Jun 16th, 10:40 AM - 12:00 PM

Accounting for Groups of Animals in QMRA of Recreational Waters

Richard W. Muirhead
Land and Environment, AgResearch Ltd, Invermay Research Centre, richard.muirhead@agresearch.co.nz

Vanessa M. Cave
Bioinformatics, Mathematics & Statistics, AgResearch Ltd, Ruakura Research Centre, vanessa.cave@agresearch.co.nz

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Environmental Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Accounting for Groups of Animals in QMRA of Recreational Waters

Richard W. Muirhead[a], Vanessa M. Cave[b]

[a] Land and Environment, AgResearch Ltd, Invermay Research Centre, Private Bag 50034, Mosgiel 9053, New Zealand. richard.muirhead@agresearch.co.nz
[b] Bioinformatics, Mathematics & Statistics, AgResearch Ltd, Ruakura Research Centre, Private Bag 3123, Hamilton 3240, New Zealand. vanessa.cave@agresearch.co.nz

Abstract: The development of QMRA for recreational waters often involves assessing the inputs from multiple sources, including human sewage, agricultural sources and wild animals. The concentrations of microorganisms in faeces from individual animals are highly variable leading to a broad spectrum of risk. When modelling the load of faecal microbes from a group of animals, based on data from individuals, the group distribution must be correctly accounted for. A further complication to QMRA is that pathogen concentrations in faeces have a high proportion of non-detects making them difficult to model using standard mathematical distributions. Moreover, when only limited concentration data are available, determining the appropriate mathematical distribution is difficult. These issues are explored in this study using the scenarios of *E. coli* and *Campylobacter* depositions into a stream from wild ducks. Using Monte Carlo methods, different models to describe the load of microorganisms deposited by ducks into a stream were compared. In addition, parametric and non-parametric methods were used to generate microbial concentrations in an individual ducks' faeces. Our study demonstrated that calculating the load of faecal microorganisms from groups of animals for QMRA requires models that accurately reflect the loads of the whole group. Non-parametric methods can be used for pathogen concentrations that do not appear to fit a standard mathematical distribution or when determining the appropriate distribution is difficult.

Key Words: QMRA; modelling; Monte Carlo; ducks; non-point pollution.

1 INTRODUCTION

Faecal microbial contamination of surface waters is of growing concern around the world (Boehm et al., 2009; Monaghan et al., 2008; Oliver et al., 2010; Till et al., 2008). This concern is driven by the desire to manage the health risk associated with contaminated water (Boehm et al., 2009; Soller et al., 2006). Quantitative microbial risk analysis (QMRA) has proven a useful technique for quantifying water quality risks (Soller et al., 2010). The QMRA approach has the advantage of being applicable to both the risk assessment associated with a single source of contamination (McBride et al., 2013) and for comparing risks from a range of different sources (Soller et al., 2010). This is important for managing recreational water quality which is a catchment scale issue that is usually impacted on by a number of point and non-point sources of pollution (Ferguson et al., 2007).

In most catchments a source of non-point contamination to streams is direct inputs from animals such as water fowl, wild or farmed animals (Davis-Colley et al., 2004; Whither et al., 2005; Zhu et al., 2011). Therefore, estimating the numbers of microorganisms shed by animals is an important step for calculating the impact of animal faecal deposition on water quality. The accuracy of these calculations is important for predicting the effect of different sources of microorganisms in QMRA. When calculating the load of microorganisms discharged from point source discharges, such as sewage, the total load is typically calculated by multiplying the microbe concentration in the sewage by the volume of sewage produced (Soller et al., 2003; Ferguson et al., 2007). Similarly, the load produced by groups of animals has been calculated by multiplying the microbe concentration in the

faeces by the weight of faeces produced and the number of animals (Zhu et al., 2001; McBride and Chapra, 2011; Muirhead et al., 2011). However, we are concerned about using microbial concentration data from individual animals to characterise the loads generated by groups of animals, particularly as the size of the group increases. To address this issue we compare the results generated by the equation used in various published models (Zhu et al., 2001; McBride and Chapra, 2011; Muirhead et al., 2011) with an alternative equation based on statistical theory.

