. Thus, for each reservoir, a new BN database entry was added each time any of the parent nodes had a positive value (i.e., occurrence of trigger event or dam volume decreasing). The respective values of the other parent nodes and child node (*E. coli*) were also recorded. The number of event entries for the BN was $n = 48$. The nodes were then discretized based on the recorded values. For total algal counts, four states were introduced as “high,” “medium,” “low,” and “no data”. The thresholds were defined based on chl-*a* concentrations and total algal counts values that yield the optimal (i.e., higher BN accuracy) discretization for populating the CPTs. Similarly, the best thresholds for discretizing the dam volume's decreasing rate were 0.5% per month and 1.2% per month of the full capacity. Such discretization allowed for an optimal population of the CPT, that is, producing the most probability distributions.

The model accuracy was tested using a number of methods. A common method to test BN classifier accuracy is to estimate the receiver operating characteristic curve (ROC) and related Area Under the Curve (AUC) (Bradley, 1997; 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 rate (i.e., $1 - \text{specificity}$) of the model, and the y axis the true positive rate (i.e., sensitivity). The line $y = x$ would represent a model that yields no information. It is commonly agreed that ROC's leading to AUC values over 0.80 are typically associated with good models, while values over 0.90 are associated with excellent models. An AUC of 0.50 would represent the accuracy of a model yielding no information. Calculations were performed in Matlab2018a (The Mathworks Inc.)

The following step consisted of finding the optimal discriminatory threshold (i.e., what is the minimum BN posterior probability value to be linked to a bloom occurrence), in order to maximize the BN models' accuracy. Both the general BN model accuracy (i.e., sum of true positive and true negatives, divided by total count) and the Youden's J statistic (i.e., sensitivity + specificity – 1) were estimated for different discriminatory thresholds. The Youden's J statistic (Youden, 1950) was preferred to the general model accuracy, as the latter would be affected by the fact that the data set only had a few exceedances, that is, it would not have been affected by very low specificity as long as there are several true negatives. By comparison, the Youden's J statistic weighs sensitivity and specificity evenly, assigning the same importance to how the BN model predicts both blooms and low values.

3. Results and Discussion

3.1. Data Analysis

The variables that correlated most strongly with an *E. coli* bloom are shown in supporting information Table 1.

Based on the results of the data analysis work as described in section 2.1, there seems to be no single variable, or combination of variables, that could account for all the blooms analyzed in the reservoirs of this study. However, there is a trend toward blooms occurring during dry periods and thus decreasing reservoir storage volumes. In addition, these blooms occurred predominantly in either spring or summer when the water temperature exceeded 18 °C (for all blooms with water temperature data). This is consistent with the literature (Power et al., 2005), which suggests that water temperature is a key factor in the growth and survival of *E. coli*. In addition, in previous studies *E. coli* that survived in dried soil showed growth upon rehydration (Ishii et al., 2009); thus, when rainfall events occur following such identified prolonged dry periods, these might have triggered growth of *E. coli* present in catchment soils, which were then transported into the reservoirs, but this hypothesis could not be verified in this study as the numbers of *E. coli* in catchment soils were not measured.

Table 1
Confusion Matrix and Accuracy at Threshold 33%, BN Model

Confusion matrix <i>E. coli</i> threshold = 0.33		Measured	
		Bloom	No bloom
Modeled	Bloom	8	7
	No bloom	3	20
Accuracy	Sensitivity	Specificity	Youden's J statistics
78.7%	0.73	0.81	0.53

The frequency of available nutrient data was insufficient to capture any sudden increase in nutrient concentrations due to wind-driven deposition events for example, and thus, any meaningful conclusions about direct correlations between nutrients and *E. coli* bloom formation could not be drawn. However, there appeared to be a trend toward *E. coli* blooms occurring during times of decreasing dam volume, algal blooms or/and after bushfires or dust storms. These latter events seem to provide an increased amount of nutrients within the catchment, which are then transported inside the reservoir following rain events (Ishii et al., 2009).

