Modelling microbial pollutant loads associated with surface water run-off in water supply catchments

Tony Miller[1] Melanie E. Roberts[2] Brooke A. Swaffer[3]
Graeme Hocking[4] Bill Whiten[5] Robert McKibbin[6]

(Received 11 April 2017; revised 17 March 2018)

Abstract

The presence of microbial pathogens in surface water run-off from water catchments is a significant problem for many Australian water supply utilities. It is known that the microbial load in surface run-off can increase rapidly during rain events, and then declines a few hours afterwards. For the treatment of such water to ensure drinking water quality to be effective, it is important to have some reliable estimate of the microbial load in the raw water. Real time assessment of microbial load is not possible as accurate laboratory assays are time-consuming and expensive. This paper considers the possible use of alternative, surrogate measures of microbial load derived from

DOI:10.21914/anziamj.v58i0.12015, Published 2018-04-16, as part of the Proceedings of the 2016 Mathematics and Statistics in Industry Study Group. ISSN 1445-8810. (Print two pages per sheet of paper.)

physical flow attributes such as volumetric flow rate and turbidity. These measures are relatively easy to obtain and can be monitored automatically to give real-time continuous data streams. We use data collected over the past 2–10 years from a number of Adelaide Hills catchments to calibrate some regression models. A log-log model for microbial load with flow rate as the explanatory variable is shown to be a good fit, but with a sizeable estimated standard deviation. Various possible factors contributing to this variability are discussed. A physical modelling approach is also used to try to understand possible microbial 'washout' associated with rain events on a seasonal scale. An improved sampling technique is also suggested, which will potentially assist with obtaining better quality data for use in developing improved regression models in the future.

Contents

1 Introduction

One of the primary responsibilities of water utilities is the maintenance of mandated water quality standards for the water that they supply to their customers. The difficulty involved in this task depends to a large extent on the source of the raw water that is used. The quality of raw surface water collected from catchments that have complex and multiple patterns of land use is particularly problematic. In particular, agricultural enterprises dominated by livestock often lead to the increased occurrence of microbes in the soil that are potential human pathogens. Rainfall events liberate these microbes from the soil and they subsequently enter streams within the catchment through surface water run-off, and ultimately find their way into storage reservoirs. Historically, heavy rainfall often preceded outbreaks of drinking water borne disease [4, 7]. Curriero et al. [5] reported that 68% of disease outbreaks were preceded by rainfall events that were above the 80th percentile of rain intensity.

One of the aims of water treatment is to remove or neutralise microbes that are potentially injurious to human health. This is usually achieved by techniques such as filtration and chlorination, amongst others. Although, in principle, it is always possible to treat raw water to achieve this goal, to do so is costly. It is therefore desirable to avoid treating water beyond the extent needed to neutralise the microbes that are present, with a suitable safety

margin included. Unfortunately, the determination of microbial population counts in raw water is a complex laboratory process that is both costly and time consuming. Turn around times for assays are typically of the order of 2–3 days. Therefore, such assays cannot be used for effective real-time control of water treatment facilities. Moreover, the microbial entities of interest are small in size (2–15 μm), and they are typically present in the raw water at low concentrations in the range 0.1–10 L^{-1}, mixed in with other particles of similar size that are present at concentrations of $1 \times 10^6 - 5 \times 10^6\,L^{-1}$ [8]. Reliable routine sampling in such environments is challenging.

Although microbial concentrations cannot be determined in real time, a number of other, more physical, properties of the surface water run-off can be automatically measured in real-time with reasonable reliability. These physical properties include (volumetric) flow rate, turbidity, and electrical conductivity of the surface water, which are typically measured at a small number of locations in the major streams or creeks within a catchment. Precipitation (rainfall) can also be measured automatically at various points in the catchment. A natural question arising from this is whether real time knowledge of these physical quantities might be useful for predicting microbial concentrations. This could then provide a means of real-time management of water treatment facilities.

The first step towards answering this question is to gather microbial population count data along with corresponding physical run-off data from key sites in a catchment, and then use this to test and calibrate a prediction model. Over recent years SA Water[1] has gathered data of this kind for a number of Adelaide Hills catchments. For some catchments this data goes back ten years or so; however, this older data mostly only contains microbe count and flow rate. SA Water has progressively rolled out automatic sensors at key locations within their catchments in the last 3–4 years, and this more recent data includes the other physical variables mentioned above.

[1] SA Water Corporation (SA Water) is a SA State Government owned water utility that manages the water collection, treatment and distribution network throughout South Australia.

Some of this data is reported on by Swaffer et al. [10]. They showed significant correlation between microbe (*Crytosporidium*) concentration and flow rate (Spearman $\rho = 0.76$) and also turbidity (Spearman $\rho = 0.63$) for all rain triggered run-off events over a six month period at one specific location in a multi-use catchment. The six months considered, late Autumn to mid-Spring, was the major rainfall period of the year for this site. Brookes et al. [2] considered various surrogate indicators for *Crytosporidium*, including the presence of fecal microbial organisms, suspended particle size, and turbidity. Using inflow data following a major rain event, Spearman rank correlations of around $\rho = 0.58$ were found between microbe (*Crytosporidium*) concentration and the concentration of some indicator organisms, and around $\rho = 0.70$ for the concentration of medium size (14–28 µm) particles. Signor et al. [9] compared rainfall triggered high flow events and baseline flows. They also found significant increases in *Crytosporidium* following a rainfall event.

At the 2016 MISG workshop, SA Water asked the workshop to further explore the potential relation between the physical surface water run-off parameters (particularly flow rate and turbidity) on the one hand, and microbe concentration on the other. This paper describes some of the outcomes from that workshop, concentrating primarily on theoretical and statistical aspects. There is an extensive existing body of work in the water resources literature on catchment models that seek to relate surface water run-off flow rates to rainfall (precipitation), catchment topography, catchment hydrogeology and measures of catchment dryness. However, these are not directly relevant to this particular problem, as in this case measured hydrograph data is assumed to be available.

The structure of this paper is as follows. Section 2 examines some of the statistical issues around the standard ColorSeed laboratory technique used to assay microbes such as *Crytosporidium*. Section 3 develops some simple regression models for predicting the microbial load from the flow rate and turbidity. These models are based on the previously mentioned historical data collected by SA Water. A physical modelling approach is described in Section 4 with particular focus on seeking to understand possible microbial

washout. Section 5 describes an automatic sampling technique for determining peak flow rate. This could have important applications in obtaining better quality data for use in developing improved regression models in the future.

2 Statistical reliability of microbe population count assays

As stated earlier, the determination of microbial population counts in raw water is a challenging laboratory process. It is commonplace for microbes to be "lost" in the assay process. ColorSeed is a widely used assay technique for water borne pathogens such as *Crytosporidium* and *Giardia*, [6, 1]. To seek to correct for losses in the course of the assay, the ColorSeed technique adds a known number (100) of "marked" control microbes (marked with a florescent dye) to the sample before assay. The number of these marked microbes that appear in the final count after laboratory processing is then used to gross up the corresponding count of unmarked microbes to arrive at a corrected value that is then taken to be the "actual" microbe count in the original sample. This is based on the assumption that equal proportions of marked and unmarked microbes will be "lost". Recovery rates for the marked microbes of 30–50% are common in laboratory practice, but both higher and lower recovery rates also occur; thus, actual counts are routinely scaled up by factors of two to three.

2.1 Microbe measurement accuracy

The measurement of microbe concentration is done by collecting 10 litre samples, adding 100 marked organisms (marked with a florescent dye) then concentrating the organisms and counting the number of marked, M, and unmarked, R, microbes recovered. The "corrected" concentration of microbes

per 10 litres in the original sample is then

$$c = R\frac{100}{M}\,, \tag{1}$$

which provides a correction for loss of microbes during the concentration process.

We assume R is Poisson distributed with parameter r (giving a variance r), and that M is binomially distributed with parameters $n = 100$ and $p = m/100$ say, so that m is the expected value of M.

The relative accuracy of the microbes per 10 litres c can be found approximately analytically or by a Monte Carlo simulation. For the analytic case, substituting the mean values into (1) and differentiating gives

$$dc = \frac{100}{m}\,dr - r\frac{100}{m^2}\,dm\,. \tag{2}$$

Using the formula for the variance of a sum

$$\operatorname{var}(c) = \operatorname{var}(dc) = \left(\frac{100}{m}\right)^2 r + \left(r\frac{100}{m^2}\right)^2 m(1 - m/100). \tag{3}$$

For one standard deviation the relative error in c is

$$\sqrt{\frac{(100/m)^2 r + (r100/m^2)^2 m(1 - m/100))}{(r100/m)}} = \sqrt{(1/r + 1/m - 1/100)}\,. \tag{4}$$

A Monte Carlo calculation takes into account the skewness of the distributions when r or m are small. A large number of samples (e.g., 100 000) from the Poisson distribution (parameter r) and the binomial distribution (parameters 100 and $m/100$) are generated. These are used to generate the same number of samples of c from the formula for c. The ratio of the standard deviation of the c samples to the mean of the c samples gives the relative one standard deviation error for c. This Monte Carlo calculation took 174 seconds compared with 0.0002 seconds for the analytic approximation.

The two methods agree well for large values of $\mathtt{r}$ and $\mathtt{m}$, but the analytic estimates are low for small values. Table 1 gives values for the analytic estimate and Table 2 gives those for the Monte Carlo method. The Monte Carlo method also allows the calculation of percentage points of the distributions for each marked microbe and unmarked microbe count. Tables 6, 7, 9 and 10 in Appendix A give the 5, 15, 85 and 95 percentage points for the concentration distributions, and for comparison Table 8 gives the calculated concentration values using (1).

These error estimates are based solely on the counting statistics and thus are the minimum size for the errors. Any other additional source of errors will add to the size of errors estimated here.

2.2 Statistical tests for repeated microbe measurements

Within the data available were 41 groups of tests for the original microbe counts and 45 for the marked microbe counts where four samples were taken and processed separately. These repeats allow the statistical assumptions of the Poisson distributions for the original microbe counts and binomial for the marked microbes to be tested. In particular it was desired to test if the data variance was larger than expected for the distribution. The statistic

$$\sqrt{\frac{\text{variance}}{\text{mean}(1-\text{probability})}}\,, \tag{5}$$

which ratios the variance to the predicted variance, was investigated with the probability as mean/100 for the binomial distribution and zero for the Poisson distribution as it is a zero probability limit of the binomial distribution. This statistic squared is asymptopically related to the Chi-squared distribution [3], but needed to be investigated due to the small number of repeats and the variation in distribution parameters. Monte Carlo simulations generated the distribution of this statistic and it was found that for the parameter range

Figure 1: Distribution of 10^6 samples of the statistic $\sqrt{\frac{\text{variance}}{\text{mean}(1-\text{probability})}}$ for binomial ($\mathrm{p} = 0.1, 0.5, 0.9$) and Poisson ($\lambda = 5, 20, 80$). This curve is to the resolution of the graph approximated by $\sqrt{\chi^2(3)/3}$, where $\chi^2(3)$ is the chi-squared distribution with three degrees of freedom.

