International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Integrated Modeling for Source Characterization of Pathogenic Contamination in Watersheds

Michael Tryby [a], S. Thomas Purucker[a], and Gene Whelan[a]

[a]*US EPA, Ecosystems Research Division, 960 College Station Road, Athens, GA 30605, USA*
(Tryby.Michael, Purucker.Tom, Whelan.Gene @epa.gov)

Abstract: The US EPA's regulatory framework for recreational waters has protected public health for decades. Pathogenic contamination of these waters, however, remains a frequent cause of impairment. Integrated modeling is being leveraged to advance the agency's understanding of pathogen fate and transport processes in watersheds and improve its ability to predict the consequences of exposure. This paper describes integrated modeling research focusing on source characterization techniques for pathogen transport scenarios in watersheds. Source characterization is a hidden requirement of Quantitative Microbial Risk Assessment, a method for estimating infection risks being evaluated across several programs within the EPA. A hybrid source characterization approach is described and demonstrated that utilizes integrated and inverse modeling methodologies to determine pathogen source allocations.

Keywords: Integrated Modeling; Multimedia Modeling; Source Characterization; Inverse Modeling; Receptor Modeling; Pathogens; Fecal Contamination; Recreational Waters; Risk Assessment

1 INTRODUCTION

The most frequent cause of impaired recreational waters in the United States is pathogen contamination. Pathogen releases are often associated with rainfall events and occur in receiving waters influenced by various sources such as septic systems, sewage discharges, combined sewer overflows, and agricultural land use. Current recreational water quality criteria established by the EPA are meant to protect public health from acute illness associated with exposure to pathogens. These criteria, however, are more than 20 years old. During the ensuing time, research has improved our fundamental understanding of the problem. In the near term, the agency seeks to apply this knowledge as it promulgates new recreational water criteria in 2012. Looking beyond, the agency seeks to advance pathogen fate and transport simulation in watersheds and exposure science with the development of tools to support a quantitative framework for assessing pathogen exposure risks in recreational waters.

Multimedia modeling is central to the approach being taken because pathogen fate and transport in watersheds encompasses several cross media processes. For example, consider a manure land application scenario as illustrated in Figure 1. The manure applied there to fertilize crops is a potential pathogen source. Transport processes acting on the manure-laden field would depend on meteorological conditions such as rainfall intensity and watershed topological conditions such as tillage practices, and crop and soil properties. These processes typically include infiltration, overland flow, in-stream dilution, advection, dispersion, and the complex mixing patterns occurring in the receiving water. Thus, the multimedia modeling domain encompasses atmospheric, surface and subsurface hydrologic and environmental hydrodynamic modeling. Furthermore, the exposure potential at the receiving water body is a complex stochastically driven function of these processes acting on multiple point and non-point pathogen sources across media boundaries.

