Methods for assessing long-term mean pathogen count in drinking water and risk management implications

James D. Englehardt, Nicholas J. Ashbolt, Chad Loewenstine, Erik R. Gadzinski and Albert Y. Ayenu-Prah Jr

ABSTRACT

Recently pathogen counts in drinking and source waters were shown theoretically to have the discrete Weibull (DW) or closely related discrete growth distribution (DGD). The result was demonstrated versus nine short-term and three simulated long-term water quality datasets. These distributions are highly skewed such that available datasets seldom represent the rare but important high-count events, making estimation of the long-term mean difficult. In the current work the methods, and data record length, required to assess long-term mean microbial count were evaluated by simulation of representative DW and DGD waterborne pathogen count distributions. Also, microbial count data were analyzed spectrally for correlation and cycles. In general, longer data records were required for more highly skewed distributions, conceptually associated with more highly treated water. In particular, 500–1,000 random samples were required for reliable assessment of the population mean ±10%, though 50–100 samples produced an estimate within one log (45%) below. A simple correlated first order model was shown to produce count series with 1/*f* signal, and such periodicity over many scales was shown in empirical microbial count data, for consideration in sampling. A tiered management strategy is recommended, including a plan for rapid response to unusual levels of routinely-monitored water quality indicators.

Key words | correlated, discrete, monitoring, sampling, scaling, Weibull

James D. Englehardt (corresponding author)
University of Miami,
PO Box 248294,
Coral Gables, FL 33124-0630,
USA
E-mail: *jenglehardt@miami.edu*

Nicholas J. Ashbolt
USEPA Office of Research and Development,
26 West Martin Luther King Drive,
Mail Code: 593,
Cincinnati, OH 45268,
USA

Chad Loewenstine
US Department of Justice,
ERF, BLDG 27958-A,
Quantico, VA 22135,
USA

Erik R. Gadzinski
HJ Foundation, 1385 NW 80th St.,
Miami, FL 33166,
USA

Albert Y. Ayenu-Prah Jr
CDM, 1515 Poydras St.,
Suite 1350, New Orleans, LA 70112,
USA

INTRODUCTION

When pathogens are directly measured in typical 1–100 L samples of treated drinking water, most samples show a zero count. Of course, disease burden due to drinking water is not zero (Craun *et al.* 2010) for several reasons. First, analytical recovery rates are generally well below 100% (Xiao *et al.* 2006; Petterson *et al.* 2007). Second, microbes are fundamentally discrete, so that a concentration of 3×10^{-7} infectious units per liter of homogeneous water (corresponding to the original goal of the US Environmental Protection Agency Surface Water Treatment Rule [US Environmental Protection Agency 1989]) would correspond to ≤ 3 in 10,000 samples containing a count, for a 1,000 L sample size. Third, drinking water at a point in time is not homogeneous with respect to organisms that tend to clump, including enteric viral, bacterial, and parasitic protozoan pathogens (particularly in the presence of coagulants, and in pipelines high in iron or copper oxides or biofilm) (Gale 1997, 2002; Cizek *et al.* 2008; Helmi *et al.* 2008). Finally, a treatment plant operating reliably under a constant set of protocols will nevertheless produce fluctuations in water quality of varying magnitude that include the more rare but more important upset conditions (Englehardt *et al.* 2009; Hijnen & Medema 2010; Englehardt & Li 2011).

Available drinking water data indicate that the majority of pathogens are released in occasional large releases, stemming for example from rain events (Signor *et al.* 2007), treatment system overload with short-term loss of treatment/disinfection efficacy (Smeets *et al.* 2008; Hijnen &

doi: 10.2166/wh.2012.142

Medema 2010), treatment system failures, and contamination during distribution (Hunter *et al.* 2005; Nygård *et al.* 2007). An example is the Milwaukee *Cryptosporidium* incident in 1993, in which >50 people died (Besner *et al.* 2011) and various sequelae were identified (Naumova 2003). The rare occurrence of such large events results in count distributions that are extremely right-skewed, with tails ranging or *scaling* over orders of magnitude over the long-term.

Scaling data are difficult to represent with most parametric distributions, because the rare events accounting for the great majority of impact are not often represented in available data. Therefore, while a distribution may be flexible enough to fit the available data (Gonzales-Barron *et al.* 2010; Francis *et al.* 2012), there may be little basis for extrapolating results to the extremely large events of interest unless a particularly long record is on hand. In fact, actuaries have long known the large number of data points required to estimate parameters of such distributions. For example, to estimate the parameter of a single-parameter Pareto II distribution within ±5%, 1,165 data points are required for with 90% confidence, 1,655 are required for 95% confidence, and so on (Reichle & Yonkunas 1985). Thus, to estimate the Dutch annual risk target of one infection per 10,000 people, some 500 samples of 2,000 L each might be needed (Teunis *et al.* 1997), and infrequent regular sampling may contribute little to risk reduction (Signor & Ashbolt 2006).

Theory can provide some basis for extrapolation beyond the range of the data. Theoretically, highly skewed distributions of non-negative outcome size are produced by multiplicative (non-linear) processes, such as first order and pseudo-first order mathematical growth processes (Englehardt & Li 2011). For example, the size of high-count events in a lake might result from a confluence of preceding causes, such as a nearby public event in the water shed, coupled with high temperatures, coupled with extreme precipitation, coupled with a lake 'turnover' providing high levels of nutrients. The sizes of such preceding causes are not additive, but are considered to act multiplicatively on the final outcome size. In multiplicative physical systems, extremes become disproportionately more extreme, producing non-negative distributions that scale. In contrast, linear (additive) physical processes tend to produce bell-shaped distributions of outcome size, positive and negative in size, as given by the central limit theorem. For example, the water level in a lake as a result of several precipitation and drainage events may have a normal distribution of uncertainty and/or variability. When linear (additive) models have been applied to explain scaling data, the sometimes incongruous results have included distributions of outcome size having zero (small) probability of zero (small) outcome sizes (e.g., the lognormal distribution); infinite variance (the Lévy distribution); and the negative-positive range, mean zero, of fractional Brownian motion models (Mandelbrot 1968).