Noticeably, there was an extremely large dust storm event in September 2009 involving thousands of kilometers of the Australian East Coast (Schmaltz, 2009), and after a few weeks of this event, *E. coli* blooms were recorded at three separate and distant reservoirs on the east coast, namely Warragamba dam, Grahamstown dam and Hinze dam, almost at the same time, following light rainfall. In contrast, algal blooms were also observed during times where *E. coli* blooms were not present, suggesting that a combination of factors may be driving *E. coli* bloom formation; that is, nutrient availability may be necessary but not sufficient on its own for bloom formation.

The findings of the data analysis are in line with the outcomes of previous research, which, as reported in the introduction, consistently pointed at warm temperature and nutrients as key for *E. coli* survival and growth, with rain/drought also playing an important role. There was no clear relationship between high solar radiation and decreasing *E. coli* concentrations: conversely, the blooms mostly occurred during the warmer months of the year, which are also when solar radiation is the strongest. Thus, it seems that other growth-promoting factors had a stronger influence than solar radiation in those instances; other concurrent variables such as high algal counts might have also contributed to decreasing the effect of solar radiation. Importantly, high *E. coli* counts were recorded at multiple water depths, as well as multiple monitoring locations during most of the bloom events.

There appears to be a strong correlation between other coliforms and *E. coli* concentrations in 7 of the 12 blooms. This was expected as the “total coliforms” (of which *E. coli* is a subset) count was much higher than *E. coli* count on those seven occasions, implying that other coliforms also bloomed. This was also in line with previous studies in Australia (Lawrence, 2001). Enterococci also seemed to bloom along with *E. coli* in a number of events, which is also in line with previous studies (Byappanahalli et al., 2012). In three of the other five blooms, sufficient data were not available to draw any conclusions; for the remaining two *E. coli* blooms, the total coliform concentration matched the *E. coli* concentration, suggesting no other coliforms bloomed. This could be due to sampling variability; however, a meaningful explanation is unlikely given the available information. Another limitation as previously mentioned is due to the lack of sufficient catchment data, given that, for instance, naturalized *E. coli* genotypes were previously reported in sediments, soil, and sand (Ishii & Sadowsky, 2008; Sadowsky & Whitman, 2011).

3.2. BN Model

The structure of the data-driven BN model developed is shown in Figure 2. The figure represents a scenario where no evidence is entered in the input nodes. To avoid a BN structure with four parent nodes, which would produce a very large CPT, the three seemingly best “trigger” factors (i.e., occurrence of dust storm, bushfire, and algal bloom) were grouped together. Thus the final BN structure consists of three levels, with two parent nodes and three grandparent nodes (all linked to the same parent). In this way, the probability of an *E. coli* bloom is directly related to (1) the rate of decrease in dam volume and (2) the occurrence of any of the three trigger factors. Essentially, with the occurrence of these conditions, there was historically a higher propensity for a bloom to occur following light rainfall events.

The green nodes represent input variables (i.e., where the user can manually enter evidence), the purple node is the target, child node (i.e., likelihood of *E. coli* bloom), and the yellow node is a calculation node that summarizes the presence of any trigger factor. Based on the data available, some nodes (e.g., dust storm or bushfire) have rather qualitative states, while others (e.g., decreasing dam volume) have quantitative states. The “chl-a or algal count” node, although its states have qualitative names, is quantitative (e.g., an algal count is considered high if it is higher than 10,000 cells per milliliter). The thresholds for the numbers defining each state were selected following training/optimization of the model. Although the outputs are

Figure 2. Final BN structure; green nodes are inputs; purple node is model output.