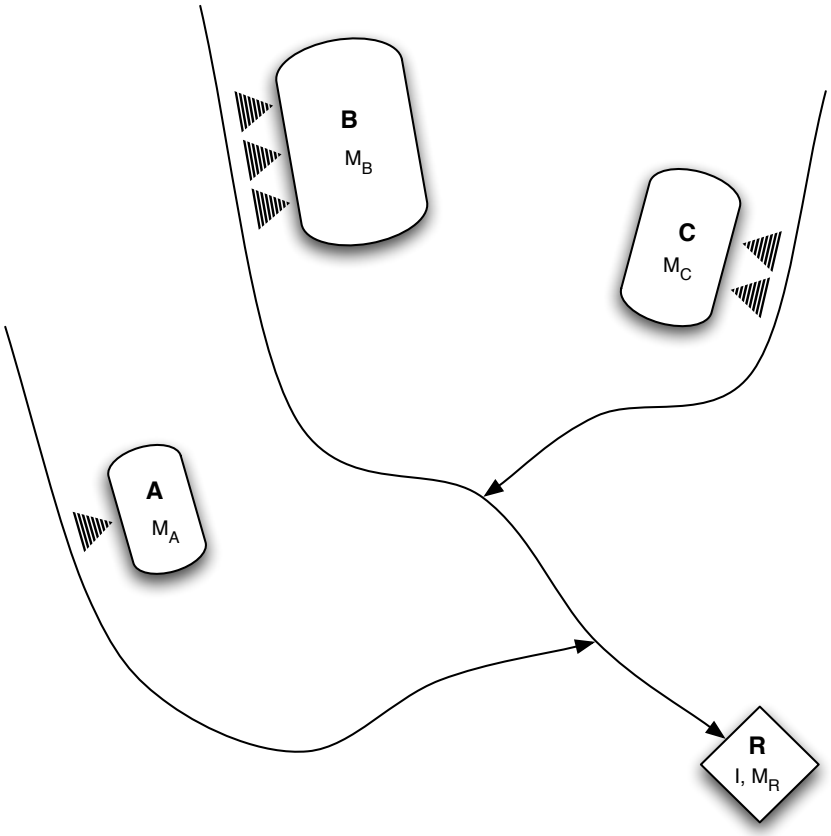

Figure 1: A hypothetical watershed impacted by agricultural activities. The scenario consists of three agricultural operations located in a watershed adjacent to each of the three reaches which make up the drainage. Farms A, B, and C represent animal feed and land application operations. Fecal contamination from Farms A, B, and C enters each adjacent reach as runoff, then flows downstream to Receiving Water R where it is observed as a mixture.

This work focuses on agricultural sources, although fecal contamination in watersheds can originate from virtually anywhere and from several source categories, including humans and native wildlife, in addition to domesticated animals. Hundreds of different pathogens can be transmitted via exposure to fecal-contaminated water. Most pathogens infect the intestines and result in gastroenteritis. With the potential for so many viral, bacterial, and parasitic pathogens to be present in contaminated waters it is impractical and cost prohibitive to directly monitor all of them. Thus, broad spectrum surrogate organisms such as Escherichia coli in fresh waters and Enterococcus in marine waters [Wade et al., 2003] are used to indicate the presence of fecal contamination and the risk of illness associated with pathogens.

A future regulatory framework for recreational waters may adopt risk based criteria that build on the historical pathogen indicator paradigm. Hunter et al. [2004] describes Quantitative Microbial Risk Analysis (QMRA) as taking the well-developed chemical risk paradigm and modifying it to access microbial risks. Using QMRA an estimate of exposure risk is calculated based on the types of pathogens present, their infectivity, concentration distribution, the exposure pathway, dosage, and characteristics of the population being exposed. Thus, quantitative characterization of pathogen sources (*i.e.* estimates of pathogen concentrations broken down by type) is a necessary step in conducting a QMRA. A host of confounding factors, however, complicates practical application of QMRA.

The main difficulty is that a loose relationship exists between indicators and pathogens,but indi-

cators are used as the basis for rules and regulations; furthermore, observations tend to focus at the receptor. For example, refocusing attention on Figure 1, samples I are taken at location R to detect the presence of fecal contamination. The results indicate their presence, and their relative abundance correlates with the total fecal load. Conventional methods for detecting indicator organisms only provide data on the abundance of specific indicator organisms associated with fecal contamination. The results do not provide any characterization of the fecal source, such as pathogen host type, nor do they provide any direct information on pathogen presence or abundance. Therefore, it is not possible to estimate with accuracy exposure risks at location R using conventional analysis of Samples I. This paper describes how an integrated modeling approach can be used to quantitatively characterize pathogenic contamination sampled at location R.

A hybrid source characterization approach is proposed that integrates receptor, forward, and inverse modeling techniques to improve risk estimates for recreational exposures to pathogens occurring in natural waters. The first phase of this approach adapts receptor modeling techniques developed for air pollution management to provide source apportionment estimates — the relative contributions from each fecal source. Fecal sterol / stanol profiles derived from multiple fecal sources (e.g. humans, cows, ducks, dogs) [Shah et al., 2007] are used as the basis for the receptor modeling calculations. Two receptor modeling approaches have been investigated for the application — deterministic and stochastic chemical mass balance receptor models. These receptor modeling approaches, however, are limited by a linear input / output assumption between source loadings and concentrations observed at the receptor; they do not capture the complex fate and transport dynamics present in watersheds, constituent reactivity, microbial reproduction and die off, although source term uncertainty and measurement errors are represented.

A forward model developed using multimedia modeling techniques, can simulate the release, fate, and transport of microbial constituents and thereby account for these unrepresented processes. A forward model is currently under development and being coupled with a QMRA model within FRAMES [Whelan et al., 2010] that can translate source term loadings into exposure risks. Future work will investigate, general inverse modeling techniques that utilize a forward model for the input / output relation, and therefore, incorporate these missing processes into source allocation estimates. We seek to integrate these source characterization approaches and the QMRA model within FRAMES to improve pathogen exposure risk estimates. The intent of this paper is to present the integrated approach currently under development.