Distributions of microbial counts in drinking and source waters were recently shown theoretically to have the discrete Weibull (DW) distribution, or the closely related discrete growth distribution (DGD) (Englehardt *et al.* 2009; Englehardt & Li 2011), assuming largely correlated first order (multiplicative) mathematical growth of outcome size from preceding causes. This result was demonstrated empirically, based on nine short-term empirical data sets and three simulated long-term data sets. Empirical fit was preferred over the negative binomial and Poisson lognormal (Masago *et al.* 2006) distributions. While continued empirical demonstration is needed, to our knowledge the DW and DGD are the only distributions or compound distributions offering theoretical basis for extrapolation to the important high count events not likely represented in typical datasets. However, the question remains as to how such distributions should be estimated from available data.

The aims of this paper are to: (a) evaluate methods of characterizing long-term mean microbial count in environmental samples, and associated data requirements, (b) understand the effect on this assessment of correlation and periodicity in drinking water microbial count data, and (c) suggest appropriate approaches to public health management. Methods of assessing long-term mean count are evaluated by simulation. Then simulated and field data on microbial counts in water are analyzed spectrally for evidence of long and short-term correlation and periodicity in counts, as a basis for the design of sampling plans. Finally, implications for public health management are discussed.

DEFINITIONS AND METHODS

In this paper, the term distributional *form* refers to the continuous or discrete, truncated or un-truncated, mathematical form

of a probability distribution having the same cumulative distribution function (CDF) subject to change in parameter values and location over a unit interval. A mean is the population (arithmetic) mean, unless specified as a sample mean, and average refers to the sample (arithmetic) average. The term $1/f$ noise refers specifically to a time series of outcome sizes (e.g., pathogen counts) for which spectral power, or squared magnitude of the Fourier transform, is roughly proportional to the inverse of frequency. The term record refers to a series of count values observed in time or space (either simulated counts, or counts measured in water samples), and the record length is the number of count values in the record. Finally, to scale is to range over orders of magnitude.

To evaluate methods of assessing long-term mean count, synthetic water quality data sets were simulated from DW and DGD distributions representative of those observed in drinking and source water in Matlab®, S-PLUS, and C++. Monte Carlo simulation was by analytical (Englehardt & Li 2011) and numerical inversion of the CDFs, respectively.

The DW has closed form probability density function (PDF) and CDF, a strong practical advantage over the DGD, and can be written (Nakagawa & Osaki 1975):

$$p(v) = q^{v^\eta} - q^{(v+1)^\eta}$$
$$P(v) = 1 - q^{(v+1)^\eta},\ 0 < q < 1, \quad 0 < \eta, \quad v = 0,\ 1,\ 2,\ \ldots,\ \infty \tag{1}$$

In Equation (1): v is the count (e.g., number of viable organisms in a water sample); $p(v)$ is the probability mass function (PMF) of the discrete variate, v; $P(v)$ is the CDF; and q and η are shape parameters. Conceptually, η is related to the number of causes of pathogen counts (e.g., number of treatment stages), with small values corresponding to more numerous causes and higher skew of the distribution. For example, more highly treated water might have smaller η, indicating higher count variability though perhaps lower mean count. The mean of the DW defined on the set {0, 1, 2, …} was given incorrectly by Nakagawa and Osaki and others, but can be found as a sum from $v = 1$, as follows (Englehardt & Li 2011):

$$E(v) = \sum_{v=1}^{\infty} q^{v^\eta} \cong \sum_{v=1}^{M} q^{v^\eta} + \frac{\Gamma\left[1/\eta,\ (M+1)^\eta(-\ln q)\right]}{\eta(-\ln q)^{1/\eta}} \tag{2}$$

in which M is a large integer. In previous work a value of $M = 1{,}000$ provided a convergent value for the mean.

The DGD can be written (Englehardt *et al.* 2009):

$$p(v) = \frac{b^{v^\beta}}{D_{q,\beta}}, \quad D_{b,\beta} = \sum_{v=0}^{\infty} b^{v^\beta},\ 0 < b < 1, \quad 0 < \beta,$$
$$v = 0,\ 1,\ 2,\ \ldots,\ \infty \tag{3}$$

in which b and β are shape parameters; β, like η, represents higher skew and conceptually more numerous causes of counts. The normalizing constant, $D_{b,\beta}$, can be computed as:

$$D_{b,\beta} \cong \sum_{v=0}^{M} b^{v^\beta} + \frac{\Gamma\left[1/\beta,\ (M+1)^\beta(-\ln b)\right]}{\beta(-\ln b)^{1/\beta}}. \tag{4}$$

The mean of Equation (3), equivalent to the concentration, d, of a pathogen, is:

$$d = \sum_{v=0}^{\infty} \frac{v b^{v^\beta}}{D_{b,\beta}} \cong \left[\sum_{v=0}^{M} \frac{v b^{v^\beta}}{D_{b,\beta}}\right] + \frac{\Gamma\left[2/\beta,\ (M+1)^\beta(-\ln b)\right]}{\beta(-\ln b)^{2/\beta} D_{b,\beta}} \tag{5}$$

Because there are no known parametric distributions for the mean of the DW or DGD, confidence intervals were estimated numerically, as follows. First, four parameter sets, having η values of 0.2, 0.3, 0.5, and 1.0, were selected for the DW as representative of waterborne microbial count data (Englehardt & Li 2011). All datasets had a population mean of 10 counts, representing a typical microbial concentration used for laboratory analysis. Sets of 100 data records of varying lengths were simulated for each representative parameter pair, and the number of sample averages out of 100 to fall within ±5, ±10, and ±50% of the population mean were recorded. Finally, the record lengths corresponding to 95 of 100 records meeting this tolerance were determined for each tolerance level and each selected parameter pair by linear interpolation.

Alternatives to the sample average for estimating sample means were evaluated for possible improvements in estimation of the population mean of scaling data. The general approach involved fitting the DW (DGD) to the simulated DW (DGD) data, and finding the sample mean from the fitted parameters for comparison with the mean

of the simulated distribution. In particular, two methods of estimating parameters were evaluated: methods of moments and maximum likelihood estimation. In addition, a third method was evaluated for estimation of DGD parameters, as proposed previously (Englehardt *et al.* 2009). The latter method is based on the fact that the DGD normalizing constant can be estimated from the fraction of observations having zero value, as $D_{b,\beta} = 1/p(0)$. Then, Equation (3) can be linearized by the following double log transformation:

$$\log\{-\log[p(v)/p(0)]\} = \beta \log(v) + \log[-\log(b)]. \quad (6)$$

Thus, parameters β and b of the DGD can be estimated from the slope and intercept of the empirical PDF by least squares linear regression, analogous to the estimation of slope parameters of power laws such as the Pareto (Johnson *et al.* 1994).