A separate complication in QMRA arises from the scarcity of data and/or high proportion of 'non-detects' in the faecal concentration datasets (Atwill et al., 2012; Hutchison et al., 2004; Moriarty et al., 2011). These pathogen datasets are typically strongly right-skewed and expert judgment is required in selecting the appropriate mathematical distribution to model the data. This judgement call is further complicated as "traditional 'goodness-of-fit' tests have difficulty rejecting any right-skewed skewed distribution" (McBride et al., 2013). To address the issues of having to select a distribution, we propose an alternative, non-parametric approach of sampling with replacement from the measured values. The multiple iterations within the Monte Carlo simulation will provide a smoothing effect on the measured data. Furthermore, using only measured values constrains the simulations to "real" values without the potential influence of extreme values generated by an unconstrained right-skewed distribution.

In this paper we validate a method to calculate the daily load of faecal microbes from different sized groups of animals using Monte Carlo simulations. Two different mathematical equations to calculate daily loads are compared along with two different methods (parametric and non-parametric) for sampling the concentrations of microbes in the faeces. The differences between these simulated distributions are explainable following standard statistical theory. The analysis is conducted using the daily loading of *E. coli* and *Campylobacter* deposited into a stream from ducks, but will be applicable to any combination of faecal microbe and animal source.

2 METHODS

The daily loads of microorganisms deposited into a stream by a group of ducks (L: # *E. coli* day^{-1}) were calculated in Monte Carlo simulations using 2 different equations. Equation (1) is referred to as the 'multiplication' equation (McBride and Chapra, 2011; Muirhead et al., 2011; Zhu et al., 2011) and Equation (2) as the 'sum' equation (based on the statistical theory in the appendix).

$$L = \alpha\ C\ U\ Z \quad (1)$$

$$L = \sum_{n=1}^{Z} \alpha_n C_n U_n \quad (2)$$

where α is the proportion of a duck's faeces deposited directly in the stream, C is the concentration of microbes in a duck's faeces (# g^{-1} wet weight), U is the weight of faeces produced by a duck (g wet weight day^{-1}) and Z is the number of ducks in the group.

For the simulations, the α and U values were sampled parametrically (i.e. from assumed distributions) and the C values were sampled in two ways, either parametrically from a fitted distribution or non-parametrically from the observed sample of concentrations. For the multiplicative equation (1), a single value of each parameter was sampled and their product was multiplied by Z, the number of ducks. For the sum equation (2), Z samples of each parameter were taken from their distributions and their product summed. So the multiplicative equation takes a single duck at random and multiplies its load by Z, while the sum equation adds the loads of Z ducks selected at random.

The proportion of faeces deposited in the stream (α) was modelled as a triangular distribution with a minimum, most likely and maximum proportion of 0.1, 0.35 and 0.6, respectively (Muirhead et al., 2011). The weight of faeces produced by a duck per day (U) was also modelled as a triangular distribution with a minimum, most likely and maximum weights of 100, 336 and 400, respectively (Muirhead et al., 2011).

The concentrations of microbes in the faeces (C) were sampled using 2 different approaches (a parametric and non-parametric method) for both *E. coli* and *Campylobacter*. The concentrations were

based on the data from 80 samples of duck faeces (Moriarty et al., 2011). *E. coli* was detected in all duck faecal samples and the concentrations were log-normally distributed with a $\log_{10}$-mean of 5.5 g^{-1} and a standard deviation of 1.5 (Figure 1). The *Campylobacter* concentrations were highly skewed due to a high proportion (70 %) of samples falling below the detection limits of 4 g^{-1} (Figure 1). We fitted an exponential distribution for the *Campylobacter* concentrations (Close et al., 2008) with β = 0.017 (Figure 1). For the parametric method, faecal microbe concentrations were sampled from their respective distributions for *E. coli* and *Campylobacter*. The non-parametric method was conducted by sampling, with replacement, from the 80 observed concentrations in the duck faeces (note, observations less than the detection limit were inputted as zero).