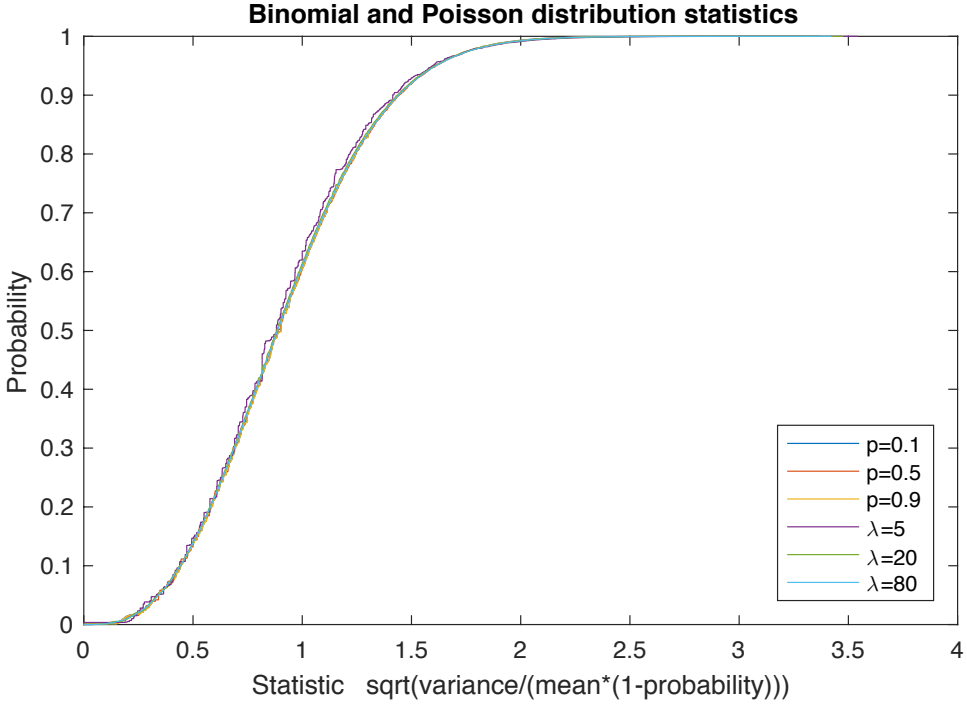
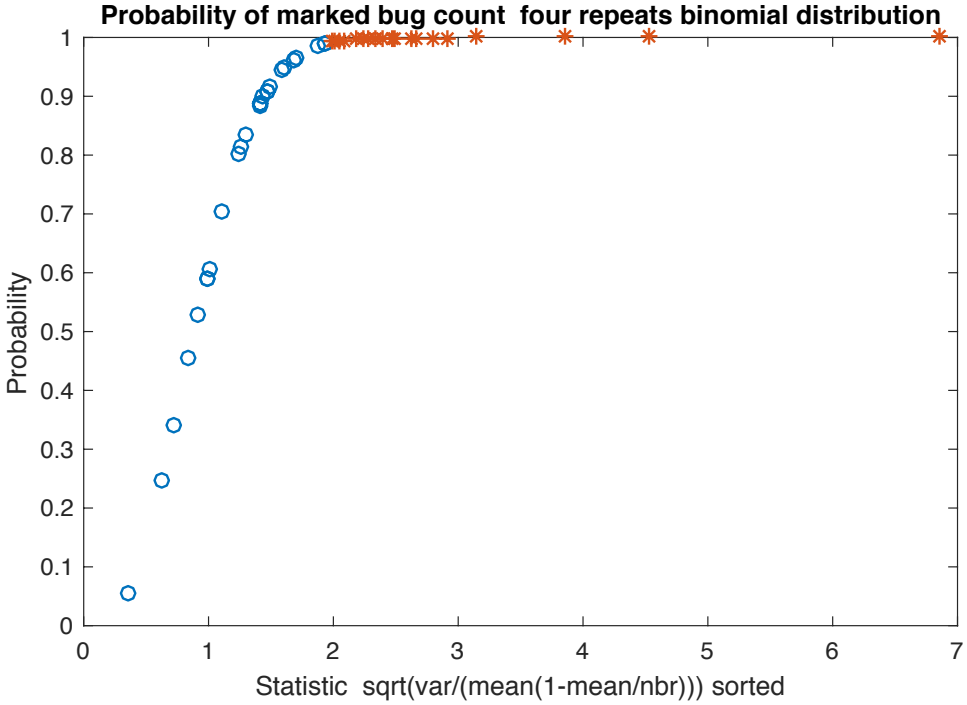


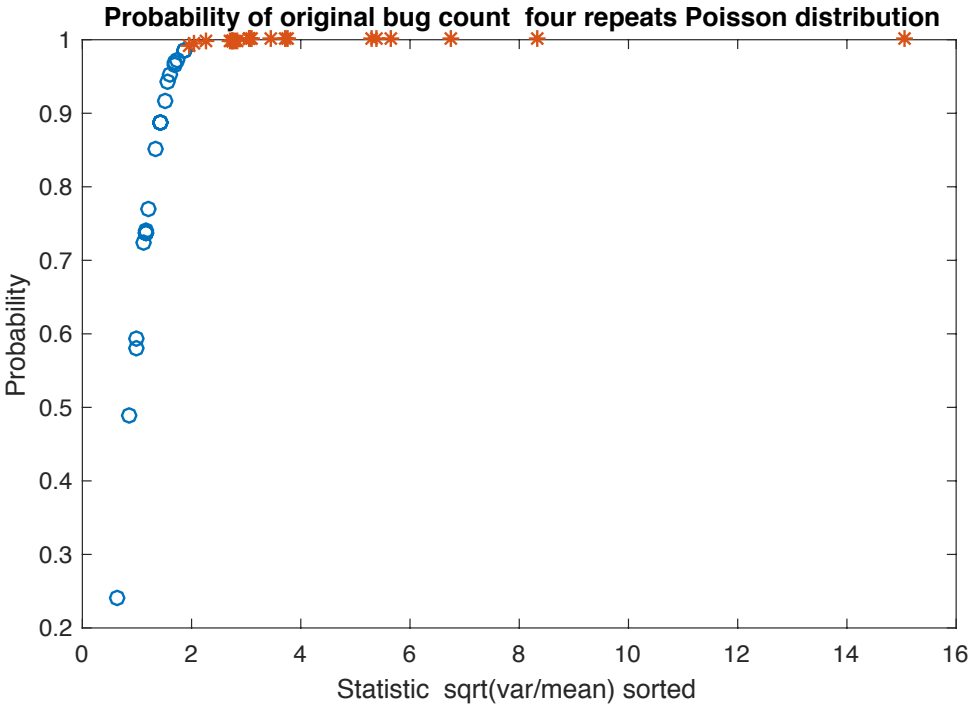
Figure 2: Probability of repeat samples of marked microbes, blue circles are those which satisfy the binomial statistical criteria.



of interest the distribution is independent of the distribution parameters as shown in Figure 1, which shows the binomial curves for $\mathfrak{n} = 100$ and the Poisson curves. The Poisson curve for $\lambda = 5$ is very marginally above the others but is well within the accuracy needed for this application, and the other curves are very similar.

The distribution curve is used to evaluate the experimental data by converting the statistic (5) to a probability value as in Figure 1. Figure 2 shows the probability distribution for the marked microbe count with the red stars indicating those that are above 99% probability and hence probably not consistent with the binomial distribution assumption. Twenty-six of the

Figure 3: Probability of repeat samples of original microbes, blue circles are those which satisfy the Poisson statistical criteria.



45 cases are considered consistent with the binomial assumption. Similarly Figure 3 shows the probability distribution for the original microbe counts. Twenty one of 41 cases (in blue) are considered consistent with the Poisson assumption (probability < 0.99). The samples not consistent with the statistic assumptions have received additional variation during analysis, or additionally from the original count during sampling or transfer. Since a reasonable proportion of the samples are within the statistical limits, the sampling and analysis method is capable of giving good results.

Table 1: Relative error in microbe concentration determined using the analytic approximation.

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	0.36	0.28	0.25	0.23	0.22	0.22	0.21	0.21	0.20	0.20
50	0.33	0.24	0.21	0.19	0.17	0.16	0.16	0.15	0.15	0.14
75	0.32	0.23	0.19	0.17	0.15	0.14	0.13	0.13	0.12	0.12
100	0.32	0.22	0.18	0.16	0.14	0.13	0.12	0.11	0.11	0.10
125	0.31	0.22	0.18	0.15	0.13	0.12	0.11	0.10	0.10	0.09
150	0.31	0.22	0.17	0.15	0.13	0.12	0.10	0.10	0.09	0.08
175	0.31	0.21	0.17	0.14	0.13	0.11	0.10	0.09	0.08	0.08
200	0.31	0.21	0.17	0.14	0.12	0.11	0.10	0.09	0.08	0.07
225	0.31	0.21	0.17	0.14	0.12	0.11	0.09	0.08	0.07	0.07
250	0.31	0.21	0.17	0.14	0.12	0.10	0.09	0.08	0.07	0.06
275	0.31	0.21	0.16	0.14	0.12	0.10	0.09	0.08	0.07	0.06
300	0.31	0.21	0.16	0.14	0.12	0.10	0.09	0.08	0.07	0.06
325	0.31	0.21	0.16	0.13	0.11	0.10	0.09	0.07	0.06	0.06
350	0.30	0.21	0.16	0.13	0.11	0.10	0.08	0.07	0.06	0.05
375	0.30	0.21	0.16	0.13	0.11	0.10	0.08	0.07	0.06	0.05
400	0.30	0.21	0.16	0.13	0.11	0.10	0.08	0.07	0.06	0.05
425	0.30	0.21	0.16	0.13	0.11	0.09	0.08	0.07	0.06	0.05
450	0.30	0.21	0.16	0.13	0.11	0.09	0.08	0.07	0.06	0.05
475	0.30	0.21	0.16	0.13	0.11	0.09	0.08	0.07	0.06	0.05
500	0.30	0.20	0.16	0.13	0.11	0.09	0.08	0.07	0.06	0.04

Table 2: Relative error in microbe concentration determined using Monte Carlo calculation.

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	0.47	0.30	0.26	0.24	0.23	0.22	0.21	0.21	0.20	0.20
50	0.43	0.27	0.22	0.19	0.18	0.16	0.16	0.15	0.15	0.14
75	0.43	0.25	0.20	0.17	0.15	0.14	0.13	0.13	0.12	0.12
100	0.43	0.25	0.19	0.16	0.14	0.13	0.12	0.11	0.11	0.10
125	0.43	0.24	0.19	0.16	0.14	0.12	0.11	0.10	0.10	0.09
150	0.42	0.24	0.18	0.15	0.13	0.12	0.11	0.10	0.09	0.08
175	0.42	0.24	0.18	0.15	0.13	0.11	0.10	0.09	0.08	0.08
200	0.42	0.23	0.18	0.15	0.13	0.11	0.10	0.09	0.08	0.07
225	0.42	0.23	0.18	0.14	0.12	0.11	0.09	0.08	0.07	0.07
250	0.42	0.23	0.18	0.14	0.12	0.10	0.09	0.08	0.07	0.06
275	0.42	0.23	0.17	0.14	0.12	0.10	0.09	0.08	0.07	0.06
300	0.42	0.23	0.17	0.14	0.12	0.10	0.09	0.08	0.07	0.06
325	0.42	0.23	0.17	0.14	0.12	0.10	0.09	0.08	0.06	0.06
350	0.42	0.23	0.17	0.14	0.12	0.10	0.09	0.07	0.06	0.05
375	0.41	0.23	0.17	0.14	0.12	0.10	0.08	0.07	0.06	0.05
400	0.42	0.23	0.17	0.14	0.11	0.10	0.08	0.07	0.06	0.05
425	0.42	0.23	0.17	0.14	0.11	0.10	0.08	0.07	0.06	0.05
450	0.41	0.23	0.17	0.14	0.11	0.10	0.08	0.07	0.06	0.05
475	0.42	0.23	0.17	0.14	0.11	0.10	0.08	0.07	0.06	0.05
500	0.42	0.23	0.17	0.13	0.11	0.10	0.08	0.07	0.06	0.04

2.3 Alternative statistical assumptions

An apparent alternative to the assumption that the count of recovered microbes (R) is distributed as a Poisson distribution, is to go back to the distribution of microbes originally collected in the 10 litre sample and assume that it has a Poisson distribution, and then adjust that number by the binomial distribution seen from the marked microbes. Let C be the number of microbes in the 10 litre sample, R be the number that were counted in the assay, and p be the fraction ($m/100$) of marked bugs that were counted in the assay out of the original 100 that were introduced. Suppose that

$$C \sim \text{Poisson}(c) \quad \text{and} \quad R \sim \text{binomial}(C, p). \tag{6}$$

From the relation for conditional binomials [12], if $R \sim \text{binomial}(C, p)$ and $C \sim \text{binomial}(c^*, q)$ then

$$R \sim \text{binomial}(c^*, pq). \tag{7}$$

Now taking limits $q \to 0$ with $c^*q = c$ constant, and $pq \to 0$ with $c^*pq = cp$ constant, shows that if $C \sim \text{Poisson}(c)$ and $R \sim \text{binomial}(C, p)$ then in the limit

$$R \sim \text{Poisson}(cp) \tag{8}$$

where cp is estimated as the number of microbes observed. (We used that a Poisson distribution $\text{Poisson}(r)$ is the limiting form of the binomial distribution $\text{binomial}(n, p)$ as $p \to 0$ with $np = r$ fixed.) Thus, this alternative assumption turns out to be identical to that used in the previous subsection, and so the same conclusions follow.

The data may also be analysed using a Bayesian approach that generates distributions for the parameters cp and p. From these distributions a distribution for the test statistic can be generated. As the two distributions ($M \sim \text{binomial}(100, p)$ and $R \sim \text{Poisson}(cp)$) are independent and each rely on a single variable, the Bayesian calculation is easily completed. However, as has been demonstrated the distributions for the test statistic ($\sqrt{\text{var}}/[\text{mean}(1-p)]$) are independent of cp and p, so simulation of the distribution in Figure 1 remains the same.