Record lengths needed to adequately assess the population mean by each method were evaluated using an alternative to confidence intervals, as follows. First, 100 records of each selected record length were generated from DW and DGD distributions having parameter values representative of waterborne pathogen count data. Specifically, values of η and $\beta = 0.2$, 0.3, 0.5, and 1, for the DW and DGD, respectively, were selected. Then, each value of η or β was paired with three appropriate q or b values to represent the range of microbial concentrations measured by standard laboratory protocols, whereby sample size is adjusted so that counts measured in any sample fall in the range ~1–1,000.

Sample means were computed for all datasets by the methods just described. Then, for each set of 100 records, the average, median, and standard deviation of the 100 estimated means and fitted η and β values were computed. Also, the record having the highest-valued DW or DGD likelihood function among each set of 100 records was identified as the most likely record, and population mean estimates obtained for those records were recorded as the *most likely* estimate for each distribution, parameter set, method, and record length. That is, the likelihood function of a sample of observed values 'gives the relative likelihood of having observed this particular sample … as a function of [the parameter vector]' (Benjamin & Cornell 1970). The data used in this task were simulated from known parameter vectors. Therefore the set of 100 samples with highest likelihood of those simulated from a particular parameter vector was identified as the most likely single dataset to be observed. The average of this record was found accordingly as the most likely average to be assessed given a single available record (of length 10–10,000) of microbial counts, e.g., from a treatment plant.

Finally, count data simulated with the correlated multiplicative model described previously (Englehardt & Li 2011) and microbial count data reported for *Giardia* in protected source water from the Kensico Reservoir, Catskill Lower Effluent Chamber (New York City Department of Environmental Protection 2006) were analyzed spectrally by spline-interpolation and fast Fourier transformation in Matlab® version R2006a. Data from January 2002 through July 2006 only were selected, because these were measured consistently for a 50 L sample volume by the same microbiological analysis technique (US Environmental Protection Agency Method 1623 HV). Also, the same data were analyzed by an empirical non-linear alternative to Fourier analysis, termed empirical mode decomposition (Huang *et al.* 1998), using a Matlab® program (Rilling *et al.* 2003).

RESULTS

The most standard way to evaluate estimates of a mean is with confidence intervals. As shown in Figure 1, extremely long data record lengths were indicated to be needed for reliable assessment of the long-term mean, except when allowing errors of ±50% for datasets having large values of η (small skew). However, confidence intervals on tolerance bands were considered difficult to interpret in terms of sampling plan guidance.

To compare alternate methods of assessing population means, means assessed from the most likely datasets were compared, as described previously. These datasets represented those most likely to be obtained when, as is generally the case, only one record is available. Selected results of simulations are plotted in Figures 2–4, and complete results are tabulated in the Appendix (available online at http://www.iwaponline.com/jwh/010/142.pdf). The average and median values of the estimated population means shown in Tables A.1 and A.2 of the Appendix

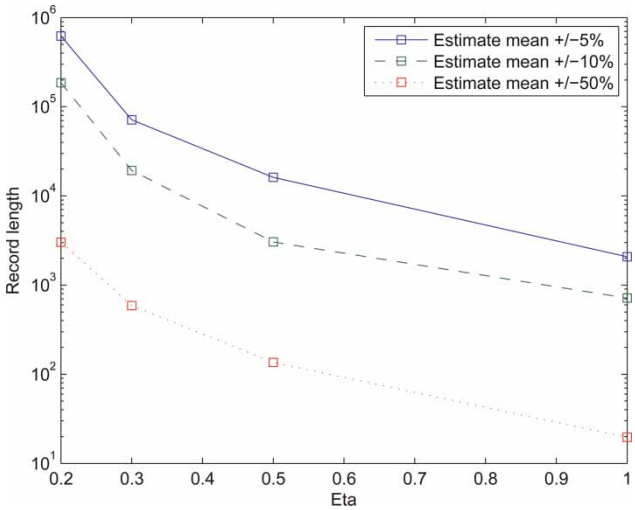

Figure 1 | Required record length to estimate DW mean to within ±5, ±10, and ±50%, with 95% confidence, versus value of η. All data are for constant DW mean = 10 counts. (Parameter values tested: $\eta = 1.0$, $q = 0.91$; $\eta = 0.5$, $q = 0.65$; $\eta = 0.3$, $q = 0.38$; $\eta = 0.2$, $q = 0.19$).

(http://www.iwaponline.com/jwh/010/142.pdf) are close to the population means, only because those values are measures of central tendency across 100 records, effectively representing record lengths of 1,000–1,000,000 samples. As shown in the figures, the most likely estimates of the population mean given a single data record increased significantly with the length of the data record. For population means on the order of one count, 50–100 samples were adequate. However, for population means on the order of 10 or greater, 500–1,000 samples were required for reliable estimation of the population mean within ±10%.

Of note, maximum likelihood estimator (MLE)-fitted means averaged across each 100 records were consistently higher than both the corresponding population mean and the average sample average, for short data records and small η or β, as shown in the Appendix. This unexpected result suggests that the MLE method is more sensitive