themselves probabilities, rather than defining the two states for *E. coli* bloom as “yes” and “no,” it was decided to name them “likely” and “unlikely”. This is to emphasize the high uncertainty associated with any trend observed, due to the paucity of data in particular. Hence, even if the model predicted a 100% chance of an *E. coli* bloom this wording ensures that the user understands that this translates to a high probability rather than certainty of a bloom. Also, with the current historical data used to develop the model, the “worst-case scenario,” that is the one leading to the highest likelihood of a bloom, yields only a 44% chance of one, which means that even under ideal conditions, there is still a higher chance of a bloom not occurring. This is most likely due to the state of other poorly characterized influencing factors (e.g., other microorganisms and catchment soil properties). Also, the variable enumeration methods used for *E. coli* may have affected the results, even though the BN model did not factor this in explicitly. Future data collection could expand the training data set, and allow the inclusion of the enumeration method as an extra input node. However, though the counts can indeed be quite different, it was rare for *E. coli* to be detected at high concentrations through one method, and not detected through another method. Hence, given that the BN model is designed to predict the likelihood of bloom occurrence rather than exact counts, merging the data from different enumeration methods would have only a minimum negative impact on the accuracy, compared to the benefits of a larger training set of data.

Figure 3 illustrates the calculated ROC for the BN. The curve is well above the random model ($y = x$), which is confirmed by the resulting high AUC value of 0.8034. Based on this metric alone, this means that the developed BN classifier is very accurate, despite the quality and quantity of the input data. This could be the consequence of a small training data set ($n = 48$), which might tend to amplify either the accuracy or inaccuracy of a model. A much larger data set would be needed to develop a truly accurate and more robust model. Data limitations were clear from the beginning of the model building process; thus, the BN’s goal was to trigger more frequent sampling activities when a higher chance of a bloom is expected; this would eventually lead to the collection of enough input data for model recalibration, or new more informative data (e.g., nutrients in the catchment) allowing it to be deployed for prediction too.

In order to find the optimal discriminatory probability threshold, the Youden’s J statistics as well as the general model accuracy for varying thresholds were calculated (Figure 4). The threshold value represents the critical probability number above which the scenario should be considered associated with a bloom occurrence and below which a bloom would not occur.

Figure 3. ROC for the developed BN.

Interestingly, there seems to be two separate threshold values (i.e., 15% and 33%) that maximize the Youden's J statistics (i.e., respectively, 0.528 and 0.533). However, the model accuracy is considerably higher (78.72% vs. 63.83%) for the 33% threshold. Based on these values, the confusion matrix, or binary contingency table (Powers, 2011), with an overview of the model performance is shown in Table 1.

In light of these accuracy index results, the BN user should use the 33% probability value for the target node *E. coli* as a discriminatory threshold and identify any input scenarios leading to a predicted posterior probability above 33%, as likely to lead to a bloom event in the following weeks. This threshold was estimated based on the features of the collected data for the given reservoirs; hence for different locations, such value should be recalculated based on the properties of the new data.

To the authors' knowledge, this study was the first to fully analyze routinely collected weather and water quality data to better understand *E. coli* blooms in reservoirs. If more data (especially nutrients) were collected when the current BN triggers a warning, the model could be refined in the future to improve its predictive capacity. However, since a number of additional influencing factors, difficult (or expensive) to monitor routinely, such as the presence and diversity of other microorganisms (Jang et al., 2017) can also affect the growth or survival of *E. coli* and thus the occurrence of a bloom, an inherent, considerable degree of uncertainty would remain in the model. Nevertheless, the degree of uncertainty is clearly represented and

Figure 4. Youden's J statistics and model accuracy (%), BN model.

quantified in a BN approach through the CPTs (Chen & Pollino, 2012; Uusitalo, 2007), unlike other types of models.

4. Conclusions

Historical data for five Australian drinking water reservoirs, routinely collected by third parties, were analyzed to potentially identify factors which could increase the chance of *E. coli* to bloom in such water bodies. In addition to water quality data, weather as well as catchment data (e.g., bushfire, land use, and dust storm occurrences) were collated where available and analyzed.