2 METHODOLOGY

Source characterization is a generic term describing methodologies that infer the characteristics of a pollutant source from observational data. In practice, source characterization can play an important role in environmental monitoring and remediation activities. To date, the work on source characterization conducted in hydrologic contexts has focused on nutrients (nitrogen and phosphorous) and heavy metal contamination. The solution of source characterization problems includes the identification of source locations, compositions, allocations, concentrations or some combination thereof. A key objective of this work is the development of a pathogen source allocation capability; specifically, quantifying the contribution of each different source species type to the total fecal load. The composition of fecal contamination is important for QMRA, however, making this determination is difficult. Host specific markers are difficult to identify and use because similar markers are present in many different species varying only by degree. The mixing of these markers due to transport phenomena in the watershed further confounds analysis. The paragraphs that follow describe each of the modeling techniques being integrated.

2.1 Receptor Modeling

Given observational data at a receptor, the goal of Receptor Modeling (RM) is to identify pollutant sources and apportion loadings to them. RM is a type of inverse model that assumes a linear input / output relationship between source loadings and concentrations observed at a receptor. RM is difficult due to complexity of transport processes, reactivity of constituents, and poor delineation

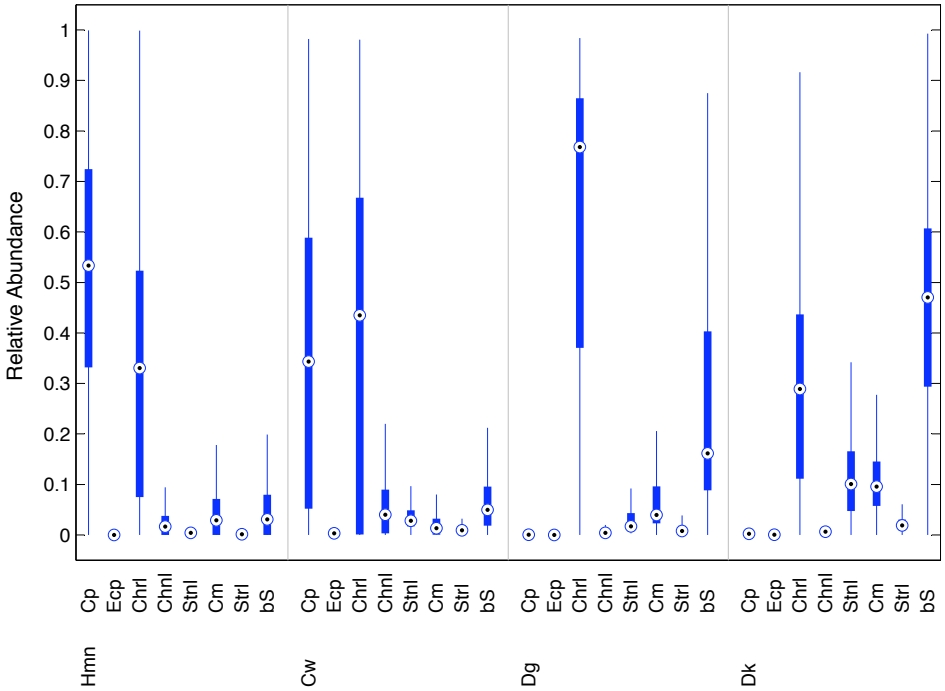

Figure 2: A maximum entropy plot derived from sampled sterol profiles for humans, cows, dogs, and ducks [Shah et al., 2007]. The relative abundance of each sterol compound is displayed. Note the amount of natural variability present in the fecal sterol sources sampled (Box edges indicate the 25th and 75th percentiles, whiskers indicate the 99th percentile range. Outlying data points have been omitted.).

of source characteristics. Different approaches for formulating a receptor model depend on the characteristics of the system being analyzed and the observational data available; they include mass balance and multivariate models. Mass balance models are predicated on the fundamental principle of mass conservation. A chemical mass balance, therefore, is assumed, meaning that the observed sample can be described mathematically as a linear combination of source components. The mass balance assumption is limiting since it means reactive contaminants cannot be used as the basis for RM. It does, however, make the problem mathematically tractable. The assumption of mass conservation complicates the selection of markers since they must be quantitative and nonreactive.