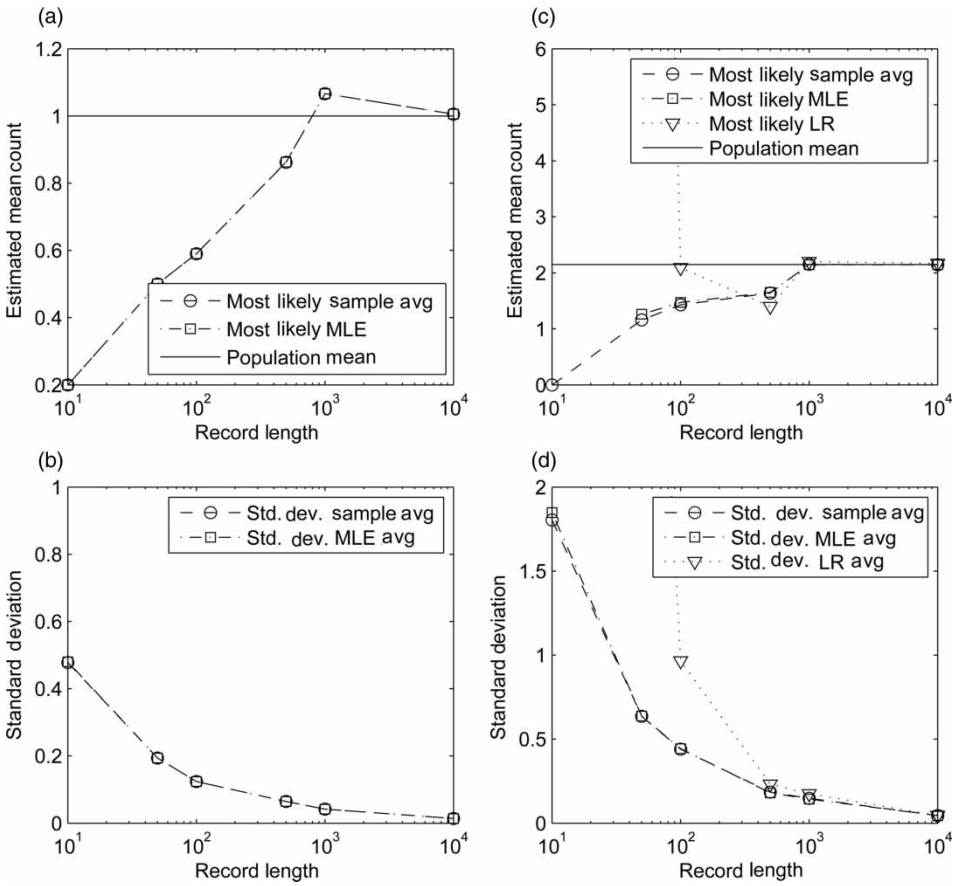

Figure 2 | Most likely sample mean estimated by sample average, maximum likelihood (MLE), and linear regression (LR), for the (a) DW ($\eta = 1.0$, $q = 0.5$) and (c) DGD ($\beta = 0.5$, $b = 0.27$), and standard deviations of 100 sample means for the same (b) DW and (d) DGD, versus record length.

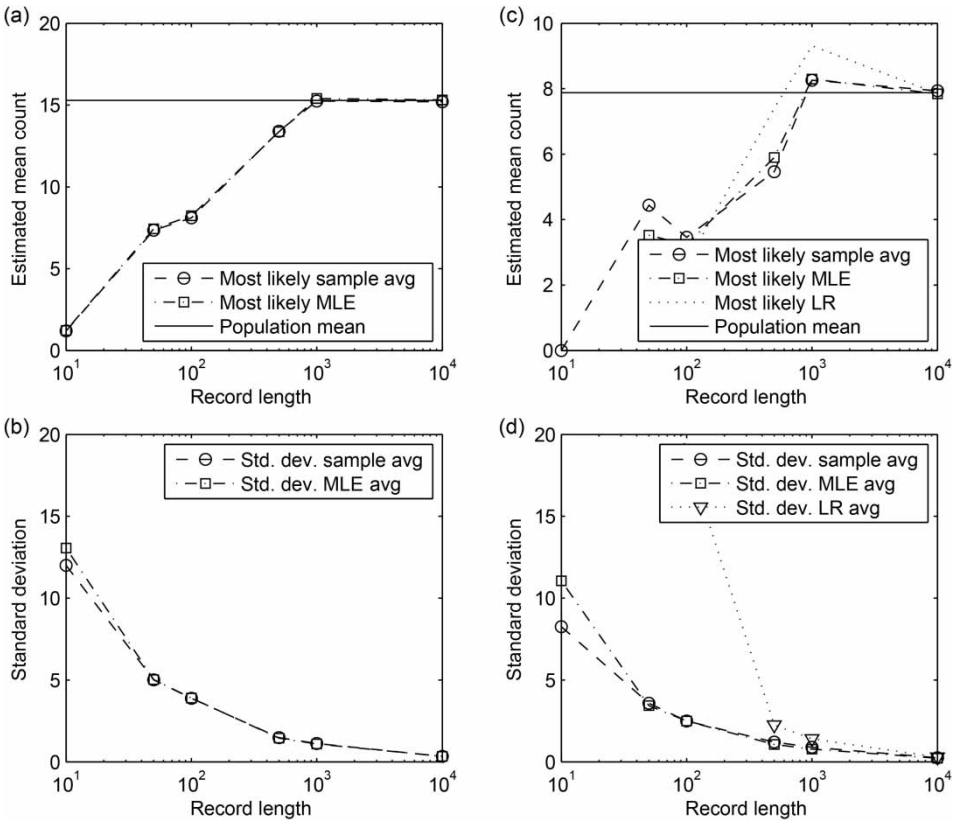

Figure 3 | Most likely sample mean estimated by sample average, maximum likelihood (MLE), and linear regression (LR), for the (a) DW ($\eta = 0.5$, $q = 0.7$) and (c) DGD ($\beta = 0.3$, $b = 0.14$), and standard deviations of 100 sample means for the same (b) DW and (d) DGD, versus record length.

than data average to the effect of uncharacteristically high counts that may be observed in particular data records. In Figures 2(b) and 2(d), 3(b) and 3(d), and 4(b) and 4(d), standard deviations of the 100 sample averages estimated by each method are shown, to depict the higher variability of the MLEs. The linear regression technique produced even less stable estimates of the population mean, apparently being highly sensitive to data noise.

Sampling frequency and duration: pathogen count cycles

Beyond record length, any cycling in counts over time would need to be considered when sampling. Seasonal and daily cycles in source water quality are known, and many natural processes show a $1/f$ power spectrum (Bak 1996). Such a spectrum signals correlation among counts decaying gradually with increasing time lag (Hooge & Bobbert 1997), suggesting cycling (or other repeated pattern) on many time-scales. Therefore, we looked for the $1/f$ spectrum in the series of counts produced by the simple first order (multiplicative) model used to derive the DW and DGD. In particular, a count series was simulated as the product of two exponential random variables. To introduce inter-cause (intra-count) correlation, the mean of the second cause was set equal to the geometric mean of the first cause and a new standard exponential variate. The distributions of these products fit the DW and DGD (Englehardt *et al.* 2009; Englehardt & Li 2011), and did not fit the Poisson, due to inter-cause correlation. However, because of the lack of correlation among final counts, the series exhibited power spectra with log-log slope near zero indicating random white noise (data not shown).

