SCADA data and the quantification of hazardous events for QMRA

P. Nilsson, D. Roser, R. Thorwaldsdotter, S. Petterson, C. Davies, R. Signor, O. Bergstedt and N. Ashbolt

ABSTRACT

The objective of this study was to assess the use of on-line monitoring to support the QMRA at water treatment plants studied in the EU MicroRisk project. SCADA data were obtained from three Catchment-to-Tap Systems (CTS) along with system descriptions, diary records, grab sample data and deviation reports. Particular attention was paid to estimating hazardous event frequency, duration and magnitude. Using Shewart and CUSUM we identified 'change-points' corresponding to events of between 10 min and > 1 month duration in timeseries data. Our analysis confirmed it is possible to quantify hazardous event durations from turbidity, chlorine residual and pH records and distinguish them from non-hazardous variability in the timeseries dataset. The durations of most 'events' were short-term (0.5–2.3 h). These data were combined with QMRA to estimate pathogen infection risk arising from such events as chlorination failure. While analysis of SCADA data alone could identify events provisionally, its interpretation was severely constrained in the absence of diary records and other system information. SCADA data analysis should only complement traditional water sampling, rather than replace it. More work on on-line data management, quality control and interpretation is needed before it can be used routinely for event characterization.

Key words | hazardous events, QMRA, quality control, SCADA, timeseries

P. Nilsson
R. Thorwaldsdotter
Division of Ergonomics and Aerosol Technology,
Department of Design Sciences,
Faculty of Engineering, Lund University,
PO Box 118, SE-221 00, Lund,
Sweden

D. Roser (corresponding author)
S. Petterson
C. Davies
R. Signor
N. Ashbolt
School of Civil and Environmental Engineering,
University of New South Wales,
Sydney NSW 2052,
Australia
Tel.: +61 2 9385 5137;
Fax: +61 2 9313 8624;
E-mail: *djroser@civeng.unsw.edu.au*

O. Bergstedt
Göteborg Vatten Projekteringsavdelningen,
PO Box 123
SE-424 23, Angered,
Sweden

INTRODUCTION

The primary objective of a drinking water supply is to provide safe drinking water for public consumption. The microbiological quality of drinking water is maintained by selecting good quality source water, application of water treatment plant processes and protection of the distribution system. Despite many advances, occasional outbreaks of waterborne diseases caused by pathogens still occur in developed as well as in developing countries (Ashbolt 2004; USEPA 2005) so new strategies are needed to understand and manage the hazardous events thought responsible. To this end a new risk-based approach, the Water Safety Plan, has been proposed by the World Health Organization (2005) for the provision of safe water (Fewtrell & Bartram 2001). The Water Safety Plan is an iterative methodology where risk assessment and health targets promote risk management. Traditionally, assessment of microbial risks and identification of hazards and hazardous events has been undertaken through end point monitoring and epidemiologic health surveillance systems. These methods have generally resulted in drinking water of high quality but have, over the years, shown some shortcomings (Medema *et al.* 2006). To manage these shortcomings Quantitative Microbial Risk Assessment (QMRA) has been proposed as the preferred way to assess potential waterborne risks in the receiving community (WHO 2004). A further development has been the introduction of Hazard Analysis and Critical

doi: 10.2166/wh.2007.138

Control Point (HACCP) concepts aimed at promoting better management of key barriers. These developments imply a need to better quantify the frequencies and magnitudes of barrier 'hazardous events', i.e. 'an incident or situation that can contribute to the presence of a hazard, where a hazard is a biological, chemical or physical agent that has the potential to cause harm and/or give rise to water quality which is unacceptable for consumers' (Nadebaum *et al.* 2004).