Receptor Modeling Example. To illustrate the type of analysis possible using this hybrid approach, a receptor modeling example has been prepared using data from the published literature. The data set consists of source profiles for humans, cows, dogs, and ducks (eight fecal sterol/stanol compounds per profile) and 39 different contrived fecal mixtures [Shah et al., 2007]. A stochastic receptor modeling formulation using bio-chemical marker profiles, specifically fecal sterols/stanols, has been developed to allocate an observed mixture of fecal contamination across a set of profiled fecal sources (see Figure 2). One important aspect of receptor modeling with bio-chemical marker signals is the high degree of natural variability present in the source profiles. A linear mass balance equation is the mathematical basis of the formulation combined with the sophisticated treatment of error, natural variability, and uncertainty possible using Bayesian inference.

The stochastic approach presented here is a modification of the Bayesian source apportionment

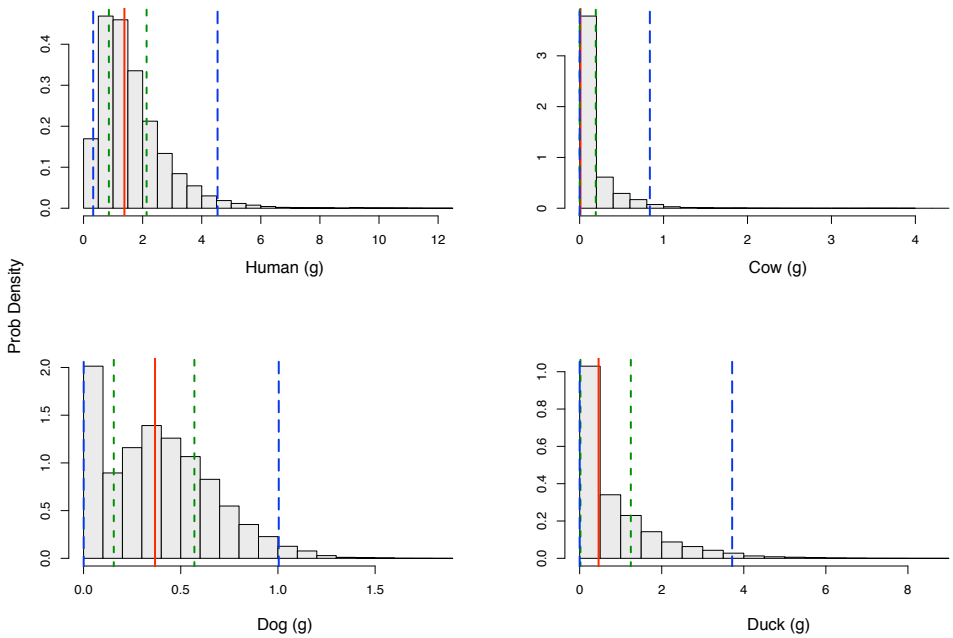

Figure 3: Stochastic source allocation estimate for a 1:1:1 Human, Cow, Dog, fecal sterol mixture. The probability density versus log concentration is shown for each source with the mean (red), the 68th (green dashed) and 95th (blue dashed) percentile intervals indicated. The true allocation (0.33, 0.33, 0.33, 0.0 g) are within the 95th credible intervals of the estimate for all sources. Note the high degree of variability present in the estimates and the false inclusion of duck in the allocation.

receptor model developed by Keats et al. [2009]. Using it, source fecal profiles, source allocations, and the observed fecal contamination are considered stationary stochastic variables. The estimate of the posterior distribution for the source allocation is proportional to the likelihood of the observations, given the source allocations and profiles, and their a priori estimates (see 3 for one of the mixtures). The formulation is solved by constructing Markov chains for each of the observed 39 fecal mixtures (with a common set of source profiles), then sampling from the domain of the prior distributions, allowing for inference on the source contribution and profiles. A Metropolis-Hasting algorithm is used to determine acceptance and construct the posterior distribution that reflects the observed mixtures. At convergence, the Markov chain approximates the solution of the source allocation problem of interest an apportionment probability distribution for each source type (i.e., the gray histograms in 3).