Data were often insufficient both temporally (e.g., low frequency or missing) and spatially, but findings still seem to point quite consistently to a combination of potential bloom predictors. Most of the blooms occurred after minor rainfall events following prolonged dry periods (with decreasing dam volume) and with water temperatures higher than 18 °C. The occurrence of a “trigger” event, such as a dust storm or a bushfire, seems to be necessary to provide a sufficient amount of nutrients in the catchment. However, the latter are often not well characterized or monitored, especially in catchment soils, thus limiting the ability to test such hypotheses. Further investigative research needs to be conducted to understand environmental *E. coli* blooms more reliably. Such efforts have to include better spatial and temporal distribution of these identified factors and other potentially important factors such as reservoir and catchment nutrients and organic matter. Most of the influential parameters, such as prolonged dry periods and bushfires, are expected to be exacerbated by climate change, thus posing a risk of an increase in bloom events, and also an opportunity for more targeted monitoring. Better spatial and temporal resolution of future *E. coli* monitoring can also prove to be useful for capturing more bloom-like events which might be missed at current monitoring frequencies.

Based on the data analysis findings, a BN model was conceptualized and trained using historical data. Despite insufficient high quality data for certain locations/periods, the model yielded a relatively high accuracy based on analyzed historical events. Given the data limitations detailed in this paper, the current model should only be used by water managers to trigger more intensive sampling programs focused on the identified predictors, to collect more information and to improve the understanding and modeling of future environmental *E. coli* bloom events.

Acknowledgments

The authors are thankful to Water Research Australia for providing financial support to this project, which was part of a larger collaborative research project (WaterRA 1101). Data used for this research project are confidential but could be obtained by contacting the relevant water utilities.

References

- Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software*, 26, 1376–1388.
- Anderson, K. L., Whitlock, J. E., & Harwood, V. J. (2005). Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. *Applied and Environmental Microbiology*, 71(6), 3041–3048. <https://doi.org/10.1128/AEM.71.6.3041-3048.2005>
- Barton, D. N., Kuikka, S., Varis, O., Uusitalo, L., Henriksen, H. J., Borsuk, M., et al. (2012). Bayesian networks in environmental and resource management. *Integrated Environmental Assessment and Management*, 8(3), 418–429. <https://doi.org/10.1002/ieam.1327>
- Bertone, E., Sahin, O., Richards, R., & 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.
- Bertone, E., Stewart, R. A., Zhang, H., & O'Halloran, K. (2016). Hybrid water treatment cost prediction model for raw water intake optimization. *Environmental Modelling & Software*, 75, 230–242.
- Blyton, M. D. J., & Gordon, D. M. (2017). Genetic Attributes of *E. coli* isolates from chlorinated drinking water. *PLoS ONE*, 12, e0169445.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145–1159.
- Byappanahalli, M. N., Nevers, M. B., Korajkic, A., Staley, Z. R., & Harwood, V. J. (2012). Enterococci in the environment. *Microbiology and Molecular Biology Reviews*, 76(4), 685–706. <https://doi.org/10.1128/MMBR.00023-12>
- Byappanahalli, M. N., Whitman, R. L., Shively, D. A., Sadowsky, M. J., & Ishii, S. (2006). Population structure, persistence, and seasonality of autochthonous *Escherichia coli* in temperate, coastal forest soil from a Great Lakes watershed. *Environmental Microbiology*, 8(3), 504–513. <https://doi.org/10.1111/j.1462-2920.2005.00916.x>
- Chaffer, C. J. (2007). Wildfire, catchment health and water quality: A review of knowledge derived from research undertaken in Sydney's Water Supply Catchments 2002–2007. Sydney Catchment Authority.
- Chen, S., & Pollino, C. (2012). Good practice in Bayesian network modelling. *Environmental Modelling & Software*, 37, 134–145.
- Cumberland, S., Bridgeman, J., Baker, A., Sterling, M., & Ward, D. (2012). Fluorescence spectroscopy as a tool for determining microbial quality in potable water applications. *Environmental Technology*, 33(4-6), 687–693. <https://doi.org/10.1080/09593330.2011.588401>
- Davies, C. M., Long, J. A., Donald, M., & Ashbolt, N. J. (1995). Survival of fecal microorganisms in marine and freshwater sediments. *Applied and Environmental Microbiology*, 61(5), 1888–1896.
- Dwivedi, D., Mohanty, B. P., & Lesikar, B. J. (2013). Estimating *Escherichia coli* loads in streams based on various physical, chemical, and biological factors. *Water Resources Research*, 49, 2896–2906. <https://doi.org/10.1002/wrcr.20265>
- Edge, T. A., Khan, I. U. H., Bouchard, R., Guo, J., Hill, S., Locas, A., et al. (2013). Occurrence of waterborne pathogens and *Escherichia coli* at offshore drinking water intakes in Lake Ontario. *Applied and Environmental Microbiology*, 79(19), 5799–5813. <https://doi.org/10.1128/AEM.00870-13>