Studying events is challenging as they often occur infrequently and unpredictably so that a low frequency grab sample style of routine monitoring has trouble detecting and characterising them. One approach which conceptually should allow the characterisation of such events is the analysis of Supervisory Control and Data Acquisition (SCADA) and similar data (e.g. river flow). SCADA data are collected in large quantities at high frequencies and are widely used for real-time online control and management of water treatment processes. At intervals as short as 1 s, SCADA systems collect flow, turbidity, pH, disinfectant residual and temperature measurements. But, despite the quantity of SCADA data available, only limited work appears to have been published on its post-collection analysis and use in risk assessment and management. LeChevallier & Au (2004) and WHO (2004) suggest better short-term monitoring and management schemes will improve health outcomes but there is little in either document on hazardous event characteristics. Haas *et al.* (1999) describe the statistical analysis of pathogen data in detail, but did not consider the analysis of surrogates or large timeseries. Lake *et al.* (2002) illustrate the use of particle size data for hazardous event management but do not report detailed statistical analysis of their own SCADA datasets. The exceptions are the work of Olofsson *et al.* (2001) and Westrell *et al.* (2003). The latter study reports statistics on coagulant dosage and chlorination failures based on analysis of SCADA data and 300 incident reports, and included QMRA simulations which estimate the impact of potential hazardous events. Westrell *et al.* (2003) conclude that most infection risk arises under nominal treatment conditions. However, other workers (e.g. Corso 2003; Smeets & Medema 2006) indicate that short-term failures can lead to disease outbreaks and be very costly. Clearly more peer-reviewed investigations of 'hazardous events' are needed.

The objective of this study was to analyse SCADA datasets obtained through the EU MicroRisk project (Medema *et al.* 2006) historical data survey and evaluate their potential use in QMRA and hence WSPs. Close attention was given to how 'hazardous events' might be recognised and quantified, and the use of timeseries analysis methods such as 'control charting' (Shewart 1931) and Cumulative Sum Control Charting (CUSUM) analysis (Taylor 2000). During the data analysis process a secondary issue considered was data reliability and quality. A preliminary assessment of its significance in SCADA data analysis was made from visual examination of SCADA traces, cross-comparisons of different SCADA records and comparison of SCADA with other plant records.

MATERIALS AND METHODS

As part of the MicroRisk project methodology a range of historical water quality data was requested from each participant. This yielded a complex mix of grab sample and on-line data reflecting all stages of water sourcing, treatment and distribution. When compiled in an MS Access™ database this still comprised ca. 1 megabyte (grab) and ca. 2 gigabytes (SCADA). This data was complemented by information on CTS configurations. Because of the extent and complexity of the data only a limited exploration proved possible so focus was put on documenting the data analysis process in principle along with its strengths and limitations and how it might be integrated with QMRA. Records were analysed from three CTSs (1, 6 and 8) with extensive SCADA records, grab sampling data and system configuration information. Water parameters monitored included flow, turbidity, free Cl_2, O_3, pH and temperature.

The initial data from even these systems did not fully meet our needs so additional data was requested for CTS 6 where electronic diary files (Word™ document and Excel™) were known to have been kept. CTS 6 treatment processes included coagulation, sedimentation, granulated active carbon filtration and chlorine disinfection, which were all monitored by SCADA systems. Diary records had been made on a daily basis. Issues noted included general treatment plant status, maintenance and small incidents.

Major events that could have an impact on the overall process performance were recorded in 'Deviation Reports'. From the available timeseries data 10 min SCADA means were extracted for the period 1 October 2004 to 19 September 2005 and matched to diary data. All information judged relevant to pathogen risk assessment was assembled in an MS Access™ database and manipulated in MS Excel™ workbooks, in particular the (1) turbidity in raw, filtrate and final drinking water; (2) Cl_2 residual in raw weir and final drinking water; and (3) pH in flocculation chamber 1.

Once the data had been collated and edited, potential SCADA events were identified empirically (i.e. by visual examination of timeseries), by the Shewart method (Figure 1) or by Cumulative Sum Control Charting (CUSUM) plots (example in Figure 2) (see StatsoftTM 2006 and http://www.statsoft.com/textbook/stquacon.html). In the Shewart method an event was provisionally identified by the occurrence of an outlier data group identified by comparison of the timeseries mode or similar statistic with 'control limit' boundaries (fixed value or statistically estimated from long term process records or special test runs). The CUSUM plotted S_1 to S_n estimated by the equation

$$S_n = S_{n-1} + (X_n - \bar{X})$$

where S is the sum of the sample measurements, n is the total number of data points, X_n the data point and $\bar{X}$ the arithmetic mean of the data points (Taylor 2000) (Figure 2). A provisional event corresponded to a major trend change where a process slowly 'slid' out of control. These were then matched to CTS 6 diary records and deviation reports to see if they were likely to reflect an operational meter problem or some other non-hazardous change, such as cleaning. For each type of event we estimated the frequency, duration and magnitude of events at the monitoring points. Five types of SCADA event were identified (Table 1).