3 Prediction of microbe concentration from flow rate and turbidity

3.1 Nature of the available data

The detailed small scale study described by Swaffer et al. [10] shows that the flow rate, turbidity and microbe concentration at a stream location can all change significantly over a period of a few hours following a rain event. Therefore, in fitting any model that seeks to relate instantaneous microbial concentration and physical stream parameters, it is essential that the collection of the water sample for assay and the measurement of the relevant physical stream parameters take place as nearly as possible at the same time. Four such synchronised data sets were made available by SA Water. Each corresponded to a stream location in one of four different catchments. Table 3 summarises the characteristics of the datasets. The data consisted of microbial concentration (expressed as corrected counts per 10 L), volumetric flow rates (as m^3s^{-1}) and turbidity values (in NTU units[2]). The collection of turbidity data only commenced following the introduction of new automatic measurement equipment, hence turbidity data is only available from around the middle of 2013 onwards. Datasets including observations prior to this only have microbial concentration and flow rate at some observation points. As part of the equipment upgrade for automated sensing, the method used to measure the flow rate also changed. This more recent flow rate data collected using the automated sensors is considered by SA Water to be more reliable than data obtained by the previous collection method.

Any samples for which the microbe concentration was recorded as zero were excluded from the data considered here, as it was not possible to distinguish between a zero concentration, and a non-zero concentration that was below the limit of detection. As significant summer rain is not common

[2]Nephelometric Turbidity Units (NTU) are based on a comparative scale of light scattering relative to scattering in a reference material and measured under some standard conditions.

Table 3: Description of the data collected at four locations in SA Water's Adelaide Hills catchments. The data was microbial concentration (corrected counts per 10 L), volumetric flow rate ($m^3 s^{-1}$) and turbidity (NTU units), unless indicated otherwise. Zero count data has been excluded.

Location code	Dates	Number	Comments
GWF	Oct 2013 – Nov 2015	$n = 44$	
KC	May 2008 – Sept 2015	$n = 81$	turbidity included from July 2013 onwards. ($n = 42$)
LP	May 2014 – Sept 2015	$n = 34$	
Myp	Aug 2000 – Nov 2015	$n = 140$	turbidity included from May 2013 onwards. ($n = 32$)

for these catchments, data collected during the summer months was also excluded, as there may be significantly different factors influencing the microbe concentration during the summer. (Annual rainfall in these catchments is typically around 800–1000 mm, 80% of which falls during April–October, peaking in the winter months of June and July.)

The microbe concentration data came from both routine samples, that is samples collected at semi-regular scheduled times, and samples that were deliberately collected in response to forecast or actual rain events. Since the collection of samples for assay has been a manual process to date, such rain event collections tend to occur at somewhat random times during the hydrograph cycle that follows the initiating rain event. One observed side effect of this is that microbial count data was mostly collected on the rising edge of the hydrograph or from the long trailing edge of the hydrograph, but seldom at or near peak flow rates. Thus, the available data may not cover

the full range of flow rates that actually occur, being biased away from peak flows, which is where we might expect microbial loadings to be the highest. We return to this point in Section 5.

3.2 Regression model of microbial load

The problem as presented to the MISG was formulated in terms of microbial concentration, as this is what is directly assayed. However, a quantity that is probably of more direct practical importance is the microbial load $Q = qC$, defined as the product of the microbial concentration C (number per 10 L) and the volumetric flow rate q. This describes the number of microbes that pass the sampling location per unit of time, and is a more direct measure of how many microbes are entering the water storage system. Using the previously stated units for q and C, the units for Q are $10^2\,\mathrm{s}^{-1}$.

A log-log relationship of the linear form

$$\log_{10} Q = a + b \log_{10} q + c \log_{10} T$$

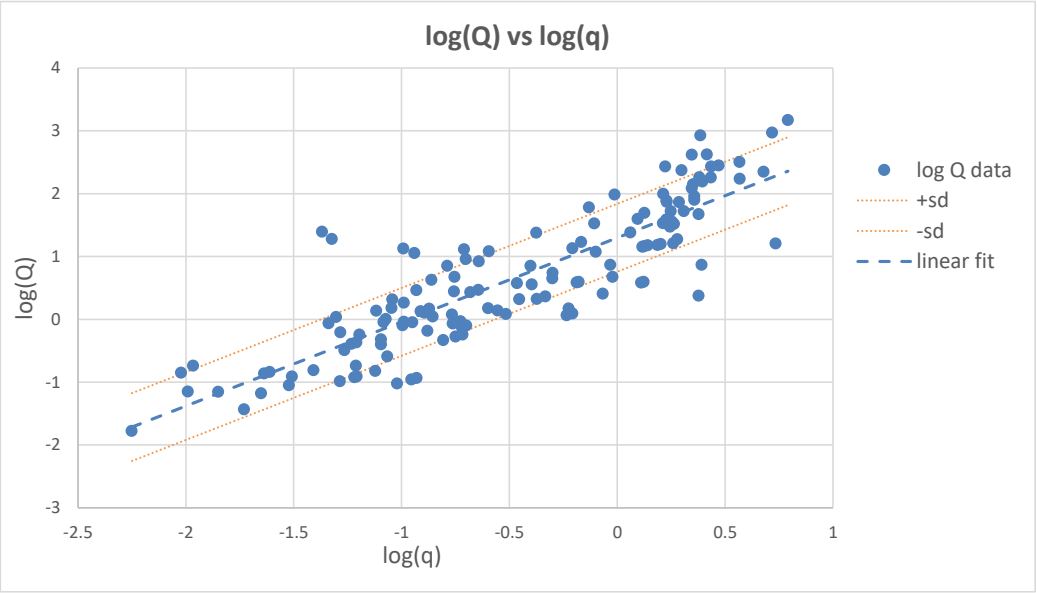
was fitted by simple regression using the four datasets, and selected subsets of them. Here T denotes the turbidity. A logarithmic transformation of Q was used as this tended to stabilise the variance of the fitted model. Also, on physical grounds, an error model that assumes a fixed relative error in Q, as opposed to a fixed absolute error, seems more reasonable. This again supports a logarithmic transformation of Q. Table 4 summarises the fitted parameters for the various catchments, or datasets.

Figure 4 shows the data and the linear fit in log-log coordinates for the first of the Myp regressions in Table 4. The plus and minus one estimated standard deviation of prediction lines are also shown around the fitted line. The estimated standard deviation of prediction and the estimated standard deviation of the data differ by less than 2% in this case. (See any standard text such as Weisberg [11] for details of the estimated standard deviation of prediction.) To assess the validity of the standard regression assumptions for

Table 4: Summary of the results of the log-log regressions for the different data sets. Here a, b, and c are the coefficients of the constant, $\log_{10} q$ and $\log_{10} T$ terms respectively; except for Myp* where the coefficient in the a and c columns are a_0 and a_1 respectively. See text for further details. No value for a or c indicates that $\log_{10} q$ or $\log_{10} T$ was not used in the regression. The adj. R^2 column shows the adjusted R^2, n is the number of data points used in the regression and est. s.d. is the estimated standard deviation of the data obtained from the residuals. A $\star$ in the comments column indicates that only data post mid-2013 is used. A * indicates that a seasonal factor is incorporated, using a 0/1 factor to indicate pre/post 1 July. Unless otherwise indicated, the complete data set is used with no adjustments.

Location	a	b	c	adj.R^2	n	est. s.d.	comments
GWF	1.42	1.86	-	0.895	44	0.553	
GWF	1.02	1.66	0.26	0.897	44	0.547	
KC	2.20	1.59	-	0.860	81	0.462	
KC	2.20	1.62	-	0.894	42	0.340	$\star$
KC	2.06	1.59	0.090	0.893	42	0.342	$\star$
LP	1.43	1.47	-	0.858	34	0.501	
Myp	1.30	1.34	-	0.754	140	0.541	
Myp*	1.24*	1.31	0.0875*	0.753	140	0.542	*
Myp	1.28	1.47	-	0.824	32	0.468	$\star$
Myp	0.188	1.20	0.883	0.844	32	0.440	$\star$
Myp	-2.76	-	2.96	0.581	32	0.722	$\star$

Figure 4: Data and linear fit of load Q versus volumetric flow rate q plotted in log-log form for the full Myp data set regression in Table 4. Plus and minus one estimated standard deviation lines are shown.



this case, Figure 5 shows a normal probability (rankit or Q-Q) plot of the regression residuals. Except for a handful of extreme values, the residuals lie close to a straight line, indicating approximate normality. Although strictly speaking, the regression residuals are not normally distributed under the standard regression assumption of normal errors, in this case the degrees of freedom $n-p = 138$ is large, and so the regression residuals are approximately normal [11]. These observations justify our use of a log transformation of the load Q data.

Transforming back to the original variables Q and q gives a power law relation $Q = 10^a q^b T^c$, which is shown graphically in Figure 6. This plot also shows the transformed +1 and +2 estimated standard deviation bands. Clearly the spread of these bands increases with q as they represent a fixed relative

Figure 5: Normal probability (rankit or Q-Q) plot of residuals from the full Myp data set regression in Table 4. The scale on the vertical axis is the inverse of the standard normal cumulative distribution function (InvNormcdf) applied to the empirical cumulative distribution of the residuals. So for example, 0 on the vertical axis corresponds to the median. A straight line through $(0, 0)$ would represent a normal distribution.

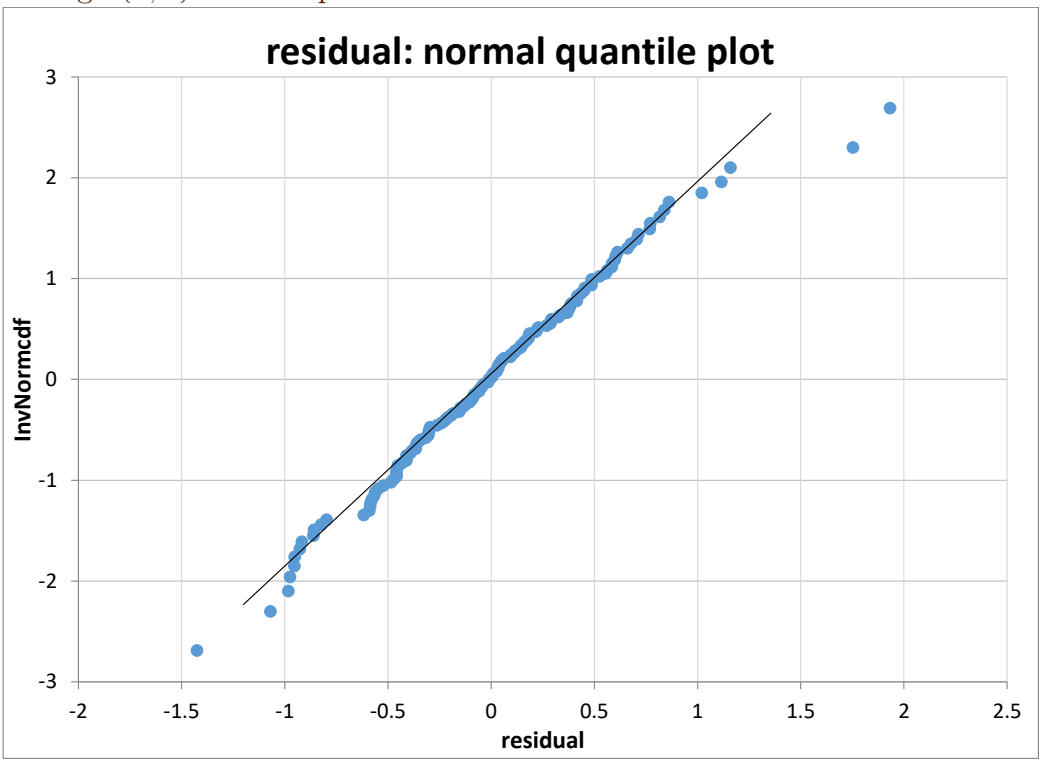
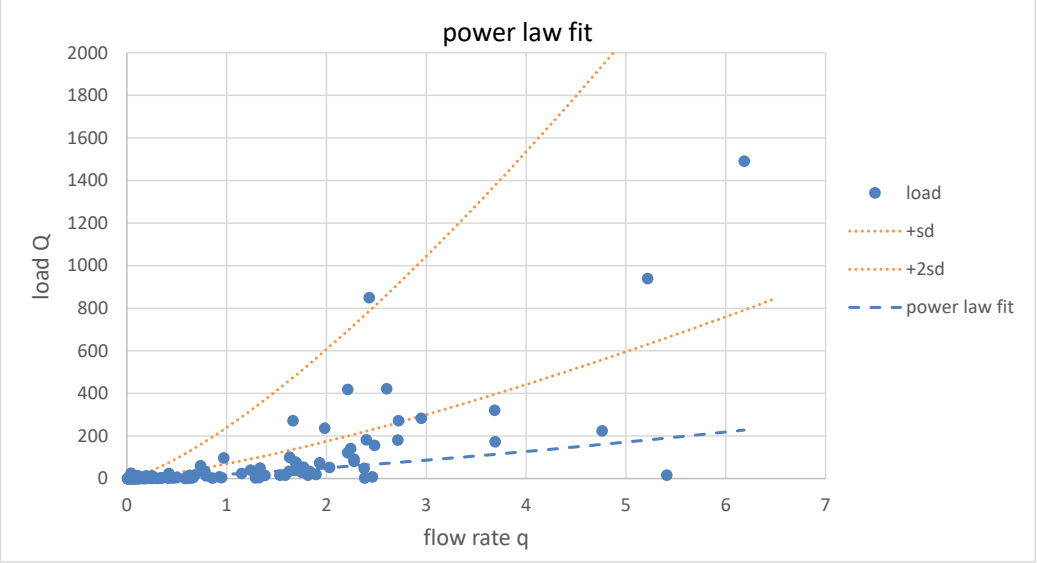


Figure 6: Plot of the load Q data and the power law fit obtained by back transforming the regression shown in Figure 4. Bands corresponding to back transformed +1 and +2 estimated standard deviations are shown.