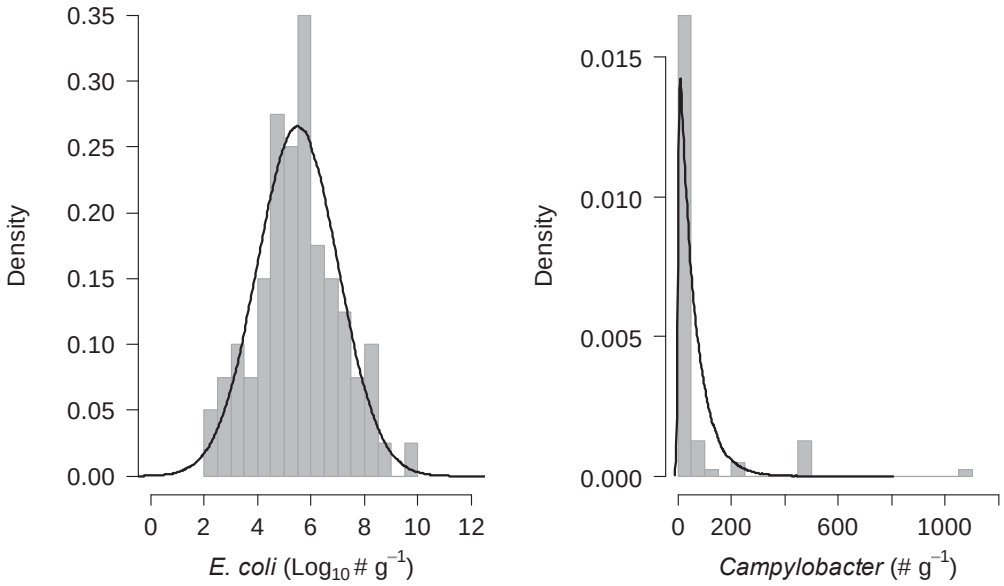

Figure 1: Histograms of the *E. coli* and *Campylobacter* concentrations measured in the duck faeces (bars) and the parametric distributions used in the Monte Carlo simulations (lines). Note that the *E. coli* graph uses a log scale on the X-axis as the data is assumed log normally distributed.

The Monte Carlo simulations were run with 100,000 iterations using the software R (R Core Team, 2013). For each microbe, Monte Carlo simulations were run as a 2x2 matrix using equations (1) and (2) with both parametric and non-parametric methods for sampling of the microbe concentrations in faeces (*C*). The simulations were repeated with the number of ducks in each group (*Z*) set at 1, 10, 100, and 1000. Note that the sum equation (2) employed a subroutine of *Z* iterations per iteration step to calculate the daily load of faecal microbes.

3 RESULTS AND DISCUSSION

3.1 *E. coli* simulations

When there was only one duck in the group, Z=1, the four different models produced essentially the same distribution of *E. coli* loads per day (Figure 2). There was little difference between the parametric and non-parametric sampling methods as *Z* was increased (Figure 2). This observation is due to the good fit of the raw data to the log-normal distribution (Figure 1). When the multiplication equation was employed the spread remained constant on the $\log_{10}$ scale as group size, Z, increased (Figure 2), meaning on the linear scale the data spread is increasing (Table 1). Furthermore, for a given Z, the median on the $\log_{10}$ scale is lower using the multiplication equation (Figure 2). For both the sum and multiplication equations, the mean simulated daily load increases by approximately 10 fold as Z increases by a factor of 10 (Table 1). Using the sum equation the variance also increases