SCADA data from CTS 1 and CTS 8 collected over 12–18 months between 2002–2004 were analysed to a more limited extent due to more limited secondary information being available for these systems. CTS 1 data was examined mainly for general patterns and internal consistency. SCADA data from CTS 8 was compared to concurrent water quality grab sample measurements of turbidity and total and residual Cl_2. In addition to turbidity, chlorine residual and pH, SCADA data from CTS 1 and 8 included ozone residual, and water supply and dosing flow rates.

RESULTS AND DISCUSSION

Timeseries analysis

Data interpretation posed three initial challenges:

1. Managing and analysing SCADA datasets in a PC environment (e.g. CTS 1 data consisted of 13 million

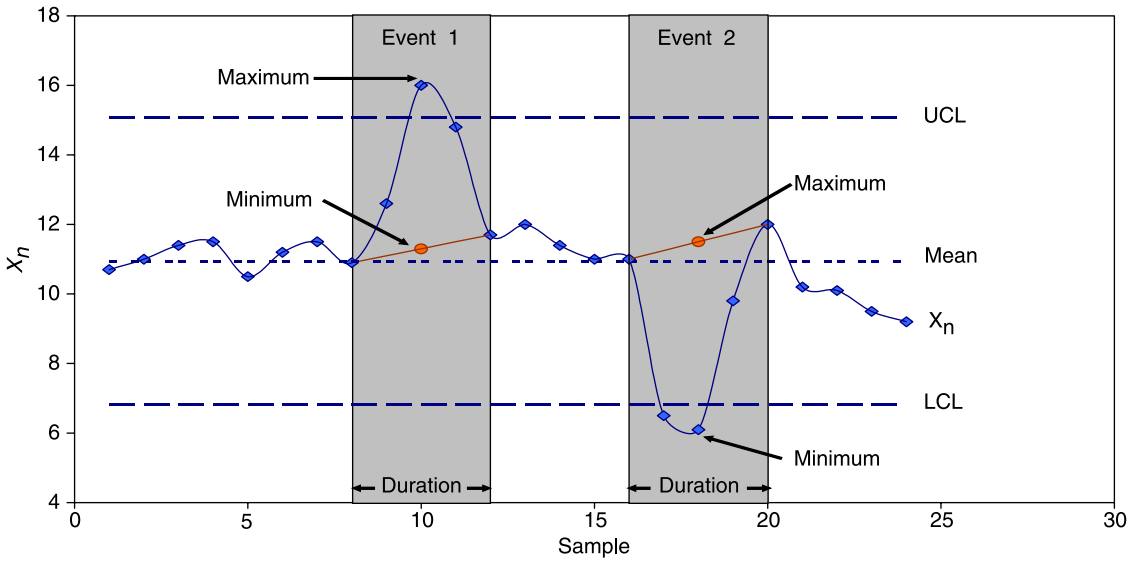

Figure 1 | Identification of event using Shewart chart (UCL = Upper Control Limit, LCL = Lower Control Limit, Mean = Arithmetic mean of samples, X_n = data point).

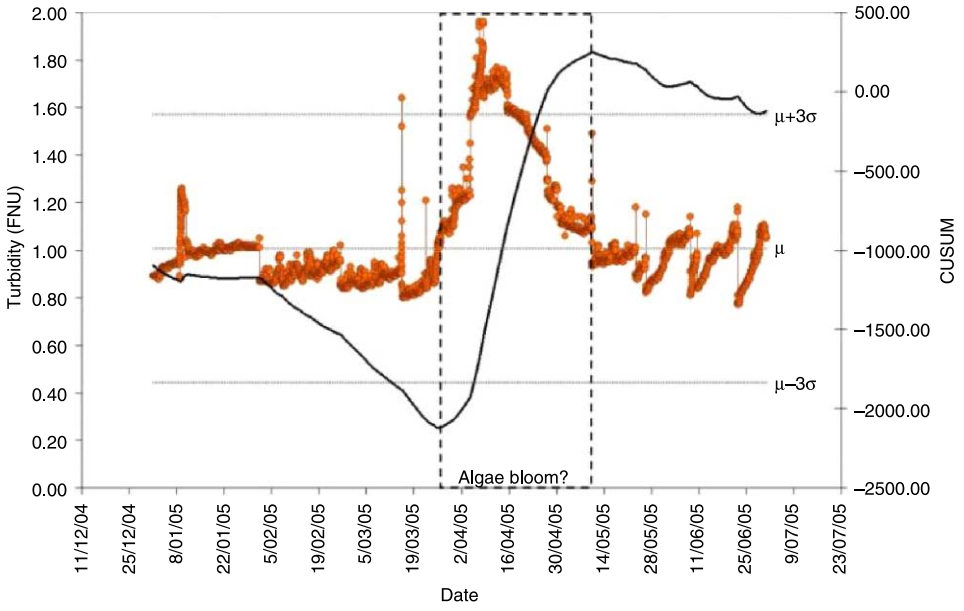

Figure 2 | Raw water turbidity (FNU) series (CTS 6) showing CUSUM plot (solid line) used to identify a possible algal bloom event period.