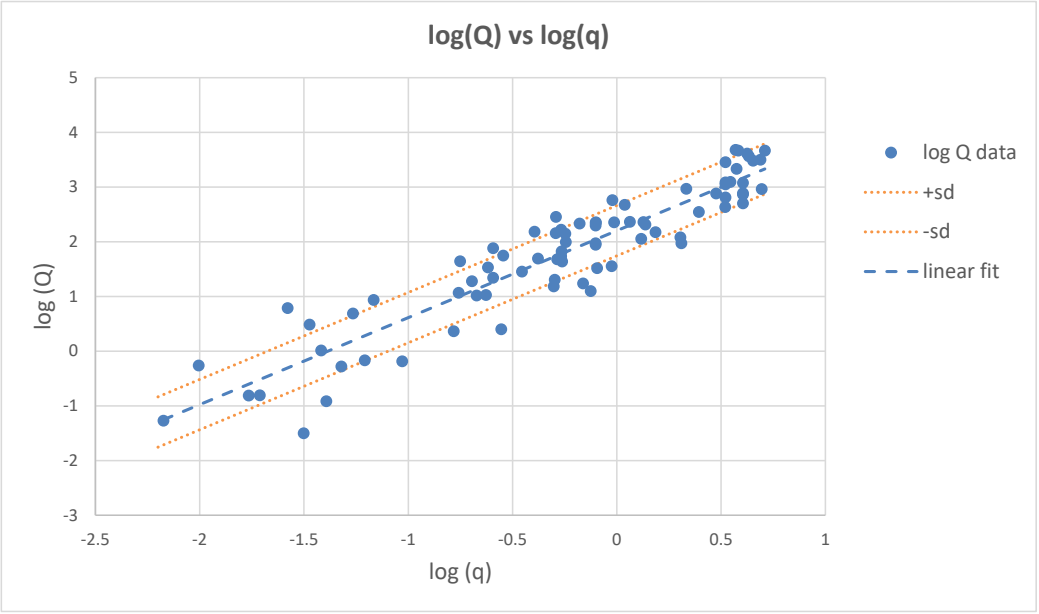


fraction of the fitted power law.

Similar plots are also obtained for the other regressions in Table 4. As one further example, Figures 7 and 8 show the log-log data and fit and the data and fit in the original variables for the full KC data set.

All of the regressions summarised in Table 4 that use $\log_{10} q$ as an explanatory variable have adjusted R^2 values over 0.82, with the exception of the full Myp data set. Attempts to improve the quality of the model fit by introducing $\log_{10} T$ as another explanatory variable had only a marginal effect. This is not completely unexpected since, as shown in Table 5, $\log_{10} q$ and $\log_{10} T$ are reasonably correlated, at least for some of the data sets. Nonetheless, it is still a little surprising to us that introducing turbidity has such a small effect on improving the regression model, since, at least intuitively, high

Figure 7: Data and linear fit of load Q versus volumetric flow rate q plotted in log-log form for the full KC data set regression in Table 4. Plus and minus one estimated standard deviation lines are shown.



turbidity, which indicates a large amount of suspended material present in the run-off stream, should mean that a high number of microbes are also present. Confirming this apparent lack of relevance of turbidity is the last Myp regression listed in Table 4, which uses $\log_{10} T$ as the only explanatory variable and gives a markedly lower adjusted R^2.

For the Myp data a very crude attempt to introduce a seasonal effect was investigated. The data was blocked according to whether the date was in the first or the second half of the year. Each block was allowed to have a different constant term in the regression. More specifically, we considered the model

$$\log_{10} Q = a_0 + a_1 M + b \log_{10} q$$

where $M = 0$ for the months Jan–June (inclusive), and $M = 1$ for the

Figure 8: Plot of the load Q data and the power law fit obtained by back transforming the regression shown in Figure 7. Bands corresponding to back transformed +1 and +2 estimated standard deviations are shown.

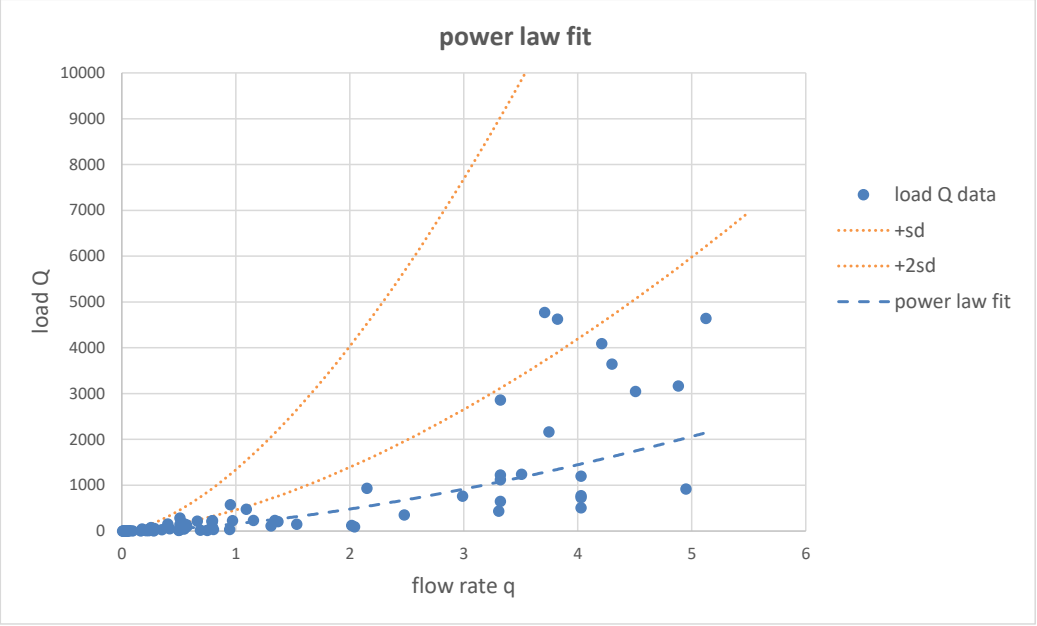


Table 5: Correlation of $\log_{10} q$ and $\log_{10} T$ for the datasets described in Table 3. Datasets are restricted to data for which both the flow rate q and the turbidity T are available.

Location code	corr.$(\log_{10} q, \log_{10} T)$	n
GWF	0.844	44
KC	0.453	42
LP	0.730	34
Myp	0.728	32

months of July–Dec. (inclusive). However, using this enhanced model had no noticeable effect on the quality of the fit, see the row marked Myp* in Table 4. Extending this to allow the coefficient of $\log_{10} q$ to also depend on M, as in

$$\log_{10} Q = a_0 + a_1 M + b_0 \log_{10} q + b_1 M \log_{10} q \,,$$

again produced no improvement ($a_0 = 1.29$, $a_1 = 0.0283$, $b_0 = 1.368$, $b_1 = -0.111$, adj. $R^2 = 0.753$, $n = 140$, est. s.d. $= 0.543$).

Another observation from Table 4 is that the estimated standard deviation of the error lies in the range of 0.35–0.55, irrespective of the catchment. Also, for the KC and Myp datasets where there was both old and more recent flow rate data, restricting to the more recent data (that is, the post mid-2013 data) decreases the estimated standard deviation noticeably, supporting the belief that the post mid-2013 data for q is more reliable. For KC and Myp, excluding the pre mid-2013 data for q hardly changes the fit parameters a and b, but, as just noted, reduces the estimated standard deviation.

3.3 Nature of the regression error

The assumption that there is a well defined functional relationship between the instantaneous values of the load Q and the flow rate q, and that the only source of error is "measurement" error arising from the assay and sampling errors, is likely to be simplistic. To gain some indication of the likely magnitude of this measurement component of the error, some repeat measurements were identified in the data for Q. Across all four catchment data sets there were five cases where four nominally identical measurements had been made, one case of three measurements, and one case of two measurements. However, whether or not these nominal repeats were truly independent repeats of the same measurement was not able to be determined. Whilst it is almost certain that the assays were independent, the sampling may not have been. For instance, the repeat samples may have just been subsamples of a single large sample. Nonetheless, pooling all these repeats and calculating a pooled variance gives

an estimated standard deviation of 0.19 (for 18 degrees of freedom). The individual unpooled standard deviations ranged from 0.04 to 0.21. Bearing in mind that the repeat measurements may not have been truly independent repeats, this estimate may err on the low side. The regressions above consistently yielded estimated standard deviations lying approximately in the range 0.35–0.55. It would therefore seem that the measurement component is a significant, but possibly not the major component, of this overall standard deviation (it is the variances, that is the square of the standard deviations, that we might expect to be approximately "additive" over various sources of error.)

There are many reasons to believe that there are other, quite significant sources of error in such a simple model. The distribution of microbes is unlikely to be uniform across a catchment, and rain need not consistently fall uniformly within a catchment. Thus, the same volume of run-off from a catchment may well carry different numbers of microbes depending upon where it falls in the catchment. Moreover, it is likely that there will be some kind of washout effect, which may operate on a number of timescales. During a heavy rainfall event, there may initially be an high rate of microbial uptake by the run-off, but this may decrease with time as the microbes in the soil become depleted and fewer are available for uptake. Depending on the source and life cycle of the microbes this depletion effect may also play out over the longer, seasonal timescales, with lower uptake rates later on in the rainy season. Although a crude seasonal factor was experimented with above, its negligible effect should not be taken to rule out more subtle seasonal effects. Thus we expect both spatial and temporal effects to be significant, neither of which are adequately captured by a simple model of Q depending only on q.

More generally, an increasing functional relationship in the log-log domain cannot apply over all ranges of q, since Q cannot increase indefinitely as the flow rate q increases since the supply of microbes must ultimately be exhausted. Thus we expect there to be a plateau in the Q versus q plot at some stage, potentially followed by a decrease. However, whether such a plateau is reached under normal rainfall conditions is likely to be catchment specific.

The data considered in this paper gives no suggestion of such a plateau.

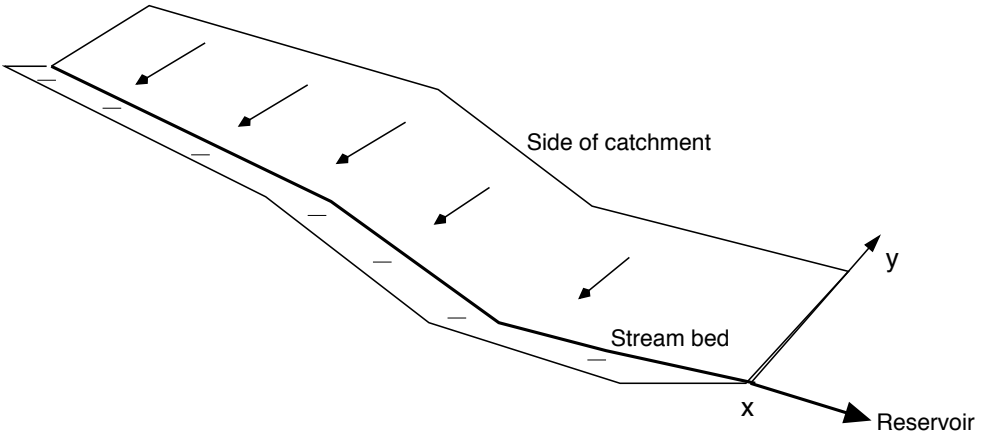
A constant standard deviation in the log-log description translates into error bands of a fixed ratio when transformed back into the original variables of Q and q. Thus a standard deviation of 0.4, say, which is broadly representative of the cases summarised in Table 4, corresponds to a ratio of $10^{0.4} \approx 2.5$. Thus the one-sigma and two-sigma error bands can become quite large for large values of the flow rate q, as demonstrated in Figures 6 and 8. These large error bands are due to the observed logarithmic nature of the error, or in other words, the error corresponds more to fixed relative errors rather than fixed absolute errors; and the large number of unknown factors, which the above simple model is, in effect, statistically averaging over. Whether, despite the large error bands, curves such as these can be useful in practice for run-off control and management by water utilities requires further investigation.