In a source water or treatment plant, there is no reason to expect that temporal correlation among causes of counts does not extend among, as well as within, the counts

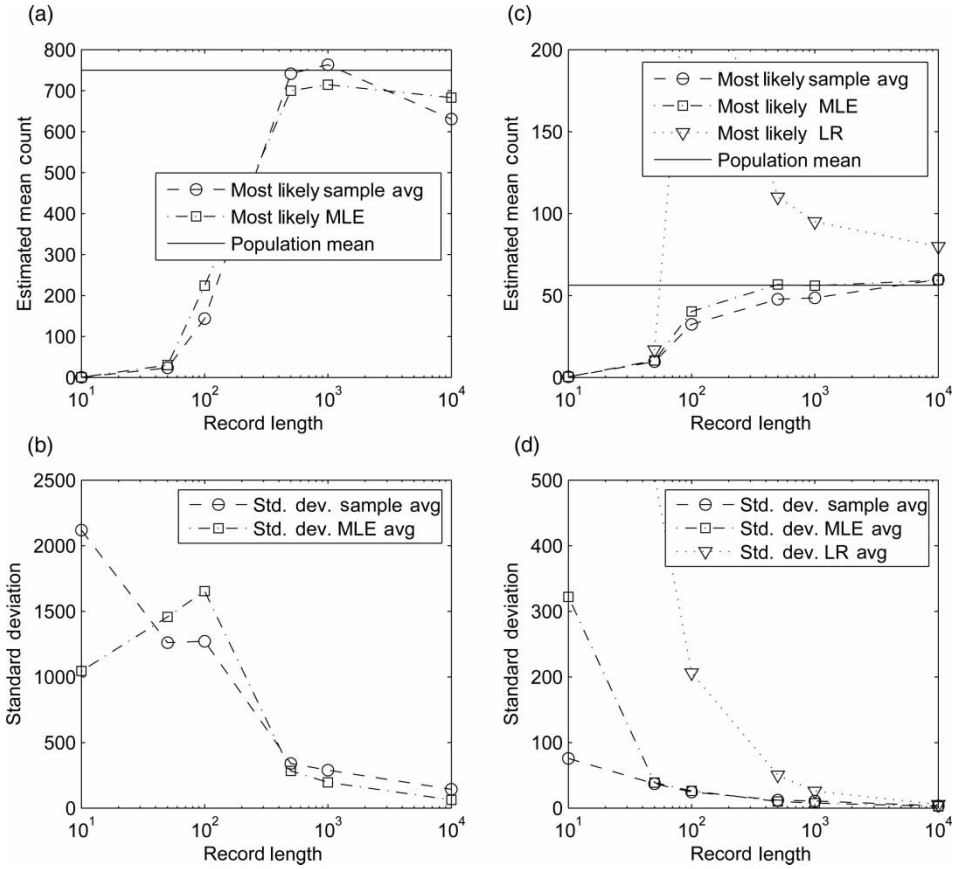

Figure 4 | Most likely sample mean estimated by sample average, maximum likelihood (MLE), and linear regression (LR), for the (a) DW ($\eta = 0.2$, $q = 0.5$) and (c) DGD ($\beta = 0.2$, $b = 0.07$), and standard deviations of 100 sample means for the same (b) DW and (d) DGD, versus record length.

produced. Moreover, correlation among counts might be expected to decay gradually with increasing time lag, therefore appearing as $1/f$ noise in the power spectrum of the count series. In fact, the same correlated first order model underlying the DW and DGD was shown to produce count series with $1/f$ spectra, when correlation among counts (products) was introduced. Specifically, the first of the two causes (except in the first product) was set equal to the second cause of the previous count, introducing inter-count correlation. Resulting power spectra had log-log slopes averaging −1.2038 with standard deviation 0.4790, over 1,000 trials, for frequency range 0.1–0.5 (periods of 2–10 outcomes). One such $1/f$ spectrum is shown in the inset to Figure 5(b). Propagation of correlation along the series was reflected in the corresponding autocorrelation function (inverse Fourier transform of the power spectrum), which showed positive correlation among products up to about 10 products apart in the series (data not shown). A similar correlated first order process may occur in water, with the range of correlation and $1/f$ character extended further due to multi-year, seasonal, weekly (anthropogenic), and daily temperature and biochemical cycles that represent causes of microbial growth and decay.

To investigate empirical evidence for temporal cycling and $1/f$ noise in *Giardia* counts at one reservoir, available data were analyzed spectrally. The data are seen in a time series in the top panel of Figure 6. The average count over the period is 2.2 cysts, with a decreasing trend averaging 0.24 cysts/year shown in the last panel of Figure 6. The time series data with a linear trend of 0.24 cysts/year removed by difference is shown in the top panel of Figure 5, together with a cubic spline fit. The lower panel of Figure 5 depicts the resulting Fourier power spectrum. Sampling was

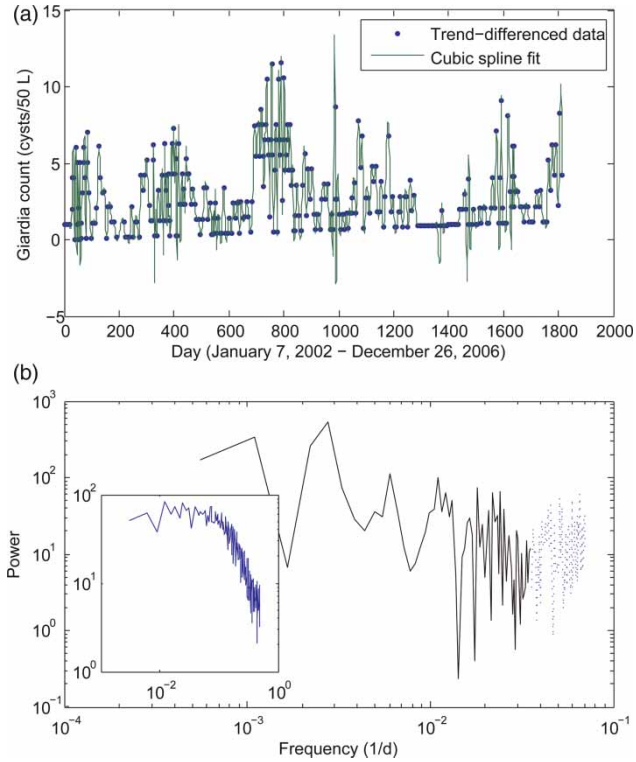

Figure 5 | *Giardia* cyst count data (New York City Department of Environmental Protection 2006): (a) data with long-term trend removed, and spline fit; (b) resulting Fourier power spectrum, showing 1/*f* character with log-log slope −0.92 over the frequency range fairly represented by regular sampling, 5.5×10^{-4}–0.036/d (–), and flatter slope over the higher frequency range 0.036–0.071 (– –). Inset: power spectrum of a simulated series of 327 correlated products of correlated exponential cause sizes, slope −1.319 over frequency range 0.1–0.5/count (wavelength 2–10 simulated counts).