records and 600 MB. Data queries required tens of minutes, slowing analysis).

2. Defining/selecting criteria for recognising Events in a timeseries – the 'change point' problem (Taylor 2000).
3. Linking timeseries data to actual plant operation and malfunction.

All three were overcome to an extent to sufficient allow calculation of general event statistics to investigate the process of hazardous event identification and characterisation. A combination of MS Access™ and MS Excel™ proved (just) adequate for managing, initial filtering and visualising the timeseries records prior to statistical analyses. The 'change point' was identified as described in the methods section (though the 'out of range' criterion did not fully capture the complexity of the issue, e.g. see Moreno *et al.* (2005) for a Bayesian perspective of the 'change point'). Links to plant operation were made by linking events identified using CUSUM and visual examination to the complementary system performance information. The CUSUM technique appeared suited to detecting rapid small shifts and slower long term changes in the timeseries as might reflect the commencement and finish of a summer algae bloom (CUSUM Figure 2; 8/1, 30/1, 26/3, 10/5) and the increasing baseline variability.

In the case of CTS 6 a total 119 possible 'events' were identified and classified. Seventy one percent of these were

Table 1 | Event classification

Event classification	Elaboration
Maintenance in diary	Day to day maintenance e.g. cleaning and calibration of meters, shift of pumps
Maintenance probable	SCADA data pattern similar to maintenance but no corresponding diary entry
Incident in diary	Event found in Diary records. For example, pump shut-down due to power cut.
Incident in report	Event found in Deviation Report, e.g. Cl_2 dosing shut-down due to malfunction.
Unknown	If the event was of unknown cause

classified non-hazardous whereas the other 29% were classified as possibly hazardous. Of those classified non-hazardous, 85% were the result of maintenance and 15% the result of incidents. Of those possibly hazardous, e.g. marked reduction of chlorine for several time steps, 76% were of unknown cause and 24% were caused by maintenance or incidents. The primary value of these statistics was that they provided estimates of the maximum frequency and duration of treatment failure that may impact on pathogen risks. Estimation of impact magnitude was more problematic but for conservative modelling purposes total process failure can be assumed to assess whether the worst case posed a problem, and hence the need for further work. For those events classified as possibly hazardous their duration, with few exceptions, ranged between 0.5 and 2.3 h.

Event characteristics based on SCADA data were then used in QMRA models to estimate the impacts of disinfection failure for historical or *simulated* hazardous events (for further information see Medema *et al.* (2006, ch 8)). Chlorination loss (Table 2) of 1.5 h at CTS 6 only marginally increased the annual infection risk for *Campylobacter* as with the events assessed by Westrell *et al.* (2003). For CTS 8, however, simulated concurrent short-circuiting with the assessed frequency, duration and magnitude from SCADA data increased the annual risk 11-fold. Event duration estimates could also be compared to varying periods of simulated plant failure to determine the extent to which such risks were tolerable or otherwise (Figure 3). Such QMRA which used hazardous event characteristics derived from SCADA data highlighted how event impact need to be considered on a system-by-system and scenario-by-scenario basis.

Data reliability

Examination of data from CTS 1 and CTS 8 highlighted its quality as a major uncertainty for data interpretation (i.e. reliability, precision and accuracy). It had been assumed that the data as received would have been quality assured and appropriately edited and the meters from which the data was sourced well maintained. This did not appear to be always the case, e.g. comparison of chlorine data for CTS 8 showed poor correlation ($r^2 = 0.09$) between spot measurements and SCADA data. And for CTS 1 inconsistencies were detected between related on-line measurements, e.g. when ozone dosing had been interrupted, as shown by pump flow, this was reflected in two residual ozone meters but a third remained fixed at the nominal dose of 2 mg L^{-1}. These