4 Physical model of the transport of microbe particles in run-off

In order to gain a better understanding of the transport of microbes in the surface run-off within a catchment, a simple physical model was developed. In this model the microbes are considered to be inert particles that have an initial population in the soil. During a rain event they enter the run-off from the soil at a rate proportional to both the local run-off flow rate and the number of microbes remaining in the soil. No regeneration of particles within the soil is allowed for during a rain event, and so "washout" may occur in some circumstances. Assuming a single or series of rainfall events and making certain simplifying assumptions about the nature of the runoff and the topography of the catchment, we compute the accumulated flow down the catchment, along the stream and entering the reservoir (Figure 9).

The problem is resolved into two components. In the first, the flow of water down the slope due to a rain event is calculated. Once the flow is known, the

Figure 9: Schematic diagram of flow. Water flows down the side of the catchment and joins the stream, eventually entering the reservoir. The distance up the slope is y, and along the streambed the distance is x to a maximum of X, the length of the catchment.



uptake of particulate during the run-off is computed. In the second phase the flow along the stream is computed using the input from the run-off.

4.1 Run off to the stream

Define the "height" of water running down the slopes in a valley in the y-direction to be $h(y,t)$. The height h is just representative of the flow down the slope rather than actual height. We compute the one-dimensional flow down a broad, flat incline. Define the distance up the slope to be y, and assume the speed of flow to be a constant, v.

In that case, the equation for flow quantity, h, is

$$h_t - v h_y = Q_R(y,t) \tag{9}$$

where $Q_R(y,t)$ can be thought of as the temporal and spatial (up the slope)

extent of a rainfall event or simply as the amount of water that eventually makes it to the stream from each location due to a rainfall event. If the result of the rainfall and consequent run off are centred over the stream and are Gaussian in time, centred at $t = T$, then

$$Q_R(y,t) = \frac{2R_0}{\kappa} e^{-[(t-T)/\sigma_T]^2} e^{-(y/\sigma_Y)^2} \tag{10}$$

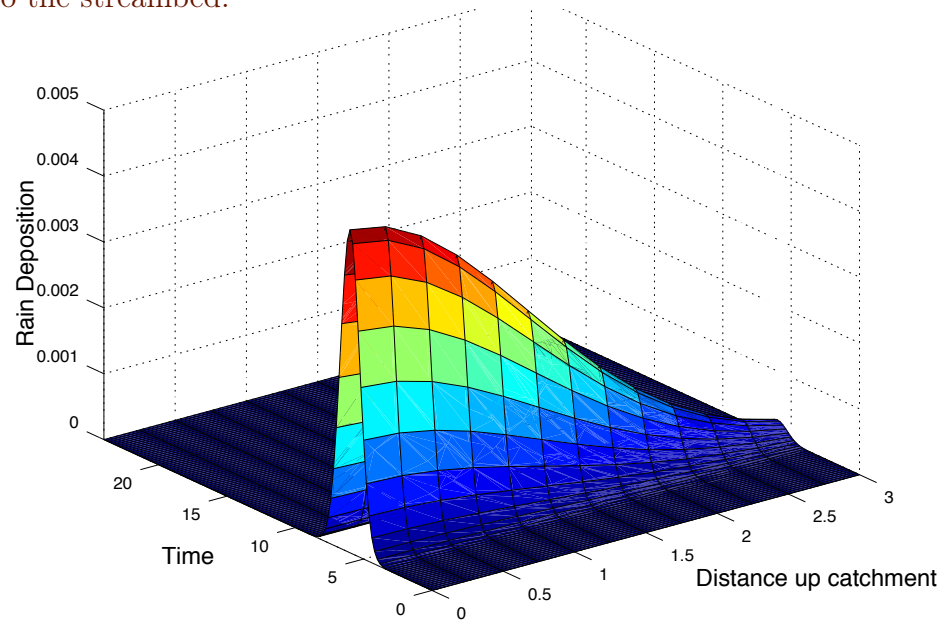
where $\kappa = \pi\sigma_Y\sigma_T\,[\mathrm{erf}(T/\sigma_T) + 1]$ standardises the total amount of rainfall for different spread values, σ_T and σ_Y for time and space respectively, and the time of maximum rainfall, T. This form assumes that water falling higher up the slopes does not all make it to the stream. R_0 is the intensity of rainfall, and doubling it will double the amount of rain. Changing the other parameters does not change the amount of rainfall, only the timing and extent. Thus the total amount of rain falling over a given catchment region is equivalent, independent of when and how it falls. This quantity is meant to be a trial function only so that we can compare behaviour. Figure 10 shows the rainfall function for a case with $R_0 = 4 \times 10^{-5}$, $\sigma_Y = 1.75$, $\sigma_T = 1$, $T = 6$, which represents about 20 mm of rain falling over a time period of about four hours, with heaviest rain six hours after midnight. The function represents that rain water closer to the bottom of the river valley is more likely to eventually reach the stream.

The level of particulate is small and so we treat it as a tracer that is simply carried along with the flow. Therefore, the level of particulate $A(y,t)$ depends on the water flow level $h(y,t)$, flow velocity v, and the concentration level on the ground, $G(y,t)$. The appropriate equation for particulate concentration in the runoff water is

$$(hA)_t - v(hA)_y = k_0 v h(y,t) G(y,t), \quad A(y,0) = 0. \tag{11}$$

The concentration of the particulate is computed by making some assumptions about the rate at which it is picked up during the downflow in the catchment. Since the concentration is always very low, we assume that it is absorbed into the downflow at a rate proportional to the "height" of the downflow (with rate

Figure 10: Shape of rainfall deposition curve (10) over the catchment upstream. Variation in y goes up the slope of the side of the catchment. The streambed is situated at $y = 0$. Less water from high up the catchment makes it to the streambed.

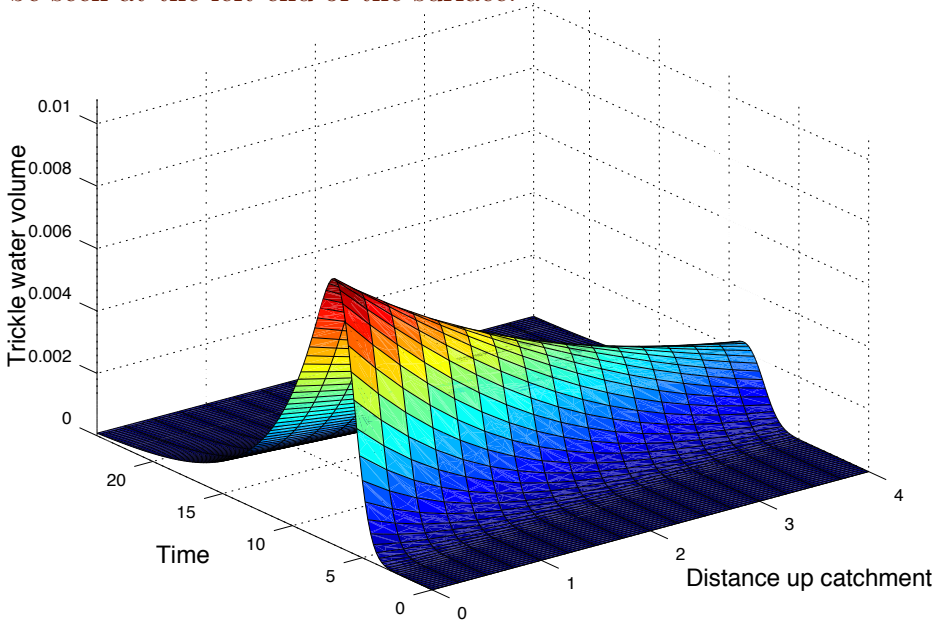


constant k_0), and to the concentration in the soil, thus at a particular point

$$\frac{dG}{dt} = -k_0 v h(y,t) G(y,t), \quad G(y,0) = G_0. \tag{12}$$

If the particulate in the soil is always constant, then we simply set $G = G_0$ at all times and the uptake continues. Otherwise, the amount of G decays exponentially depending on the flow rate. The longer the flow occurs, the more of the particulate is washed away, so that eventually this term drops to zero. The equations for the water flow decouple from the particulate concentration, and we obtain the solution for $h(y,t)$ by determining the

Figure 11: Shape of runoff volume up the catchment and over time. The streambed is situated at $y = 0$, so the amount of water entering the stream can be seen at the left end of the surface.



characteristics of the equation in the usual way. The solution for $h(y,t)$ is

$$h(y,t) = \frac{2R_0 \exp\left\{-\left[\frac{y-v(t-T)}{D}\right]\right\}}{\sqrt{\pi}D\left[\operatorname{erf}(\frac{T}{\sigma_T})+1\right]} \tag{13}$$
$$\times\left\{\operatorname{erf}\left[\frac{\left(\frac{\sigma_T}{\sigma_Y}\right)vy+\left(\frac{\sigma_Y}{\sigma_T}\right)(t-T)}{\sqrt{D}}\right]-\operatorname{erf}\left[\frac{\left(\frac{\sigma_T}{\sigma_Y}\right)v(y-vt)-\left(\frac{\sigma_Y}{\sigma_T}\right)T}{\sqrt{D}}\right]\right\}.$$

where $D = c^2\sigma_T^2 + \sigma_Y^2$. Setting $y = 0$ in this expression gives the level of water entering the stream as a function of time. Figure 10 shows the shape of the rainfall function, Q_R, and Figure 11 shows the resulting runoff graph

for $h(y,t)$. The sharp rise is due to the sudden rainfall event while the long tail is due to the time taken for the runoff to trickle down the side of the catchment after the rain has stopped.

Once the value of $h(y,t)$ is known we solve PDE (11) for $A(y,t)$ by rearranging (using PDE (9)) to give

$$A_t - \nu A_y = k_0 \nu G(y,t) - \frac{Q_R(y,t)A(y,t)}{h(y,t)}, \quad A(y,0) = 0, \tag{14}$$

a linear, but very nasty, equation for particulate concentration in the runoff water. Evaluating $G(y,t)$ exactly, using ODE (12), gives

$$G(y,t) = G_0 \exp\left[\frac{2R_0(E+F)e^{-y/D}}{\sqrt{\pi D}}\right], \tag{15}$$

where

$$E = \left(\frac{N}{M} + t\right)\left[\operatorname{erf}\left(\frac{Mt+N}{\sqrt{D}}\right) - \operatorname{erf}\left(\frac{N}{\sqrt{D}}\right)\right]$$

and

$$F = \frac{D}{\sqrt{\pi}M}\left[\exp\left(-\left(\frac{Mt+N}{\sqrt{D}}\right)^2\right) - \exp\left(-\frac{N^2}{D}\right)\right]$$

with

$$M = \nu^2\left(\frac{\sigma_T}{\sigma_Y}\right) + \left(\frac{\sigma_Y}{\sigma_T}\right) \quad \text{and} \quad N = (y - \nu t)\nu\left(\frac{\sigma_T}{\sigma_Y}\right) - \left(\frac{\sigma_Y}{\sigma_T}\right)T.$$

This function can be used to evaluate the remaining particulate concentration at any location on the slope, and is also required to solve for the particulate in the trickle water. The solution for $A(y,t)$ must be obtained from PDE (14), and this is done very accurately by integrating along the characteristics of the PDE, in this case lines on which $\alpha = y - ct$ are constant. The load of particulate entering the stream is therefore

$$L(t) = A(0,t)h(0,t). \tag{16}$$

4.2 Stream flow

In the second part of the calculation we use the water level and particulate load entering the stream to compute the stream flow and particulate load flowing down to the dam. In this case, the input flow is that from above computed as the sum of the catchment downflow at location $y = 0$ (the stream location). Thus the PDE for the stream flow is

$$S_t - uS_x = [1 - H(x - X)]h(0, t), \quad S(x, 0) = 1\,, \tag{17}$$

where u is the stream flow velocity (assumed constant), and $S(x, t)$ is the flow volume in the stream. The assumption that $S(x, 0) = 1$ is simply to say that before the rainfall event the stream is flowing with unit depth. This does not affect the load entering the dam. The water inflow from the catchment at each location, x, is $Q(t) = h(0, t)$; that is, evaluating Section 4.1 at $y = 0$ gives the flow entering the stream. In this simple model the water inflow is assumed to be the same at each point on the stream, but in principle there is no reason why it could not be a function of x also.