on an irregular weekly schedule, though daily samples were occasionally collected; therefore, the Fourier analysis was considered most meaningful for wavelengths from 5 years (frequency 5.5×10^{-4}/d) to 4 weeks (Nyquist frequency 0.036/d, equal to one half of a regular two-week sampling frequency). Over this range of frequencies, the power spectrum shown has a log-log linear trend with slope −0.9166, classical 1/*f* noise. The corresponding autocorrelation function indicates positive correlation among counts more than 400 days apart (data not shown).

In Figure 6, the issue of sampling frequency irregularity is eliminated through analysis of the intrinsic mode functions, non-linear signal components of the same data found by empirical mode decomposition (Huang *et al.* 1998). As mentioned, the first plot represents the raw data time series. Following that are seven panels containing plots of the seven non-linear cyclic modes found. The first mode (top panel) is noise below the sampling frequency, as is generally reported. The strongest mode is the second (second panel), showing roughly bi-weekly cycling possibly related to the sampling period. The next strongest is the sixth (sixth panel), generally corresponding with the annual water temperature cycle. As mentioned, the last mode shows data non-stationarity (trend over the period), perhaps climate-driven or attributable to land-use changes in the watershed. Overall, modes having average periods ranging from ca. 2 weeks to >5 years emerged, suggesting correlations across all sampled time scales.

DISCUSSION

The long record length shown in this work to be needed to reliably assess the long-term mean of a scaling distribution stems from the fact that the large count values controlling the mean generally appear only in longer records. Use of shorter records may coincidentally produce an accurate assessment if: (a) the unknown distribution is not highly dispersed, or (b) the available dataset is not a highly likely set, such as one focused on events in the watershed or in the treatment plant leading to higher counts. However, such information cannot be known from short records. On the other hand it should be noted that risk is generally evaluated only in orders of magnitude, and in that light it can be seen that 50–100 samples produced an estimate within 1-log (45%) of the true mean, albeit lower than the true mean. It should also be noted that some pathogen datasets may require adjustment for bias, due to low or highly variable analytical recovery rates.

Sampling should ideally be frequent enough to detect high-count transients, and long enough in duration to capture long-term variability. In addition, periodicity such as shown in Figures 5 and 6 should be considered. For example, to avoid consistent sampling in peaks or valleys of count intensity, sampling can be regularly timed, at a frequency not equal to an important annual or other cycle. Estimated long-term mean pathogen density is then computed as the population mean count divided by the analytical sample volume.

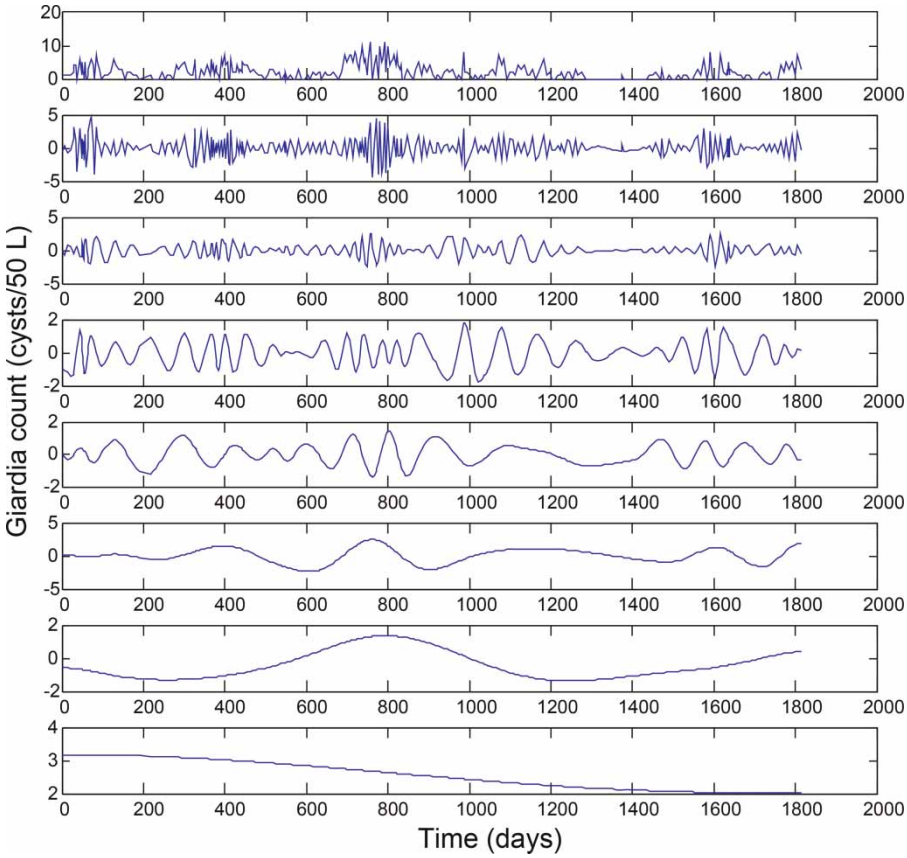

Figure 6 | Intrinsic mode functions, representing non-linear cycles found in the *Giardia* count data of Figure 5.

CONCLUSIONS AND RAMIFICATIONS FOR PATHOGEN MANAGEMENT

Results presented in this paper provide some insight into the use of environmental data for the assessment of pathogen exposure due to consumption of contaminated water. Principal conclusions are:

- To assess the long-term mean microbial count in source water and drinking water, relatively long records of pathogen count data are needed. In particular, 50–100 data points may be needed to reliably estimate the population mean to within 1-log below actual, and 500–1,000 are probably needed for accuracy within ±10%, as microbes are not homogeneous, with counts scaled over orders of magnitude in environmental waters.
- For estimation of population mean from short records, the sample average provides a more stable and accurate estimate than maximum likelihood, with parameter estimation by linear regression being the least reliable.
- Microbial counts in environmental waters cycle on many time scales, as evidenced by a $1/f$ power spectrum and explained by a simple correlated first order model. Ideally, then, water sampling should be: (a) representative across seasonal, daily, and other known water quality cycles; (b) long enough in duration to capture long-term cycles; and (c) frequent enough to capture short-term transients, such as during rain events.