Table 2 | Illustrative hazardous event impacts on pathogen risks at MicroRisk Catchment-to-Tap Systems (CTS)

CTS	Pathogen/barriers altered	Simulated event (= variations from baseline)	Total duration of event conditions	Average baseline risk (Prob. infection.person^{-1} y^{-1})	Average baseline + hazardous event risk (Prob. infection.person^{-1} y^{-1})
CTS 6	*Campylobacter*/ disinfection	Total suboptimal chlorination period based on SCADA data – worst case of total loss of disinfection assumed	1.5 h (based on SCADA)	2.5×10^{-6}	3.2×10^{-6}
CTS 8	*Campylobacter*/ catchment and disinfection	Short-circuiting reduced raw water reservoir protection to decimal reduction of 1.0 for each of nine 24 hour events over 1 year. Concurrent chlorination loss was simulated to occur due to power failure for 0.1 d.	2.4 h (based on SCADA + Scenario simulation)	1.7×10^{-5}	1.8×10^{-4}

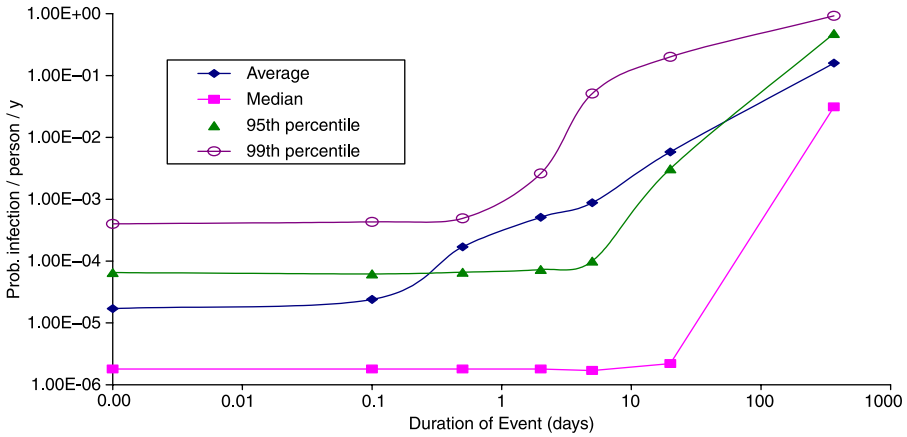

Figure 3 | Infection risk at CTS 8 in response to simulated chlorination failure alone.

problems reflected a need to systematically account for on-line/SCADA meter data quality during event identification and characterisation. The potential scale of quality assurance problems was identified by Scott *et al.* (1999) who showed that >25% of on-line monitoring systems used in the UK water industry probably generated unreliable data due to poor management, amounting to a loss of data worth hundreds of millions of Euros. Since then the UK Environment Agency has tried to improve the management situation through its MCERTs program (www.MCERTS.net). As of April 2007 only 3 types of flow meters and 2 on-line quality monitors (1 TOC, 1 turbidity) had been certified. This uncertainty further complicates recognition of a 'change point' due to the potential for both false positive and false negative events occurring with in-line data.

CONCLUSIONS

When water companies improve their record-keeping and data quality control, analysis of SCADA data should provide a powerful tool to manage drinking water safety, both for risk assessment and for risk control. The key attributes of SCADA data were confirmed as (1) its potential as a primary source of numerical data on hazardous event frequencies, durations and magnitudes which could be used in QMRA simulations; and (2) the potential for enhancing the benefits of on-line monitoring to include hazardous event identification and response triggering beyond that based on recognition of 'out-of-range' value occurrence. Limitations of the SCADA data included (1) the inability to detect all events due to sensor limitations (e.g. no *E. coli* meter); (2) SCADA 'incidents' were necessarily only provisional as the raw data were not strictly quality assured and could have interpretations other than hazardous events; (3) full interpretation needed complementary grab sampling, laboratory testing and plant diary records if 'SCADA' events were to be satisfactorily interpreted; and (4) the magnitude of a risk posed by a SCADA event was not necessarily clear from the on-line measurements.

It had been hoped that using mainly SCADA data we could estimate frequencies, durations and magnitudes of hazardous events occurring for the MicroRisk CTSs for use in QMRA. In principle this appeared feasible, but the process of getting to a sufficient level was too time-consuming to be practical with available resources. Conceptually improvements might be achieved through automation of data analysis. But the necessary programming/data management systems do not appear available except in a basic form (incident/maintenance reports as text-processor and spreadsheet files versus hand-written records) which does not yet routinely recognise the on-line data quality assurance issue.