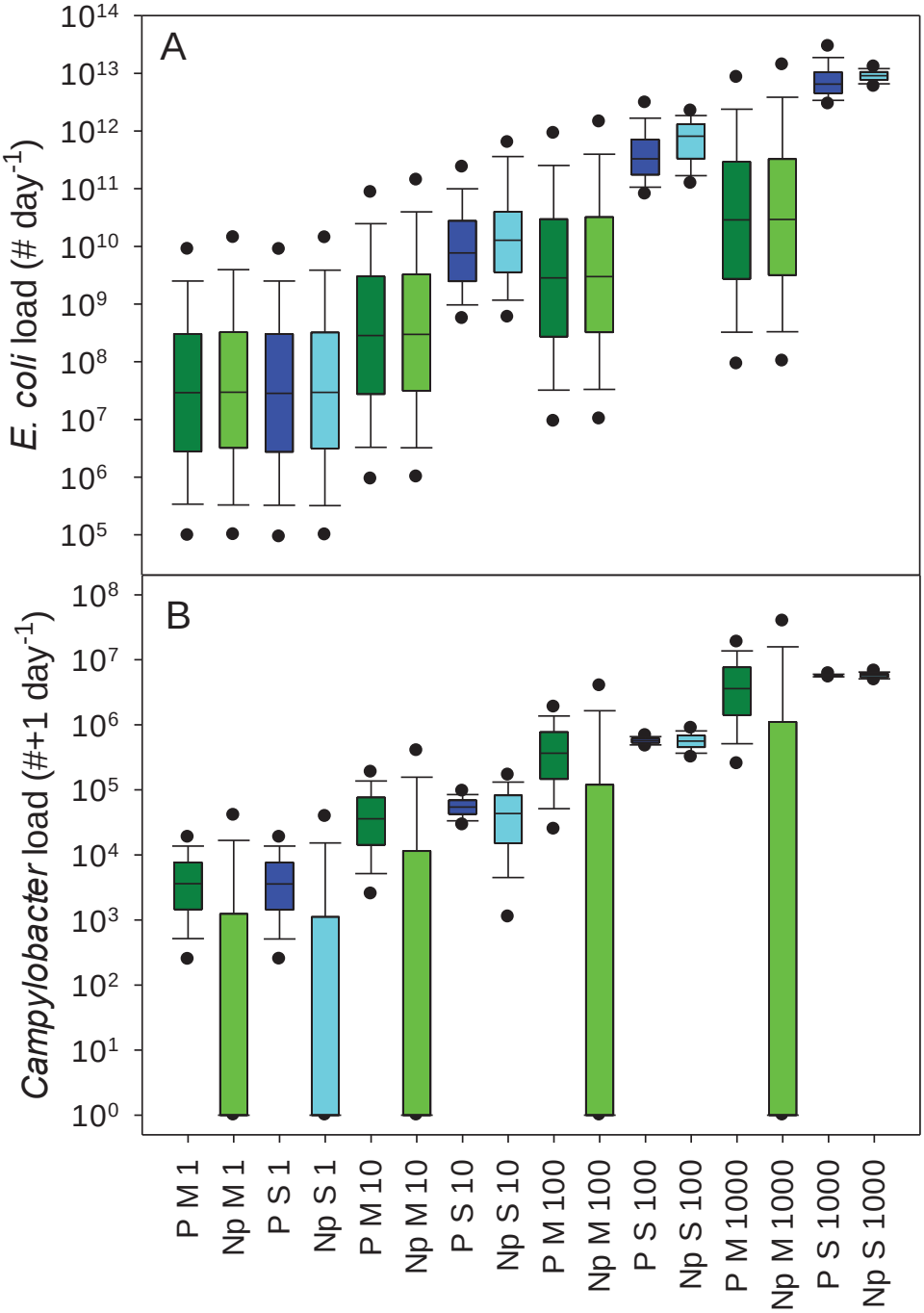

Figure 2: Modelled distributions of the daily load of (A) *E. coli* and (B) *Campylobacter* deposited in a stream by groups of ducks using the 4 different calculation methods. The notations M and S refer to the use of the multiplication and sum equations, respectively. The use of the parametric and non-parametric methods for sampling of the microbe concentrations (*C*) in the duck faeces are denoted by P and Np, respectively. The number refers to the number of animals in each group (*Z*). The horizontal line in the box represents the median value, the boxes span the interquartile ranges, the whiskers span the 10th to 90th percentiles and the dots represent the 5th and 95th percentiles. Note that to avoid problems with *Campylobacter* counts of zero on a log scale, a value of 1 was added to all *Campylobacter* loads before analysis and graphing.

by approximately 10 fold as Z increases by a factor of 10 (Table 1). However, using the multiplication equation, the variance increases by a greater amount: approximately 100 fold as Z increases by a factor of 10. These finding are to be expected given the underlying statistical theory outlined in the appendix.

Table 1: Mean and variance of the simulated daily load of *E. coli* produced by groups of ducks using the 4 different calculation methods. The parametric and non-parametric methods are denoted by P and Np, respectively.