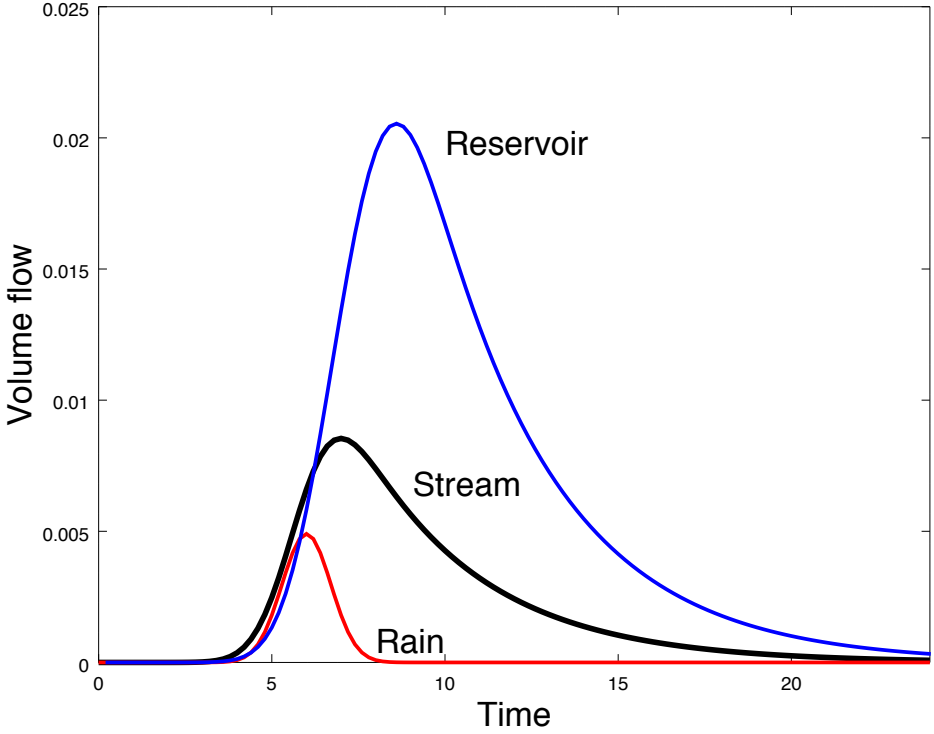
The Heaviside step function is defined as

$$H(x - X) = \begin{cases} 0, & \text{if } x < X\,, \\ 1, & \text{if } x > X\,, \end{cases} \tag{18}$$

where X is the length of the stream within the catchment. This restricts the flow domain to remain within the catchment.

This set of parameters can be estimated in a real situation, but our goal here is to test the concept and so we make what we believe to be reasonable choices to compute the flow and subsequent particulate load in the reservoir. The flow from the stream into the reservoir is computed by solving PDE (17) and then finding $S(0, t)$. Figure 12 shows the time series of flow values for a rainfall event shown with the timing and volumes entering the stream at any point and then the reservoir.

Figure 12: Comparison of magnitude and timing of flows; rainfall, flow entering the stream and flow entering the reservoir for $R_0 = 4 \times 10^{-5}$, $\sigma_Y = 1.75$, $\sigma_T = 1$, $T = 6$, $v = 1.75\,\text{km/hr}$, $u = 4.5\,\text{km/hr}$, $X = 12\,\text{km}$, and $Y = 4\,\text{km}$.



The PDE for the particulate concentration in the stream is

$$(SC)_t + u(SC)_x = [1 - H(x - X)]h(0,t)A(0,t), \tag{19}$$

with $C(x,0) = 0$, $0 < x < X$, $t > 0$, assuming there is initially no particulate in the stream. As above, after rearrangement, this gives

$$C_t + uC_x = [1 - H(x - X)]\left[\frac{L(t) - Q(t)C(x,t)}{S(x,t)}\right]. \tag{20}$$

The final step is to compute the flow and load at the downstream location

$x = 0$, assumed to be the entry to the reservoir. This is achieved again by integrating along the characteristic curves, in this case where $\gamma = x - ut$, subject to the condition that $t < X/u$. If $t > X/u$, then for the first period of time $0 < t < X/u$ there is no water entering the stream, and so integration must then be from $t = X/u$ to the current value of t. The load of particulate entering the reservoir is

$$L_R = S(0,t)C(0,t). \tag{21}$$

Due to the complexity of these calculations, we used the package `Octave` for the calculations.

4.3 Results

It is of interest to compute the stream flow and particulate load at the reservoir entry point. In particular, we are interested in the effect of washout of the particulate from the catchment. Comparing different cases we determine if there is a typical signature to the load of particulate entering the reservoir. The case used here is just a simple example, but does show some interesting effects.

Figure 13 shows the particulate load entering the reservoir for a case with constant amounts of particulate on the slope compared to a case of diminishing amount due to washout. The input to the reservoir starts the same as the flow begins but in the case of washout diminishes as the particulate is flushed off the slopes. In this example all of the particulate is washed out during this rain event. Parameter values were $R_0 = 4 \times 10^{-5}$, $\sigma_Y = 1.75$, $\sigma_T = 1$, $T = 6$, $v = 1.75\,\text{km/hr}$, $u = 4.5\,\text{km/hr}$, $X = 12\,\text{km}$, and $Y = 4\,\text{km}$.

Figure 14 shows an example with the same parameter values as the case above, except that the trickle velocity down the slope has been reduced from $v = 1.75\,\text{km/hr}$ to $1.1\,\text{km/hr}$. All other parameters were kept the same. The load has a slight double bump due to a slight difference in timing of peak flows between the slope runoff and the stream flow.

In this case the lower trickle velocity on the slopes reduces the uptake of

particulate so that after the rainfall event there is still some remaining. Assuming two more (identical) rainfall events over subsequent days, the three solid curves represent the load reaching the reservoir on the three days. At each stage more of the particulate is washed out, leaving a much smaller load to enter the dam. After the third rainfall event, there is no particulate remaining, so in the short term subsequent events does not lead to any particulate reaching the level. However, if there is sufficient time for the particulate levels to rebuild on the slope, then the situation returns to the original. If the particulate is not being washed out, that is there is an unlimited supply on the slopes of the catchment, then the curves for each rain event would look identical.

4.4 Comments

This simple model is limited by its assumptions, but does provide an interesting study of the events that lead to load reaching the reservoir. If there is an unlimited microbe concentration on the slopes, then each event leads to a similar load curve. However, if the microbe concentration is washed away, that is locally depleted, then depending on the rate of uptake, subsequent events may have much reduced load reaching the reservoir. Given the nature of the microbes it is quite likely that the time between events may see a rebuilding of the levels so that the next event is likely to see higher load delivered to the reservoir again. Variations in the stream and slope runoff velocities have quite a significant effect on both the stream flow characteristics and the load delivered to the reservoir. Thus it is important to know the time scales on which a build up of the ground level concentration G occurs in the catchment, so that estimates can be made of build up before subsequent rain events. It is beyond the scope of this study to do a full analysis of the variations in flow, rainfall and load washout, but the simple model does provide a tool for such analysis in the future.

Figure 13: Load of particulate entering the reservoir for undiminished load on the slopes (dashed line), and a case with uptake rate $k_0 = 10$. In this example, all of the particulate is washed out in the single rain event. Other parameters were $R_0 = 4 \times 10^{-5}$, $\sigma_Y = 1.75$, $\sigma_T = 1$, $T = 6\,\text{hr}$, $v = 1.75\,\text{km/hr}$, $u = 4.5\,\text{km/hr}$, $X = 12\,\text{km}$, and $Y = 4\,\text{km}$.

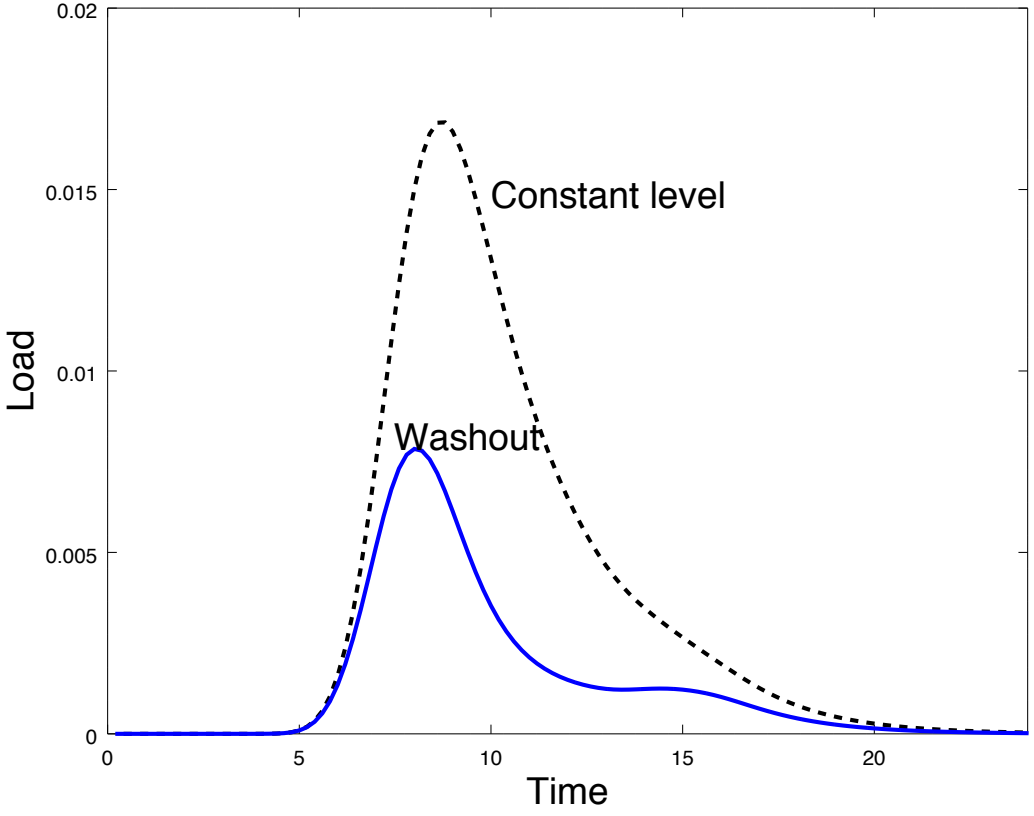
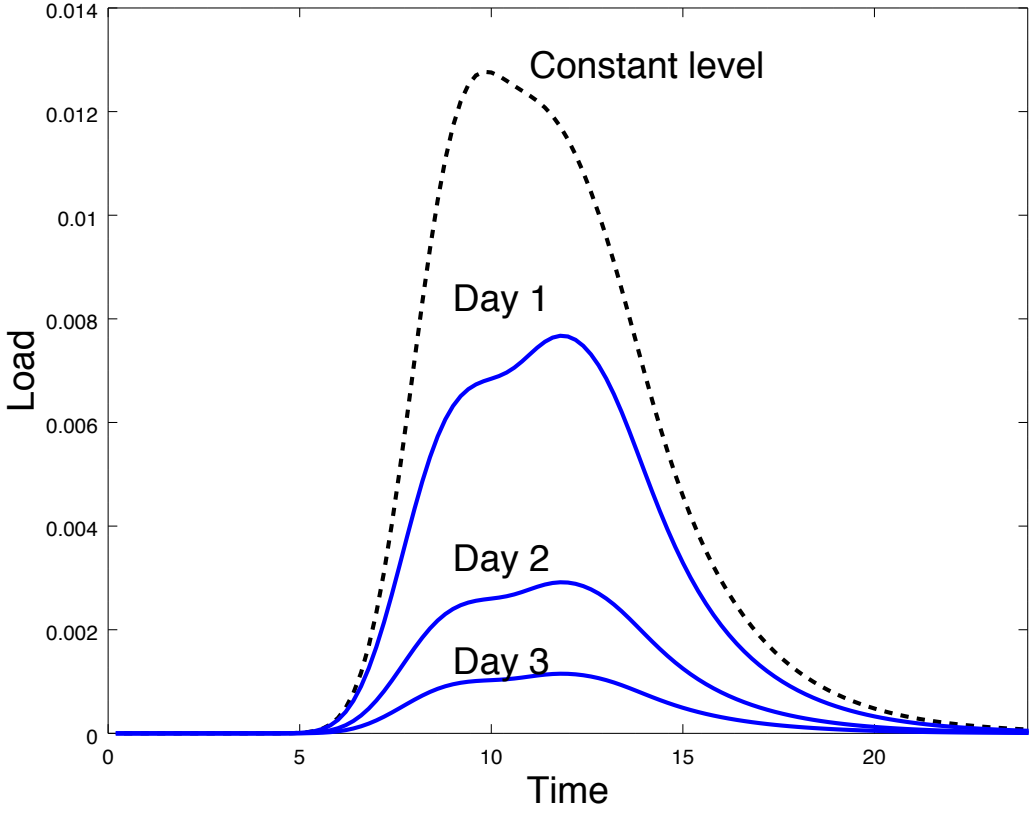


Figure 14: Load of particulate entering the reservoir for undiminished load (dashed line) on the slopes, and a case with uptake rate $k_0 = 10$. Parameters are as in Figure 13 except that $\nu = 1.1\,\mathrm{km/hr}$. Not all particulate is washed out, so the load from subsequent (identical) rainfall events on successive days is shown (time shifted to match).