- Faust, M. A., Aotaky, A., & Hargadon, M. (1975). Effect of physical parameters on the in situ survival of *Escherichia coli* MC-6 in an estuarine environment. *Applied Microbiology*, 30(5), 800–806.
- Friesema, I. H. M., Van De Kastele, J., De Jager, C. M., Heuvelink, A. E., & Van Pelt, W. (2011). Geographical association between livestock density and human Shiga toxin-producing *Escherichia coli* O157 infections. *Epidemiology and Infection*, 139(7), 1081–1087. <https://doi.org/10.1017/S0950268810002050>
- Hondzo, M., & Al-Homoud, A. (2007). Model development and verification for mass transport to *Escherichia coli* cells in a turbulent flow. *Water Resources Research*, 43, W08413. <https://doi.org/10.1029/2006WR005613>
- Ishii, S., Ksoll, W. B., Hicks, R. E., & Sadowsky, M. J. (2006). Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior watersheds. *Applied and Environmental Microbiology*, 72(1), 612–621. <https://doi.org/10.1128/AEM.72.1.612-621.2006>
- Ishii, S., & Sadowsky, M. J. (2008). *Escherichia coli* in the environment: Implications for water quality and human health. *Microbes and Environments*, 23(2), 101–108. <https://doi.org/10.1264/jsme2.23.101>
- Ishii, S., Yan, T., Vu, H., Hansen, D. L., Hicks, R. E., & Sadowsky, M. J. (2009). Factors controlling long-term survival and growth of naturalized *Escherichia coli* populations in temperate field soils. *Microbes and Environments*, 25(1), 8–14.
- Jang, J., Hur, H. G., Sadowsky, M. J., Byappanahalli, M., Yan, T., & Ishii, S. (2017). Environmental *Escherichia coli*: Ecology and public health implications—A review. *Journal of Applied Microbiology*, 123, 570–581.
- Johnson, J. R., Sannes, M. R., Croy, C., Johnston, B., Clabots, C., Kuskowski, M. A., et al. (2007). Antimicrobial drug-resistant *Escherichia coli* from humans and poultry products, Minnesota and Wisconsin, 2002–2004. *Emerging Infectious Diseases*, 13(6), 838–846. <https://doi.org/10.3201/eid1306.061576>
- Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21, 1–6.
- Lawrence, I., (2001). Bacteria Bloom in Lake Burley Griffin probably not caused by sewage. May 04 2001, CRCFE, Canberra - Media Release. Cooperative Research Centre for Freshwater Ecology.
- Lim, N., Munday, C. I., Allison, G. E., O’Loingsigh, T., De Deckker, P., & Tapper, N. J. (2011). Microbiological and meteorological analysis of two Australian dust storms in April 2009. *Science of the Total Environment*, 412, 223–231.
- Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A better measure than accuracy in comparing learning algorithms. In *Conference of the Canadian society for computational studies of intelligence*, (pp. 329–341). Springer, Berlin, Heidelberg.
- Liu, L., Phanikumar, M. S., Molloy, S. L., Whitman, R. L., Shively, D. A., Nevers, M. B., et al. (2006). Modeling the transport and inactivation of *E. coli* and enterococci in the near-shore region of Lake Michigan. *Environmental Science & Technology*, 40(16), 5022–5028. <https://doi.org/10.1021/es060438k>
- Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., & Konstantinidis, K. T. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences*, 108, 7200–7205.
- Mannina, G., & Viviani, G. (2009). Separate and combined sewer systems: A long term modelling approach. *Water Science and Technology*, 60(3), 555–565. <https://doi.org/10.2166/wst.2009.376>