	Mean							
	Natural Scale				Log_{10} Scale			
	Multiplication		Sum		Multiplication		Sum	
Z	P	Np	P	Np	P	Np	P	Np
1	1.3×10^{10}	9.3×10^{9}	9.1×10^{9}	9.1×10^{9}	7.5	7.5	7.5	7.5
10	1.1×10^{11}	9.5×10^{10}	1.1×10^{11}	9.2×10^{10}	8.5	8.5	10.0	10.1
100	9.2×10^{11}	9.4×10^{11}	1.2×10^{12}	9.2×10^{11}	9.5	9.5	11.6	11.8
1000	9.1×10^{12}	9.3×10^{12}	1.2×10^{13}	9.2×10^{12}	10.5	10.5	12.9	13.0

	Variance							
	Natural Scale				Log_{10} Scale			
	Multiplication		Sum		Multiplication		Sum	
Z	P	Np	P	Np	P	Np	P	Np
1	1.2×10^{24}	4.7×10^{21}	7.3×10^{22}	4.5×10^{21}	2.28	2.40	2.29	2.39
10	2.1×10^{25}	4.8×10^{23}	7.7×10^{24}	4.5×10^{22}	2.28	2.40	0.65	0.75
100	6.0×10^{26}	4.7×10^{25}	9.9×10^{25}	4.6×10^{23}	2.29	2.40	0.25	0.16
1000	6.7×10^{28}	4.7×10^{27}	1.9×10^{27}	4.6×10^{24}	2.28	2.39	0.10	0.01

3.2 *Campylobacter* simulations

Analogous differences between the sum and multiplication methods were observed in the simulated *Campylobacter* loads (Figure 2, Table 2). However, unlike the *E. coli* simulations there were noticeable differences between the parametric and non-parametric sampling methods (Figure 2) due to both the sparsity of observed data and the discrepancies between the observed data and theoretical distribution (Figure 1). These differences were greater under the multiplicative equation than the sum equation. The use of the non-parametric method of sampling the concentration of *Campylobacter* in the faeces resulted in a number of daily loads of zero due to selecting a sample where *Campylobacter* was not detected (Figure 2). Using the multiplication equation, this faecal concentration of zero would then be applied to all ducks in the group, which is not sensible. For groups of ducks the multiplication equation produced higher 95^{th} percentile values for the *Campylobacter* loads than the sum equation (Figure 2). This observation for the *Campylobacter* loads is in contrast to that observed for the *E. coli* loads and is due to the highly skewed distribution of the *Campylobacter* concentrations (Figure 1).

Table 2: Mean and variance of the simulated daily load of *Campylobacter* produced by groups of ducks using the 4 different calculation methods. The parametric and non-parametric sampling methods are denoted by P and Np, respectively. Due to the presence of 0 loads, 1 was added before transforming to the $\log_{10}$ scale.

Mean

	Natural Scale				Log_{10}(#+1) Scale			
	Multiplication		Sum		Multiplication		Sum	
Z	**P**	**Np**	**P**	**Np**	**P**	**Np**	**P**	**Np**
1	5.8×10^{3}	5.9×10^{3}	5.7×10^{3}	5.7×10^{3}	3.5	1.2	3.5	1.2
10	5.8×10^{4}	5.8×10^{4}	5.7×10^{4}	5.8×10^{4}	4.5	1.4	4.7	4.4
100	5.8×10^{5}	5.8×10^{5}	5.7×10^{5}	5.8×10^{5}	5.5	1.8	5.8	5.7
1000	5.8×10^{6}	5.7×10^{6}	5.7×10^{6}	5.8×10^{6}	6.5	2.0	6.8	6.8

Variance

	Natural Scale				Log_{10}(#+1) Scale			
	Multiplication		Sum		Multiplication		Sum	
Z	**P**	**Np**	**P**	**Np**	**P**	**Np**	**P**	**Np**
1	4.3×10^{7}	3.1×10^{8}	4.3×10^{7}	3.0×10^{8}	0.34	3.28	0.34	3.24
10	4.3×10^{9}	3.1×10^{10}	4.2×10^{8}	3.0×10^{9}	0.34	5.07	0.02	0.86
100	4.2×10^{11}	3.0×10^{12}	4.3×10^{9}	3.0×10^{10}	0.35	7.36	0.00	0.02
1000	4.3×10^{13}	3.0×10^{14}	4.2×10^{10}	3.0×10^{11}	0.34	9.96	0.00	0.00