This study should be a seen as a first-time exploration of what is required to utilise SCADA information and what issues 'data-miners' should consider in planning their analyses. Clearly more research into the use of SCADA data is needed before exploitation for QMRA will be reliable. Some questions which might be addressed include the following:

- How exactly does SCADA parameter variability reflect microbial risk probability?
- What scale of change in a parameter is a cause for concern?
- What are the best physical points to locate SCADA data measuring equipment?
- Can event identification and diary records be better automated and integrated?
- What are the best statistical analysis techniques for identifying 'change points' in SCADA data? CUSUM or other statistical process control methods?
- How should on-line monitoring be integrated with Water Safety Planning?

REFERENCES

Ashbolt, N. 2004 Microbial contamination of drinking water and disease outcomes in developing regions. *Toxicology* **198**(1–3), 229–238.

Corso, P. 2003 Cost of Illness in the 1993 waterborne Cryptosporidium outbreak Milwaukee, Wisconsin. *Emerg. Infect. Dis.* **9**(4), 426–431.

Fewtrell, L. & Bartram, J. (eds) 2001 *Water Quality: Guidelines Standards and Health. Assessment of Risk and Risk Management for Water-related Infectious Disease*. WHO, Geneva Switzerland.

Haas, C., Rose, J. & Gerba, C. 1999 *Quantitative Microbial Risk Assessment*. John Wiley & Sons, New York.

Lake, R., Agutter, P. & Burke, T. 2002 Using percentile analysis for determination of alarm values. *Wat. Sci. Technol. Wat. Supply* **2**(2), 145–150.

LeChevallier, M. & Au, K-K. 2004 *Water Treatment and Pathogen Control – Process Efficiency in Achieving Safe Drinking Water*. WHO, Geneva, Switzerland.

Olofoson, B., Tideström, H., Willert, J. 2001 Risk identification of urban water and wastewater systems (Risk identifiering or urbana VA-system). *Urban Water Report* **2** (in Swedish).

Medema, G., Loret, J.-F., Stenström, T. & Ashbolt, N. 2006 *Quantitative Microbial Risk Assessment in the Water Safety Plan. Final Report on the EU MicroRisk Project*. European Commission, Brussels.

Moreno, E., Casella, G. & Garcia-Ferrar, A. 2005 An objective Bayesian analysis of the changepoint problem. *Stochast. Environ. Res. Risk Assess.* **19**, 191–204.

Nadebaum, P., Chapman, M., Morden, R. & Rizak, S. 2004 *A Guide to Hazard Identification & Risk Assessment for Drinking Water Supplies*. Cooperative Research Centre for Water Quality and Treatment, Salisbury, Australia.

Scott, M., Bogue, R., Marshall, D., Thomas, C. & Whitworth, C. 1999 *On-line Instrumentation Standards and Practices*. UK Water Industry Research Limited, London.

Shewhart, W. 1931 *Economic Control of Quality of Manufactured Product*. American Society for Quality, Milwaukee, WI.

Smeets, P. & Medema, G. 2006 Combined use of microbiological and non-microbiological data to assess treatment efficacy. *Wat. Sci. Technol.* **54**(3), 35–40.

StatSoft, Inc 2006 *Electronic Statistics Textbook*. StatSoft, Tulsa, OK, http://www.statsoft.com/textbook/stathome.html.

Taylor, W. 2000 Change-Point Analysis: a powerful new tool for detecting changes. *Qual. Engng*. Available at: http://www.variation.com/cpa/tech/changepoint.html (accessed 15/11/05).

USEPA 2005 *Water Distribution System Analysis: Field Studies, Modelling and Management. A Reference Guide for Utilities*. EPA/600/R-06/028. USEPA, ORD, Cincinnati, OH.

Westrell, T., Bergstedt, O., Stenström, T. & Ashbolt, N. 2003 A theoretical approach to assess microbial risks due to failures in drinking water systems. *Int. J. Environ. Health Res.* **13**(2), 181–197.

WHO 2004 *Guidelines for Drinking-water Quality*, 3rd edn. **vol. 1**. Recommendations WHO, Geneva, Switzerland.

WHO 2005 *Water Safety Plans. Managing Drinking-water Quality from Catchment to Consumer*. WHO, Geneva, Switzerland.