5 Identifying peak flow events for automated sampling

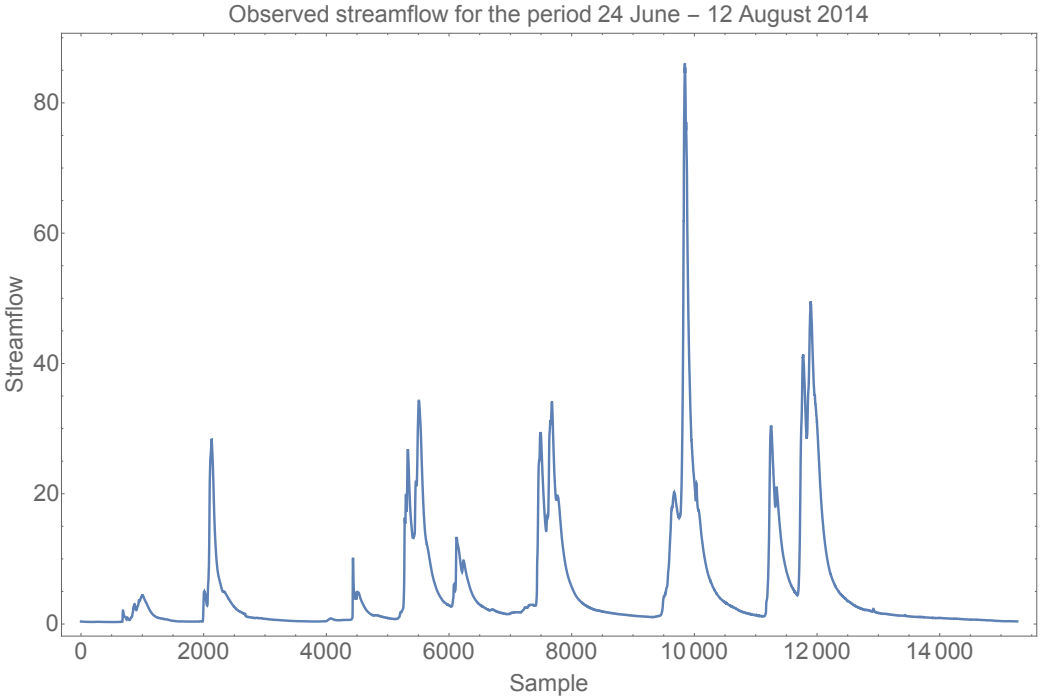
One of the shortcomings in the data collection process has been the difficulty in collecting samples for microbe population assays at the best times. It is desirable to collect samples at the peak of the hydrograph, as SA Water believes this is when the microbe count is likely to be at a maximum. The introduction of automatic sample collection presents the opportunity to refine the sampling process in order to better achieve this goal. An approximate real-time technique to determine the local maxima in the hydrograph is described in this section.

This technique is based on approximating the slope of the hydrograph using discrete time steps. Successive falls in the hydrograph over a given number of time steps is taken as an indication of a local maximum, which is used to trigger the taking of an automatic sample. Strictly speaking, such a local maximum would only be detected after the event; however, if the size of the time step and the number of time steps needed to make a decision are small enough, then this delay in detection will not be significant.

As shown in Figure 15, the hydrograph is noisy, with many localised peaks that are not of interest for sampling purposes. One category of localised peaks that is ignored for sampling purposes are those occurring during low flow events. As interest is in high flow events, a lower streamflow threshold is identified, below which samples are not of interest. This sampling threshold is specific to each catchment and should be identified from the historical record of events and past sampling.

During high flow events, the hydrograph typically contains many localised peaks during both the rise and fall. Localised peaks during the fall of the hydrograph are discounted by ignoring identified peaks corresponding to a lower streamflow than previously sampled. The presently implemented automated sampling method collects point-time samples from a single location,

Figure 15: Observed streamflow for one catchment over the period 21 June to 12 August 2014, with observations at five minute intervals.



which are then collected for processing. The sampler is therefore simply re-set to identify new peaks on collection of previous samples from the site.

One challenge with this approach is the limited capacity of the automated sampler. Once taken, samples cannot be automatically discarded, and therefore the number of samples that can be collected is limited. Attendance at the site is required to empty or replace the sampling buckets before additional samples are taken. It is therefore necessary to balance the duration of the falling streamflow, and hence the delay after a peak where a sample is taken, with the risk of exceeding the capacity of the sampler due to a high number of localised peaks during the rising hydrograph. Examination of the historical

record for each catchment will assist in identifying a suitable number of repeat falls to trigger a sample.

The proposed automated sampling method is illustrated by the pseudo-code:

```
if (d_t ⩾ sampling.threshold and d_t < d_{t−1}) then
    falling.count = falling.count + 1
    if falling.count = falling.threshold then
        if d_t > previous.sample then
            take sample
            previous.sample = d_t
        end
    end
else
    falling.count = 0
end
```

In this code

- d_t is the streamflow at the current timestep,
- sampling.threshold is the minimum streamflow below which samples are not taken,
- falling.count records the number of successive falling streamflow measurements,
- falling.threshold is the number of successive falling streamflow measurements that will trigger a sample, and
- previous.sample is the discharge associated with the previous sample taken.

The above pseudo-code applies to the automated sampler, when applied to historical data it is necessary to include a sampling window, after which the previous.sample is reset to zero.

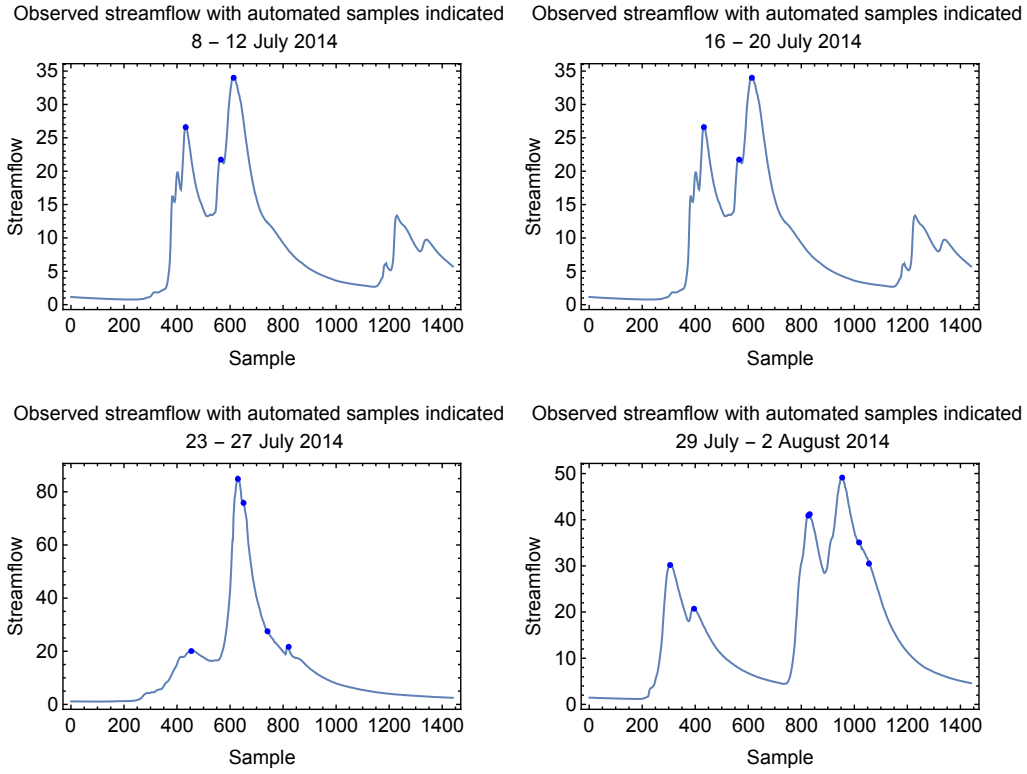
The above method is illustrated on four local peaks observed for one sample

catchment during July 2014, and is shown in Figure 16. Each subfigure shows a five-day window of streamflow observations (in $\mathrm{m\,s^{-1}}$), with the identified sampling times indicated by the points. In these examples two successive falls in the streamflow were used to trigger a sample, with a minimum sampling threshold of $20\,\mathrm{m\,s^{-1}}$. A falling.threshold $= 2$ was chosen based on a manual review of the results of this method on the catchment data. It was found that, given the sampling interval of five minutes, a high falling threshold, for example 5, missed many of the peaks, whereas the chosen threshold of 2 did not lead to an unacceptable number of false positives during the rising hydrograph. All other parameters were chosen in an analogous manner, that is by reviewing the data and trailing different parameter values. As seen in Figure 16, the selected parameter values are effective in identifying the peak, which is the desired time in the hydrograph at which to take a sample. Increasing the sampling window when reviewing historical data, here just one hour, would further reduce false positives during the falling hydrograph.

This method provides a simple means by which automated samples can record close to the peak of the hydrograph, which will hopefully correspond to the peak microbe load. Since the rainfall record and the hydrograph often seem to have the same shape, the pattern of local peaks and troughs should be quite similar between the two. As there is a delay between them, it may be possible to perform an analysis similar to that just mentioned on the rainfall record, and use this to predict beforehand when the peak flows will occur. It would be useful to explore this approach further.

Although the focus here is on collecting samples as close to the peak as possible, automated sampling provides for alternative collection strategies to be explored. For example, in some cases it might be of more interest to obtain a flow rate weighted accumulated microbe count over an entire rain event as this would more accurately represent the number of microbes in the run-off that will accumulate in the reservoir. One possibility might be to collect samples at regular time intervals, and then mix them in proportions determined by the flow rate at the time each sample was taken. This one mixed sample could then be assayed for the microbe count, which could then

Figure 16: Illustrated peak identification for a series of high streamflow events in the month of July 2014 for one catchment. The observed discharge, recorded at five minute intervals, is shown for a five day period in each plot by the solid line, with the indicated sampling locations shown by the blue points. The plots are generated with sampling.threshold = 20, falling.threshold = 2, and with a sampling window for considering previous samples of one hour.



be scaled up by the total flow volume to give an overall microbe load.

6 Conclusions

SA Water captures water from a diversity of sources, including from surface water catchments located in the Adelaide Hills. This region is known to host microbes in the soil, which are transported into the water supply, and if untreated, present a risk to human health. A range of water treatment options are used by SA Water to ensure water supplies are safe, and while it is always possible to use the most extreme treatment options, such an approach is inefficient and costly. Although it is possible it identify the concentration of microbes in a water supply, these laboratory techniques take time and are expensive. It would therefore be advantageous to use alternative characteristics of the water supply that are more easily measurable to supplement the laboratory process in understanding water quality. To this end, SA Water asked the 2016 MISG workshop to investigate connections between observed microbe concentrations and other physically observed properties such as streamflow and turbidity. A variety of approaches were taken to understand the potential connection between microbe concentrations and the observed streamflow properties.

The laboratory technique used to identify the microbe concentration involves adding a known number of marked microbes to the sample, and using the retained number of marked microbes as a proxy for the expected loss of unmarked microbes, under the assumption that marked and unmarked microbes are lost during processing at the same rate. Statistical analysis of this process in Section 2 indicates that where only a small proportion of marked microbes, say 20% or less, are recovered, errors in the unmarked microbe count can typically exceed 20% and thus care should be taken when using the microbe count.

Simple catchment specific log-log regression models for the microbe load

as a function of the streamflow and turbidity are presented in Section 3 together with the coefficients for each of the example catchments. This regression analysis demonstrates that the microbe load is best described by the streamflow and turbidity, with streamflow being the dominant factor in the models. Although the models that are described fitted the available data well in an overall statistical sense, with an adjusted R^2 of around 0.85 for most of the catchments, because of the logarithmic transformations that are involved, the model prediction errors when expressed in absolute terms are large at high flow rates and high microbe loadings (typically around 200–300% for the one-standard deviation error band). Such large absolute errors in part reflect that the various sources of variability which the simple models do not explicitly account for are more likely to show themselves as fixed relative errors in prediction rather than absolute errors. These simple regression models, which only use flow rate or turbidity as the explanatory variables, in effect statistically average over all the other sources of variability that are not explicitly accounted for. Examples of these unaccounted for sources of variability include spatial variability within a catchment, and time effects ranging from the short term during a rain event, to longer term seasonal effects. Whether the simple prediction models presented in Section 3, despite their large error bands, can still be useful in practice as management tools by water utilities such as SA Water requires further investigation. To improve the models' predictive power, further explanatory variables will need to be introduced, presumably addressing the spatial and time factors just mentioned.