3.3 General discussion

The typical approach to assigning numeric values for samples where no microbe was detected is to use a value of half the detection limit. In many studies this applies to only a small proportion of the samples. In this current study, 70% of the *Campylobacter* samples were below the detection limits. Therefore, it was assumed that the measured values represented only the upper tail of the true concentration distribution. Assigning a value of half the detection limit to 70% of the samples is likely to overestimate the true concentration for most of these samples. In QMRA we are typically most interested in the upper percentiles of the distributions, as these contribute to most of the risk (Soller et al., 2010). In this analysis the *Campylobacter* concentrations that were less than the detection limits were assigned a value of zero. Replacing the zero's, with a value of 2 (half the detection limit), only increased the mean concentration from 59 to 61 *Campylobacter* g^{-1} and did not affect the upper percentiles of the modelled distributions (data not shown). Analysis of alternative methods for accounting for sample results below detection limits may be a fruitful area for further study.

This study highlights the fact that simulating the effect of a group of animals cannot be done by selected one at random and multiplying its effect by the number of animals; doing so leads to results with incorrectly high variances (*Z* times too large) which leads to a distribution for the load that is mostly too low and is more highly skewed (see appendix). Note that the arithmetic mean of the distribution derived from the multiplicative model is that same as that from the sum model but all its percentiles are too low. Note also the extreme skew on the distributions of the load, with means being around the 95^{th} percentiles of the distributions. This is of concern, for in QMRA we are typically most interested in the upper percentiles of the distributions, as these contribute to most of the risk (Soller et al., 2010).

The parametric method depends on the validity of the assumed distribution. The non-parametric approach offers a viable alternative to the parametric approach when the underlying distribution is unknown. However, the non-parametric method relies on the availability of a representative dataset to resample from, one which is sufficiently large enough to represent the true distribution well.

4 CONCLUSIONS

The sum equation will generate the most accurate results when using Monte Carlo simulations to calculate the loads of faecal microbes from groups of animals. If there is uncertainty of appropriate fit of a distribution to the raw data, then the non-parametric method can be used if the raw dataset is sufficiently robust.

ACKNOWLEDGMENTS

This work was funded by the Ministry for Business, Innovation and Employment (MBIE) contract C10x1006. Thanks to Elaine Moriarty for kindly providing the raw data from her publication, Moriarty et al., (2011). We are also grateful to Neil Cox and Jen Robson for their insightful comments on the draft manuscript and feedback from the 3 anonymous reviewers.

APPENDIX: UNDERLYING STATISTICAL THEORY

Result 1: Sum of n i.i.d. random variables

Let $X_1, X_2, \ldots, X_n$ be a set of n independent and identically distributed (i.i.d.) random variables with expectation $E[X_i] = \mu$ and variance $V[X_i] = \sigma^2$. Their sum,

$$S_n = \sum_{i=1}^{n} X_i$$

has expectation $E[S_n] = n\mu$ and variance $V[X_i] = n\sigma^2$

Result 2: Multiplying a random variable by a constant value

Let X_i be a random variable with expectation $E[X_i] = \mu$ and variance $V[X_i] = \sigma^2$, and let c denote a constant. Then $E[cX_i] = c\mu$ and $V[cX_i]=c^2\sigma^2$

Hence result 2 shows that if the constant value is the number of animals (i.e. c=n) then the variance is multiplied by n^2 as opposed to n from result 1.

REFERENCES

Atwill, E.R., Li, X., Grace, D., Gannon, V., 2012. Zoonotic waterborne pathogen loads in livestock. In: (Eds.), Dufour, A., Bartram, J., Bos, R., Gannon. V., Animal Waste, Water Quality and Human Health, IWA Publishing, London, UK. pp. 73-114.

Boehm, A.B., Ashbolt, N.J., Colford J.M.Jr., Dunbar, L.E., Fleming, L.E., Gold, M.A., Hansel, J.A., Hunter, P.R., Ichida, A.M., McGee, C.D., Soller, J.A., Weisberg, S.B., 2009. A sea change ahead for recreational water quality criteria. Journal of Water and Health 7 (1), 9-20.