Even in a laboratory setting with the Shah data set and an artificially restricted set of predictors, this is a high dimensional problem. There are 35 parameters in each of the 39 simulation sets corresponding to the observed mixtures, resulting in $> 1K$ parameters. Field applications will be expected to deal with an even greater set of potential predictors. Dimensionality causes problems across the board: for chemical mass balance approaches to the receptor problem (summarized in Christensen and Gunst [2004]), fate and transport model calibration, as well as for conventional Bayesian Markov chain Monte Carlo applications. In addition, due to correlation in the Markov chains, simulations may run significantly longer than for direct Monte Carlo sampling, with runs $> 100K$ simulations necessary to have a significant chance of convergence. Therefore, solution approaches must be able scale well for both parameters and sampling. A great advantage of the Keats et al. [2009] approach that we adopt is its use of Hamiltonian Monte Carlo

techniques [Hajian, 2007] that increase the rates of acceptance and allow for larger jumps in the multi-dimensional parameter space, leading to much more efficient parameter estimation, compared to traditional techniques, as dimensionality increases.

The linear chemical mass balance receptor model is appropriate for this laboratory scenario where fate and transport processes do not reduce or transform the sterol/stanol compounds between the source and receptor. However, the dynamics of fate and transport processes requires additional algorithms and parameters for effective inference in QMRA applications. Therefore, a research question that the authors are addressing is replacing the linear mass balance assumptions with more realistic fate and transport process algorithms. Current multi-media forward models cannot be calibrated on the observations of simultaneous multiple predictors necessary for QMRA and typical mass balance receptor modeling approaches do not faithfully represent known environmental processes. Combination of these two approaches is therefore necessary and would be a significant advance for QMRA, source identification, allocation, and remediation.

2.2 Multimedia Forward Modeling

Multimedia modeling can leverage existing modeling capabilities to address complex cross-media problems. Linking the contamination source to a receptor using multimedia modeling is a new tool for the risk assessment process that may improve the correlation between indicator and pathogen predictions. A multimedia model is currently being developed to express pathogen loading, transport, and fate (source-to-receptor modeling) and will provide source modeling capabilities that enable QMRA. The Framework for Risk Analysis in Multimedia Environmental Systems (FRAMES) will support pathogen fate and transport model development and QMRA [Whelan et al., 2010]. The flexibility and extensibility of integrated multimedia modeling enables the forward modeling approach outlined by Whelan *et al.* to be extended, thereby providing a more comprehensive understanding of pathogen fate and transport through an integrated analysis with receptor and general inverse modeling techniques. Furthermore, the powerful analysis capabilities of mature integrated modeling frameworks like FRAMES can aid in the solution of the source characterization approaches discussed and improve handling of uncertainties through application of sophisticated tools such as optimization algorithms, Monte Carlo methods, and uncertainty analysis. The proposed inverse modeling approach is briefly described in the paragraphs that follow.

2.3 General Inverse Modeling

Solution of source characterization problems often involves formulating and solving an inverse problem, since the monitoring data is sparse and contains only faint signatures of the desired system information. Inferring source characteristics, such as source locations, pathogen release, and total fecal loading from the observational data collected at the receptor (scenario illustrated in Figure 1), is a canonical inverse problem. Problems such as this are difficult to solve since they are frequently ill-posed and computationally intensive.

There are several approaches for solving inverse problems encountered as part of a source characterization analysis. Deterministic and stochastic inverse problem formulations are powerful tools that infer input characteristics, given sufficient output data. These formulations, however, rely on a forward model, usually a system of partial differential equations, that describes the dynamic processes of the environmental system and defines the relationship between model inputs and outputs. Development of a forward model would allow us to formulate and solve more general inverse problems than the receptor modeling formulations discussed previously. For example, an integrated pathogen fate and transport model-based inverse problem could incorporate more diverse sources of information — such as chemical kinetics, transport delays, and flow paths — and thus produce better solutions. Solving inverse problems with integrated forward models may prove particularly challenging due to the characteristics of the coupled system of model components that describe transport processes in watersheds. The solution of inverse problems is also several orders of magnitude more computationally intensive than solution of the corresponding

integrated model, since thousands of forward model evaluations are typically required.

Deterministic inverse problem formulations couple the forward model with formal mathematical or heuristic search procedures to determine the model parameters or boundary conditions that best approximate the observed data. Gradient-based search techniques represent the state-of-the-art search procedure for solving inverse problems using the optimization approach. Using a integrated forward model, however, may be difficult as gradient computations would occur across model boundaries, which may lead to solution instabilities or excessive computational effort. Furthermore, gradient-based search techniques, while efficient for well formed problems, are often poorly suited for ill-posed environmental problems with multiple optima and non-linear, discontinuous, and discrete features; and thus, can converge erroneously to local optima, missing the global optimal solution of the problem. One alternative is the use of global optimization techniques, such as evolutionary algorithms, which can provide a more robust search of the decision space [Tryby et al., 2009]. This alternative, however, has a high computational cost that require the application of distributed computing techniques for tractability.