Also in Section 3 an initial simple analysis of seasonal effects within the regression model concluded that these effects were not significant. However, this simple analysis only considered an early and late season binary classification as the seasonal variable, and that this had a negligible effect should not rule out the possibility of more subtle seasonal or other time related effects being significant.

The land-based population of microbes is shown in Section 4 to be important for the resulting microbe load that reaches reservoir entry points. Where the

population of microbes is limited, our physical model demonstrates how the profile of microbe concentration, and hence the final concentration in the reservoir, is reduced if the land-based population is limited and subject to washout. Improved information about the land-based lifecycle of the relevant microbes would be useful in understanding the expected microbe concentration in samples after rainfall events, particularly where rainfall is ongoing for some period of time. Such understanding could throw light on whether the maximum microbe concentration tends to occur close to the peak streamflow. Whether or not this holds is potentially an important consideration in deciding upon a water sampling schedule, as discussed briefly below.

In addition to the things considered at MISG, there are several other avenues that might be explored further. One possibility is to consider the size of the reservoir and the physical processes therein, for example, for a larger reservoir with a relatively small inflow, depending on the properties of the inflow, new water is unlikely to reach the outlet in a short time, and consequently the presence or otherwise of particulates may not matter. Models exist that study this behaviour under different conditions. SA Water already has some computational models of the flow within reservoirs. It may be useful to seek to incorporate the transport of microbes and other particulates from input streams into theses models.

One challenge in identifying correlates and predictors of high microbe concentrations has been the number and the timing of the available microbe samples. Currently SA Water takes water samples both routinely, that is according to a schedule, and in response to significant streamflow or rainfall events. However, in most instances, samples are not taken at the peak of the hydrograph, which is when the microbe concentration is widely believed to peak. Section 5 presents an algorithm to detect near-peak streamflow events in real-time. This algorithm has been successful in identifying peaks in some examples taken from the historical record. It could potentially be used to trigger automatic sampling at near-peak streamflow, although in-service trials would be needed to confirm this.

Further research by SA Water is necessary to develop and deploy a predictive model for microbe concentration. The investigations reported here provide a basis to guide future sampling, observation and analysis. Although most of the work described here has been specific to the SA Water catchments for which data was available, the same approaches could be applied to other catchments. Clearly, the various model parameters will be catchment specific, depending on rainfall, weather patterns, catchment topography, the type of land cover and the catchment hydrogeology amongst other factors.

A Percentage points of microbe concentration

Table 6: 5% Percent points of microbe concentration.

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	143	78	54	41	34	28	24	22	19	17
50	307	170	119	91	75	63	55	48	43	39
75	471	263	184	143	117	98	86	76	68	61
100	636	356	250	194	159	134	117	104	93	84
125	800	450	317	245	200	171	149	131	118	107
150	964	542	382	298	244	207	180	160	143	130
175	1131	635	450	349	286	244	212	188	169	154
200	1293	729	515	402	328	280	244	217	195	177
225	1460	822	581	453	372	317	276	245	221	201
250	1623	915	649	504	415	353	308	273	246	225
275	1787	1010	715	555	457	390	340	302	272	249
300	1957	1100	779	608	500	426	372	331	299	272
325	2120	1193	847	660	543	462	404	360	324	296
350	2287	1288	911	712	585	499	436	388	351	320
375	2453	1384	978	763	629	536	469	417	376	344
400	2607	1477	1046	815	672	573	501	445	402	368
425	2780	1570	1113	867	715	610	532	474	429	392
450	2933	1663	1179	920	758	646	565	503	454	416
475	3106	1756	1245	971	800	683	597	533	481	440
500	3279	1850	1311	1024	843	719	629	561	507	464

Table 7: 15% Percent points of microbe concentration.

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	175	93	64	49	39	33	28	25	22	20
50	364	196	134	102	83	70	60	53	47	43
75	554	300	206	158	128	108	93	82	73	66
100	746	400	277	213	173	145	126	111	99	90
125	936	505	349	268	218	183	159	140	125	114
150	1123	609	421	323	262	222	192	169	152	137
175	1317	712	494	378	308	260	225	199	178	161
200	1508	816	565	434	353	298	258	228	204	185
225	1700	919	636	489	398	337	292	258	231	209
250	1886	1021	708	545	443	375	325	287	257	234
275	2077	1126	781	600	489	413	358	317	284	258
300	2269	1229	852	656	535	452	392	346	311	282
325	2458	1330	924	711	579	490	425	376	338	306
350	2654	1435	994	766	624	528	458	405	364	331
375	2843	1541	1068	821	669	567	492	435	391	355
400	3031	1645	1140	877	716	605	525	465	417	379
425	3223	1746	1211	933	760	644	558	494	444	404
450	3400	1848	1284	988	805	681	592	524	471	428
475	3607	1954	1356	1044	850	720	625	554	498	452
500	3800	2058	1426	1100	896	758	659	583	524	477

Table 8: Calculated microbe concentration per ten litres

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	250	125	83	62	50	42	36	31	28	25
50	500	250	167	125	100	83	71	62	56	50
75	750	375	250	188	150	125	107	94	83	75
100	1000	500	333	250	200	167	143	125	111	100
125	1250	625	417	312	250	208	179	156	139	125
150	1500	750	500	375	300	250	214	188	167	150
175	1750	875	583	438	350	292	250	219	194	175
200	2000	1000	667	500	400	333	286	250	222	200
225	2250	1125	750	562	450	375	321	281	250	225
250	2500	1250	833	625	500	417	357	312	278	250
275	2750	1375	917	688	550	458	393	344	306	275
300	3000	1500	1000	750	600	500	429	375	333	300
325	3250	1625	1083	812	650	542	464	406	361	325
350	3500	1750	1167	875	700	583	500	438	389	350
375	3750	1875	1250	938	750	625	536	469	417	375
400	4000	2000	1333	1000	800	667	571	500	444	400
425	4250	2125	1417	1062	850	708	607	531	472	425
450	4500	2250	1500	1125	900	750	643	562	500	450
475	4750	2375	1583	1188	950	792	679	594	528	475
500	5000	2500	1667	1250	1000	833	714	625	556	500

Table 9: 85% Percent points of microbe concentration

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	380	169	107	79	62	52	44	38	34	30
50	743	327	208	152	119	98	84	72	64	57
75	1100	486	307	224	176	144	122	106	94	84
100	1467	644	407	295	232	190	161	140	123	110
125	1829	800	504	368	288	236	200	173	153	137
150	2186	960	604	439	343	282	239	206	182	163
175	2557	1117	704	511	400	328	277	240	211	189
200	2914	1272	804	583	455	373	316	273	240	215
225	3286	1433	900	654	511	419	354	306	270	241
250	3629	1587	1000	725	567	464	393	340	299	267
275	3988	1750	1100	797	623	510	431	373	328	292
300	4357	1906	1200	869	679	557	470	405	357	318
325	4714	2062	1296	942	735	602	508	439	386	344
350	5086	2218	1396	1014	790	647	546	472	415	369
375	5457	2380	1496	1085	847	693	584	505	444	395
400	5829	2533	1596	1156	902	739	623	538	472	421
425	6183	2693	1692	1226	958	783	662	571	501	446
450	6514	2856	1792	1303	1013	829	700	604	531	472
475	6886	3012	1892	1371	1070	875	738	637	560	498
500	7229	3169	1988	1442	1126	921	776	671	588	523

Table 10: 95% Percent points of microbe concentration

Original unmarked microbe count	Marked microbe count out of 100									
	10	20	30	40	50	60	70	80	90	100
25	500	200	126	91	71	58	49	43	38	34
50	980	387	238	170	133	108	91	79	70	62
75	1443	573	348	250	193	157	132	115	101	90
100	1917	757	461	329	253	206	173	149	131	117
125	2383	940	571	406	314	255	214	184	162	144
150	2860	1127	683	484	373	304	255	219	192	171
175	3340	1312	795	563	434	352	295	254	222	197
200	3820	1493	905	641	493	400	335	288	252	224
225	4300	1677	1016	719	555	448	375	323	282	250
250	4780	1850	1125	797	613	496	416	357	312	277
275	5240	2050	1233	875	674	545	456	392	342	303
300	5760	2223	1348	955	734	593	497	425	371	329
325	6220	2415	1457	1032	793	641	536	461	401	355
350	6740	2593	1568	1112	855	689	577	495	431	382
375	7200	2783	1683	1188	913	738	616	529	461	407
400	7720	2964	1791	1266	974	787	658	563	491	434
425	8200	3143	1900	1344	1033	835	698	597	520	460
450	8620	3336	2012	1427	1093	882	738	631	550	485
475	9160	3529	2118	1500	1154	931	778	665	580	511
500	9660	3700	2233	1577	1212	980	818	700	609	537

References

[1] BTF. *ColorSeed: Revolutionary internal standard for Cryptosporidium and Giardia testing.* http://btfbio.com/products/colorseed/technical-articles/ (cit. on p. M72).

[2] Justin D. Brookes et al. "Relative Value of Surrogate Indicators for Detecting Pathogens in Lakes and Reservoirs". In: *Environ. Sci. Technol.* 39 (2005), pp. 8614–8621 (cit. on p. M71).

[3] L. D. Brown and L. H. Zhao. "A test for the Poisson distribution". In: *Sankhya: The Indian Journal of Statistics* 64 Series A, Pt. 3 (2002). http://www-stat.wharton.upenn.edu/~lzhao/papers/MyPublication/Newtest_Sankhya_2002.pdf, pp. 611–625 (cit. on p. M74).

[4] K. F. Cann et al. "Systematic review: extreme water-related weather events and waterborne disease". In: *Epidemiol Infect* 141.4 (2013), pp. 671–86 (cit. on p. M69).

[5] Frank C Curriero et al. "The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994". In: *American journal of public health* 91.8 (2001), pp. 1194–1199 (cit. on p. M69).

[6] D. S. Francey et al. "Effects of Seeding procedures and Water Quality on Recovery of *Crytosporidium* Oocysyts from Stream Water by U.S. Environmental Protection Agency Method 1623". In: *Applied and Environmental Microbiology* 70.7 (2004), pp. 4118–4128 (cit. on p. M72).

[7] Gordon Nichols et al. "Rainfall and outbreaks of drinking water related disease and in England and Wales". In: *Journal of Water and Health* 7.1 (2009), pp. 1–8. ISSN: 1477-8920. DOI: 10.2166/wh.2009.143. (Cit. on p. M69).

[8] Jerry E. Ongerth. "The concentration of *Crytosporidium* and *Gardia* in water—The role and importance of recovery efficiency". In: *Water Research* 47 (2013), pp. 2479–2488. DOI: 10.1016/j.watres.2013.02.015 (cit. on p. M70).

[9] R. S. Signor et al. "Quantifying the impact of runoff events on microbiological contaminant concentrations entering surface drinking waters". In: *Journal of Water and Health* 3.4 (2005). DOI: 10.2166/wh.2005.052 (cit. on p. M71).

[10] Brooke A. Swaffer et al. "Investigating source water *Crytosporidium* concentration, species and infectivity rates during rainfall-runoff in a multi-use catchment". In: *Water Research* 67 (2014), pp. 310–320. DOI: 10.1016/j.watres.2014.08.055 (cit. on pp. M71, M81).

[11] Sanford Weisberg. *Applied Linear Regression*. Wiley Series in Probability and Mathematical Statistics. ISBN:0-71-04419-9. New York: John Wiley and Sons, 2004 (cit. on pp. M83, M85).

[12] Wikipedia. *Binomial distribution*. https://en.wikipedia.org/wiki/Binomial_distribution (cit. on p. M80).

Author addresses

1. **Tony Miller**, Flinders University
mailto:tony.miller@flinders.edu.au
orcid:0000-0003-2646-8034

2. **Melanie E. Roberts**, IBM Research - Australia.
mailto:melanie.roberts@au1.ibm.com
orcid:0000-0003-4027-9651

3. **Brooke A. Swaffer**, SA Water
mailto:brooke.swaffer@sawater.com.au

4. **Graeme Hocking**, Murdoch University
mailto:G.Hocking@murdoch.edu.au
orcid:0000-0002-5812-6015

5. **Bill Whiten**, 4 Magnet Cl, Riverhills 4074, Australia.
mailto:billwhiten@tpg.com.au
orcid:0000-0002-9778-3632

6. **Robert McKibbin**, Massey University
mailto:R.McKibbin@massey.ac.nz