Close, M., Dann, R., Ball, A., Pirie, R., Savill, M., Smith, Z., 2008. Microbial groundwater quality and its health implications for a border-strip irrigated dairy farm catchment, South Island, New Zealand. Journal of Water and Health 6 (1), 83-98.

Davies-Colley, R.J., Nagels, J.W., Smith, R.A., Young, R.G., Phillips, C.J., 2004. Water quality impact of a dairy cow herd crossing a stream. New Zealand Journal of Marine and Freshwater Research 38, 569-576.

Ferguson, C.M., Croke, B.F.W., Beatson, P.J., Ashbolt, N.J., Deere D.A., 2007. Development of a process-based model to predict pathogen budgets for the Sydney drinking water catchment. Journal of Water and Health 5 (2), 187-208.

Hutchison, M.L., Walters, L.D., Avery, S.M., Synge, B.A., Moore, A., 2004. Levels of zoonotic agents in British livestock manures. Letters in Applied Microbiology 39, 207-214.

McBride, G.B., Chapra, S.C., 2011. New hydroepidemiological models of indicator organisms and zoonotic pathogens in agricultural watersheds. Ecological Modelling 222, 2093-2102.

McBride, G.B., Stott, R., Miller, W., Bambic, D., Wuertz, S., 2013. Discharge-based QMRA for estimation of public health risks from exposure to stormwater-borne pathogens in recreational waters in the United States. Water Research 47, 5282-5297.

Monaghan, R.M., de Klein, C.A.M., Muirhead, R.W., 2008. Prioritisation of farm scale remediation efforts for reducing losses of nutrients and faecal indicator organisms to waterways: a case study of New Zealand dairy farming. Journal of Environmental Management 87 (4), 609-622.

Moriarty, E.M., Karki, N., Mackenzie, M., Sinton, L.W., Wood, D.R., Gilpin, B.J., 2011. Faecal indicators and pathogens in selected New Zealand waterfowl. New Zealand Journal of Marine and Freshwater Research 45 (4), 679-688.

Muirhead, R.W., Elliott, A.H., Monaghan, R.M., 2011. A model framework to assess the effect of dairy farms and wild fowl on microbial water quality during base-flow conditions. Water Research 45, 2863-2874.

Oliver, D.M., Page, T., Hogson, C.J., Heathwaite, A.L., Chadwick, D.R., Fish, R.D., Winter, M., 2010. Development and testing of a risk index framework to determine field-scale critical source areas of faecal bacteria in grassland. Environmental Modelling & Software 25, 503-512.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (Last accessed 04.03.14).

Soller, J.A., Olivieri, A.W., Crook, J., Cooper, R.C., Tchobanoglous, G., Parkin, R.T., Spear, R.C., Eisenberg, J.N.S., 2003. Risk-based approach to evaluate the public health benefit of additional wastewater treatment. Environmental Science & Technology 37 (9), 1882-1891.

Soller, J.A., Eisenberg, J.N.S., DeGeorge, J.F., Cooper, R.C., Tchobanoglous, G., Olivieri, A.W., 2006. A public health evaluation of recreational water impairment. Journal of Water and Health 4 (1), 1-19.

Soller, J.A., Schoen, M.E., Bartrand, T., Ravenscroft, J.E., Ashbolt, N.J., 2010. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Research 44, 4674-4691.

Till, D., McBride, G., Ball, A., Taylor, K., Pyle, E., 2008. Large-scale microbiological study: Rational, results and risks. Journal of Water and Health 6 (4), 443-460.

Wither, A., Greaves, J., Dunhill, I., Wyer, M., Stapleton, C., Kay, D., Humphrey, N., Watkins, J., Francis, C., McDonald, A., Crowther, J., 2005. Estimation of diffuse and point source microbial pollution in the Ribble catchment discharging to bathing waters in the North West of England. Water Science and Technology 51 (3-4), 191-198.

Zhu, Z., Broersma, K., Mazumder, A., 2011. Model assessment of cattle and climate impacts on stream fecal coliform pollution in the Salmon River Watershed, British Columbia, Canada. Water, Air and Soil Pollution 215, 155-176.