Another potential approach is statistical inversion. Statistical and deterministic inverse models are related and differ in their treatment of information and error. Deterministic inversion is a special case of statistical inversion that takes a frequentist view of probability and assumes errors are normally distributed and additive. Statistical inversion represents information generically using probability distributions and is rooted in a Bayesian view of probability. Bayes' theorem makes it possible to move back and forth between the conditional and marginal probabilities that are used to represent the observational data and the source model. The forward relationship between the model and the data is captured in the likelihood function that is also a conditional probability. Problem structure dictates the solution technique employed and ranges from analytical solutions to demanding Monte Carlo simulations. Statistical techniques are better able to handle some types of ill-posed problems than deterministic inversion. Given the non-linearities and uncertainties associated with the pathogen source characterization problem, investigation of statistical inversion methodologies is planned.

3 CONCLUSIONS

Utilizing source characterization methodology that combines bio-chemical MST markers with receptor modeling techniques, the authors have demonstrated the potential for pathogen source apportionment. Indeed, these results are promising and motivate the authors to continue development of the integrated approach we have described. A multimedia pathogen fate and transport model is currently under development; once it is completed more general inverse modeling methodologies can be explored that will allow the assumption of linear input/output relations to be relaxed and alternate sources of information to be utilized to eliminate false allocations and reduce the uncertainty associated with allocation estimates.

Although receptor, forward, and inverse modeling are different approaches they share the same objectives; integrating them will provide insights not otherwise possible. For example, integrating modeling approaches can inform the QMRA process by improving source term estimation and illuminating the differential fate and transport of indicators and pathogens. Combining source allocation estimates generated using receptor and general inverse modeling with pathogen fate and transport forward modeling and QMRA, the authors have described an approach by which the quality of pathogen exposure risk estimates may be improved.

ACKNOWLEDGMENTS

The views expressed in these Proceedings are those of the individual authors and do not necessarily reflect the views and policies of the United States Environmental Protection Agency. These Proceedings have been reviewed in accordance with EPA's peer and administrative review policies and approved for presentation and publication.

REFERENCES

Christensen, W. F. and R. F. Gunst. Measurement error models in chemical mass balance analysis of air quality data. *Atmospheric Environment*, 38(5):733–744, Feb 2004.

Hajian, A. Efficient cosmological parameter estimation with hamiltonian monte carlo technique. *Physical Review D*, 75(8):083525, Apr 2007.

Hunter, P. R., P. Payment, N. Ashbolt, and J. Bartram. *Assessing microbiological safety of drinking water: improving approaches and methods*, chapter Assessment of Risk, pages 79–109. Organisation for Economic Co-operation and Development/World Health Organization, Paris, France, 2004.

Keats, A., M. Cheng, E. Yee, and F. Lien. Bayesian treatment of a chemical mass balance receptor model with multiplicative error structure. *Atmospheric Environment*, 43(3):510–519, Jan 2009.

Shah, V. G., R. H. Dunstan, P. M. Geary, P. Coombes, T. K. Roberts, and E. V. Nagy-Felsobuki. Evaluating potential applications of faecal sterols in distinguishing sources of faecal contamination from mixed faecal samples. *Water Research*, 41(16):3691–700, Aug 2007.

Tryby, M., B. Y. Mirghani, G. K. Mahinthakumar, and S. R. Ranjithan. A solution framework for environmental characterization problems. *International Journal of High Performance Computing Applications*, 00(0):1–19, Oct 2009.

Wade, T. J., N. Pai, J. N. S. Eisenberg, and J. M. Colford. Do u.s. environmental protection agency water quality guidelines for recreational waters prevent gastrointestinal illness? a systematic review and meta-analysis. *Environ Health Perspect*, 111(8):1102–9, Jun 2003.

Whelan, G., M. Tryby, M. Pelton, J. Soller, and K. Castleton. Using an integrated, multidisciplinary framework to support quantitative microbial risk assessments. In *Proceedings of the 2010 International Congress on Environmental Modelling and Software*, Ottawa, Canada, July 5-8 2010.

LIST OF ABBREVIATIONS

bS	β-Sitosterol
Chnl	Cholestanol
Chrl	Cholesterol
Cm	Campesterol
Cp	Coprostanol
Ecp	Epicoprostanol
Stnl	24-ethylcoprostanol
Strl	Stigmasterol