Results of this work coupled with the expense of enteric viral, parasitic protozoan and specific bacterial pathogen analyses suggest the need for cost-effective alternatives for source water and drinking water quality monitoring. This endeavor is further complicated in that in addition to human-infectious pathogens originating from human excreta/sewage, various zoonotic pathogens

(e.g., *Escherichia coli* O157:H7, *Giardia* and *Cryptosporidium* spp.) may also be episodic within animal populations (e.g., Dowd *et al.* 2010; Xiao 2010) and hence, not indexed by traditional fecal indicator bacteria. Therefore, monitoring alternatives may be better derived from the same multiplicative correlated nature of the processes that lead to high-count microbial releases. For example, large pathogen releases may be managed by:

- Breaking the causal chain as it develops, e.g., by switching to alternate water source(s) during higher risk periods (such as rain events or known sewage discharges [Åström *et al.* 2007]).
- Introducing negative correlation to the causal chain, such that downstream protective barriers are proportionate to upstream loadings. This could be undertaken by implementing a plan for rapid response to large sentinel fluctuations in one or more routinely monitored, cost-effective water quality indicators, such as turbidity, flow, total organic carbon or salt ratios that may be indicative of a water quality changed or contamination.

To address improved water quality management the water safety plan/hazard assessment, critical-control-point type of management system has been advocated in many countries and promoted via the World Health Organization (WHO 2009). Specifically, a monitoring scheme designed to notify management personnel of a hazardous event in real-time has been suggested, as a basis for a tiered approach to managing pathogen risks. For example, in source waters impacted by cattle manure, monitoring of cattle water troughs could provide indication of high risk periods from *E. coli* O157:H7 (Ayscue *et al.* 2009), and at water treatment works turbidity, chlorine residual, and pH can be monitored continuously (Nilsson *et al.* 2007; Åström *et al.* 2009) to signal in real-time any need for additional (treatment) response.

The key message from our paper is that pathogen events in drinking water that are important to public health are episodic and short-lived, intrinsically punctuating a much lower equilibrium density. The ramification for improved drinking water management is that improved control of waterborne pathogens, for example in potable reuse systems coming online, is unlikely to be achieved through traditional water sampling plans. Though still required by various regulatory agencies, routine microbiological sampling typically misses (spatially and/or temporally) important fecal indicator and/or pathogen counts (Signor & Ashbolt 2006; Teunis *et al.* 2010). With the reality of limited microbiological sampling, effort would be more productively focused on identifying the likely indicators or surrogates of increased pathogen occurrence. To some degree these indicators are site-specific, but in general relate to knowledge of the water supply system's vulnerabilities to human and important animal species/periods of fecal contamination. Following an understanding of ones' system, sampling for pathogen management should then focus on close to real-time monitoring of those vulnerabilities using simple and inexpensive surrogates.

ACKNOWLEDGEMENTS

This research was supported in part by the US Environmental Protection Agency, National Center for Environmental Assessment, and hosted in part by the University of New Hampshire and the University of Delaware during sabbatical residence. Mary Rothermich, Jeff Swartout, Norden Huang, Jim Nearing, and Nii Attoh-Okine are thanked for their generous help and contributions. The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the US Environmental Protection Agency.

REFERENCES

Ayscue, P., Lanzas, C., Ivanek, R. & Gröhn, Y. T. 2009 Modeling on-farm *Escherichia coli* O157:H7 population dynamics. *Foodborne Pathog. Dis.* **6**, 461–470.

Åström, J., Petterson, S. R., Bergstedt, O., Pettersson, T. J. R. & Stenström, T. A. 2007 Evaluation of the microbial risk reduction due to selective closure of the raw water intake before drinking water treatment. *J. Water Health* **5** (S1), 81–97.

Åström, J., Pettersson, T. J., Stenström, T. A. & Bergstedt, O. 2009 Variability analysis of pathogen and indicator loads from urban sewer systems along a river. *Water Sci. Technol.* **59**, 203–212.

Bak, P. 1996 *How Nature Works: the Science of Self-Organized Criticality*. Copernicus, Springer-Verlag New York, Inc., New York

Benjamin, J. & Cornell, C. A. 1970 *Probability, Statistics, and Decision for Civil Engineers*. McGraw-Hill, New York, 397.

Besner, M. C., Prevost, M. & Regli, S. 2011 Assessing the public health risk of microbial intrusion events in distribution systems: conceptual model, available data, and challenges. *Water Res.* **45**, 961–979.

Cizek, A. R., Characklis, G. W., Krometis, L. A., Hayes, J. A., Simmons 3rd, O. D., Di Lonardo, S., Alderisio, K. A. & Sobsey, M. D. 2008 Comparing the partitioning behavior of *Giardia* and *Cryptosporidium* with that of indicator organisms in stormwater runoff. *Water Res.* **42**, 4421–4438.

Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., Calderon, R. L., Roberts, J. M., Beach, M. J. & Roy, S. L. 2010 Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. *Clin. Microbiol. Rev.* **23**, 507–528.

Dowd, S. E., Crippen, T. L., Sun, Y., Gontcharova, V., Youn, E., Muthaiyan, A., Wolcott, R. D., Callaway, T. R. & Ricke, S. C. 2010 Microarray analysis and draft genomes of two *Escherichia coli* O157:H7 lineage II cattle isolates FRIK966 and FRIK2000 investigating lack of shiga toxin expression. *Foodborne Pathog. Dis.* **7**, 763–773.

Englehardt, J. & Li, R. 2011 The discrete Weibull distribution: an alternative for correlated counts with verification for microbial counts in water. *Risk Anal.* **31**, 370–381.

Englehardt, J., Swartout, J. & Loewenstine, C. 2009 A new theoretical discrete growth distribution with verification for microbial counts in water. *Risk Anal.* **29**, 841–856.

Francis, R. A., Geedipally, S. R., Guikema, S. D., Dhavala, S. S., Lord, D. & LaRocca, S. 2012 Characterizing the performance of the Conway-Maxwell Poisson generalized linear model. *Risk Anal.* **32**, 167–183.