- Marcot, B. G., Holthausen, R. S., Raphael, M. G., Rowland, M. M., & Wisdom, M. J. (2001). Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management*, 153, 29–42.
- NHRMC, (2011). Australian Drinking Water Guidelines 6. In: Council, N.H.a.M.R. (Ed.).
- Odonkor, S. T., & Ampofo, J. K. (2013). *Escherichia coli* as an indicator of bacteriological quality of water: An overview. *Microbiology Research*, 4, e2–e2.
- Paix, B., Ezzedine, J. A., & Jacquet, S. (2019). Diversity, dynamics, and distribution of *Bdellovibrio* and like organisms in Perialpine Lakes. *Applied and Environmental Microbiology*, 85, e02494-02418.
- Percec-Merien, A. M., & Lewis, G. D. (2013). Naturalized *Escherichia coli* from New Zealand wetland and stream environments. *FEMS Microbiology Ecology*, 83(2), 494–503. <https://doi.org/10.1111/1574-6941.12010>
- Pinto, D., Almeida, V., Almeida Santos, M., & Chambel, L. (2011). Resuscitation of *Escherichia coli* VBNC cells depends on a variety of environmental or chemical stimuli. *Journal of Applied Microbiology*, 110(6), 1601–1611. <https://doi.org/10.1111/j.1365-2672.2011.05016.x>
- Power, M. L., Littlefield-Wyer, J., Gordon, D. M., Veal, D. A., & Slade, M. B. (2005). Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environmental Microbiology*, 7(5), 631–640. <https://doi.org/10.1111/j.1462-2920.2005.00729.x>
- Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2, 37–63.
- Sadowsky, M. J., & Whitman, R. L. (2011). *The fecal bacteria*. Washington, DC: ASM Press.
- Schmaltz, J., (2009). Dust over eastern Australia. <http://earthobservatory.nasa.gov/NaturalHazards/view.php?id=40274>.
- Sinclair, M., (2019). Discussion paper—Identification and management of environmental *E. coli* blooms. WaterRA Project #1101, p. 22.
- Tyagi, V., Chopra, A., Kazmi, A., & Kumar, A. (2006). Alternative microbial indicators of faecal pollution: Current perspective. *Journal of Environmental Health Science & Engineering*, 3, 205–216.
- Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, 203, 312–318.
- Wei, C., & Zhao, X. (2018). Induction of viable but nonculturable *Escherichia coli* O157: H7 by low temperature and its resuscitation. *Frontiers in Microbiology*, 9. <https://doi.org/10.3389/fmicb.2018.02728>
- Whitman, R. L., Nevers, M. B., Korinek, G. C., & Byappanahalli, M. N. (2004). Solar and temporal effects on *Escherichia coli* concentration at a Lake Michigan swimming beach. *Applied and Environmental Microbiology*, 70(7), 4276–4285. <https://doi.org/10.1128/AEM.70.7.4276-4285.2004>
- Whitman, R. L., Shively, D. A., Pawlik, H., Nevers, M. B., & Byappanahalli, M. N. (2003). Occurrence of *Escherichia coli* and enterococci in *Cladophora* (Chlorophyta) in nearshore water and beach sand of Lake Michigan. *Applied and Environmental Microbiology*, 69, 4714–4719.
- Winfield, M. D., & Groisman, E. A. (2003). Role of nonhost environments in the lifestyles of salmonella and *Escherichia coli*. *Applied and Environmental Microbiology*, 69(7), 3687–3694. <https://doi.org/10.1128/aem.69.7.3687-3694.2003>
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::aid-cnrcr2820030106>3.0.co;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cnrcr2820030106>3.0.co;2-3)