Gale, P., Pitchers, R. & Gray, P. 2002 The effect of drinking water treatment on the spatial heterogeneity of micro-organisms: implications for assessment of treatment efficiency and health risk. *Water Res.* **36**, 1640–1648.

Gale, P., van Dijk, P. & Stanfield, G. 1997 Drinking water treatment increases micro-organism clustering. *J. Water SRT – Aqua* **46**, 117–126.

Gonzales-Barron, U., Kerr, M., Sheridan, J. J. & Butler, F. 2010 Count data distributions and their zero-modified equivalents as a framework for modelling microbial data with a relatively high occurrence of zero counts. *Int. J. Food Microbiol.* **136**, 268–277.

Helmi, K., Skraber, S., Gantzer, C., Willame, R., Hoffmann, L. & Cauchie, H. M. 2008 Interactions of *Cryptosporidium parvum*, *Giardia lamblia*, vaccinal poliovirus type 1, and bacteriophages phiX174 and MS2 with a drinking water biofilm and a wastewater biofilm. *Appl. Environ. Microbiol.* **74**, 2079–2088.

Hijnen, W. A. M. & Medema, G. J. 2010 *Elimination of Micro-organisms by Drinking Water Treatment Processes: A Review*. 3rd edition, IWA Publishing, London.

Hooge, F. & Bobbert, P. 1997 On the correlation function of 1/*f* noise. *Physica B* **239**, 223–230.

Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N.-C., Tung, C. C. & Liu, H. H. 1998 The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R Soc. London, Ser A* **454**, 903–995.

Hunter, P. R., Chalmers, R. M., Hughes, S. & Syed, Q. 2005 Self-reported diarrhea in a control group: a strong association with reporting of low-pressure events in tap water. *Clin. Inf. Dis.* **20**, 32–34.

Johnson, N., Kotz, S. & Balakrishnan, N. 1994 *Continuous Univariate Distributions.* 2nd edition, John Wiley & Sons, Inc., New York, 1, 580.

Mandelbrot, B. B. & Van Ness, J. W. 1968 Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* **10**, 422–437.

Masago, Y., Hiroyuki, K., Watanabe, T., Hashimoto, A., Omura, T., Hirata, T. & Ohgaki, S. 2006 Quantitative risk assessment of noroviruses in drinking water based on qualitative data in Japan. *Environ. Sci. Technol.* **40**, 7428–7433.

Nakagawa, T. & Osaki, S. 1975 The discrete Weibull distribution. *IEEE T. Reliab.* **R-24**, 300–301.

Naumova, E. M. 2003 The elderly and waterborne *Cryptosporidium* infection: Gastroenteritis hospitalizations before and during the 1993 Milwaukee outbreak. *Emerg. Infect. Dis.* **9**, 418–425.

New York City Department of Environmental Protection 2006 *Cryptosporidium* and *Giardia* Monitoring Data, New York City's Water Supply System, *Cryptosporidium* and *Giardia* Background Information and Monitoring Program. Available from: http://home2.nyc.gov/html/dep/html/drinking_water/pathogen.shtml [accessed 22 February 2012].

Nilsson, P., Roser, D., Thorwaldsdotter, R., Petterson, S., Davies, C., Signor, R., Bergstedt, O. & Ashbolt, N. 2007 SCADA data and the quantification of hazardous events for QMRA. *J. Water Health* **5** (Suppl 1), 99–105.

Nygård, K., Wahl, E., Krogh, T., Tveit, O. A., Bøhleng, E., Tverdal, A. & Aavitsland, P. 2007 Breaks and maintenance work in the water distribution systems and gastrointestinal illness: a cohort study. *Int. J. Epidemiol* **36**, 873–880.

Petterson, S. R., Signor, R. S. & Ashbolt, N. J. 2007 Incorporating method recovery uncertainties in stochastic estimates of raw water protozoan concentrations for QMRA. *J. Water Health* **5**, 51–65.

Reichle, K. A. & Yonkunas, J. P. 1985 Discussion of: a practical guide to the single parameter Pareto distribution by Stephen W. Philbrick. *Proc Casualty Actuarial Soc*, 72 (137 & 138), 85–123, Casualty Actuarial Society, Arlington, VA, USA.

Rilling, G., Flandrin, P. & Gonçalvès, P. 2003 On empirical mode decomposition and its algorithms. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (I), June 8 – 11.

Signor, R. S., Ashbolt, N. J. & Roser, D. J. 2007 Microbial risk implications of rainfall-induced runoff events entering a

reservoir used as a drinking-water source. *J. Water Supply Res. Technol. AQUA* **56**, 515–531.

Signor, R. S. & Ashbolt, N. J. 2006 Pathogen monitoring offers questionable protection against drinking-water risks: a QMRA (Quantitative Microbial Risk Analysis) approach to assess management strategies. *Water Sci. Technol.* **54**, 261–268.

Smeets, P. W. M. H., Medema, G. J., Dullemont, Y. J., Van Gelder, P. H. A. J. M. & Van Dijk, J. C. 2008 Case study of *Campylobacter* reduction by filtration and ozonation. *J. Water Health* **6**, 301–314.

Teunis, P., Medema, G., Kruidenier, L. & Havelaar, A. 1997 Assessment of the risk of infection by *Cryptosporidium* or *Giardia* in drinking water from a surface water source. *Water Res.* **31**, 1333–1346.

Teunis, P. F., Xu, M., Fleming, K. K., Yang, J., Moe, C. L. & LeChevallier, M. W. 2010 Enteric virus infection risk from intrusion of sewage into a drinking water distribution network. *Environ. Sci. Technol.* **44**, 8561–8566.

US Environmental Protection Agency 1989 Drinking water; national primary drinking water regulations; filtration, disinfection; turbidity, *Giardia lamblia*, viruses, *Legionella*, and heterotrophic bacteria. *Federal Register* **54**, 27486–27603.

WHO 2009 *Water Safety Plan Manual: Step-by-step risk management for drinking-water suppliers*. World Health Organization, Geneva.

Xiao, L., Alderisio, K. & Jiang, J. 2006 Detection of *Cryptosporidium* oocysts in water: Effect of the number of samples and analytic replicates on test results. *Appl. Environ. Microbiol.* **72**, 5942–5947.

Xiao, L. 2010 Molecular epidemiology of cryptosporidiosis: an update. *Exper. Parasitol.* **124**, 80–89.

First received 27 August 2011; accepted in revised form 15 January 2012. Available online 8 March 2012