

**D.C. Water Resources Research Center
Annual Technical Report
FY 2009**

Introduction

This report summarizes the activities of the District of Columbia (DC) Water Resources Research Institute (the Institute) for the period March 1, 2009 through February 28, 2010. The Institute is one of a network of 54 such entities at land-grant universities in the nation which constitutes a federal/state partnership in research, information transfer and education regarding water related issues. The Institute provides DC with interdisciplinary research support to identify city water and environmental resources and problems and contribute to their solution.

In recent years, there have been increasing opportunities for the Institute to step into a major role as a vibrant center for water-related research in the DC area; to serve as a conduit for communication between and among agencies, universities, and other area researchers; and to support the development of an integrated water research community within the District of Columbia. DCWRRRI is strengthening an Advisory Board consisting of a committed group of high-level representatives from several important water use and water quality protection agencies and other water-related stakeholder organizations. There has been increased activity within water-related disciplines at universities within the DC University Consortium, and increased opportunities for coordination with Universities throughout the DC University Consortium, particularly to address issues involving multiple disciplines. There have also been increased opportunities for the Institute to coordinate with locally-based federal water research facilities and agencies.

Interest in, and opportunities for, research and outreach activities with local and federal agencies in DC reflects shifts in federal and local research priorities and opportunities such as the American Recovery and Reinvestment Act, the US Department of Interior WaterSMART program, revisions to the US Army Corps of Engineers' Principles and Guidelines under the Water Resources Development Act, and other federal spending and activities related to water infrastructure and resource management, climate change, water and energy efficiency, public health, food safety and security, and general interest in sustainable development, particularly related to water quantity and quality and related ecosystems. DC Universities have a unique role to play in offering opportunities for continued education, demonstration sites for new technologies, and serving as host to events on water issues of national and international interest for members of the federal family who have made DC their home during periods of public service. These federal agency personnel often come from parts of the country with very different climate, hydrology, water law, and water use practices from those in the DC area.

The Water Resources Research Institute coordinates and facilitates water resources related research projects through seed grants provided to faculty members from the consortium of universities in the District. Presently, these universities include the University System of the District of Columbia, Howard University, George Washington University, The Catholic University, Georgetown University, and American University. The opportunity to train students through development and implementation of practical applications of water science in multi-disciplinary programs is a major accomplishment of the Institute. Through these research projects, students also interact with employers at federal and local agencies essential for future job opportunities. The seed grant program allows faculty members access to new technologies and equipment that develop their expertise in water resource management. Results of each project are reported and disseminated through published studies, technical reports, seminars, newsletters, brochures, through an information transfer partnership with the Cooperative Extension Service Water Quality Education Program and via our website.

Many positive changes have taken place within the host university that will greatly enhance the Institute's ability to fulfill its mission. The University name has recently changed to the University System of the District of Columbia, which now includes both a Flagship University and a Community College. As UDC has evolved through three different Presidents and three Provosts, we have seen a growing recognition of the central role of water resources education as a cornerstone of the new College of Agriculture, Urban Sustainability and

Environmental Sciences (CAUSES), where the Institute will be located when the new College initiates its academic programs in the fall of 2010. In the fall, CAUSES will introduce a new B.Sc. Program in Environmental Science with emphasis in Water Quality and a Professional Science Masters Program in Water Resource Management. UDC's commitment to this program is clearly evidenced on the home page of the University's website <http://www.udc.edu/>, which features the Fall 2010 Course Outline http://www.udc.edu/docs/course_schedule_fall2010.pdf with the phase, College of Agriculture, Urban Sustainability and Environmental Sciences, from Architecture to Water Resource Management .

The Institute website, <http://www.udc.edu/wrri/>, provides updated information about current activities. The Institute also completes bi-seasonal issues of the Water Highlights Newsletter. These documents are very informative and highlight current research and educational projects sponsored by the Institute along with interactions among faculty members and their student interns on projects and conferences. An electronic mailing list of over 150 Water Resources faculty and experts in the consortium of universities in Washington DC is maintained and regularly updated and disseminated via email to report updates on local, regional, and national water issues received by the Institute. This line of information transfer has enhanced the visibility and credibility of the Institute among its stakeholders.

Research Program Introduction

The District of Columbia (DC) has a total area of 68.3 square miles (177 km²), of which 61.4 square miles (159 km²) is land and 6.9 square miles (18 km²) (10.16%) is water. The District is unique among the states included in the Institutes of Water Resources programs. The water resources problems facing DC include those facing most large municipalities, particularly in the Eastern states, including maintenance of high drinking water quality, storm water management, an increasing need for watershed-based water resources management, aging infrastructure, and the continuous need for expanding water supplies in the face of population growth and the added pressures of climate change. With the increase in population and aging storm water infrastructure, the impact of storm water management on the quality and quantity of the District's water resources remains a major issue.

The water resource issues and problems of DC are uniquely impacted by its location, climate, and hydrology. The District also experiences issues that stem from its transient residential population and job base (including federal agencies and universities) and its unique government structure. These characteristics impact the manner in which government institutions and agencies can manage and protect its water resources, as well as the manner in which water issues can be communicated to area residents. Water use in the city is significantly impacted by commuters from surrounding suburbs. DC has a resident population of 599,657 but during the workweek, that population rises to over one million. The Washington Metropolitan Area, of which the District is part, has a population of 5.3 million, the ninth-largest metropolitan area in the country. The federal government is a major employer, accounting for about 27% of the jobs in Washington, D.C. in 2008. In an area of less than 70 square miles, DC has six major universities, including American University (American); The Catholic University (CU); George Washington University (GWU), Georgetown University (Georgetown); Howard University (Howard), and The University System of the District of Columbia (UDC). These institutions are not only significant in DC as centers of research and education, but also as major employers; the top four non-federal employers in DC are GWU, Georgetown, Washington Hospital Center (the teaching hospital for Georgetown University School of Medicine), and Howard.

Critical current issues facing the District of Columbia are reflected in the research priorities identified by the DCWRRRI Stakeholder Advisory Board, including: Improved freshwater monitoring systems, particularly with respect to the Chesapeake Bay; Toxics monitoring and prevention and river restoration, particularly with respect to the Anacostia River; Potomac drinking water source protection partnership, including collaborative planning between upstream agricultural communities and downstream urban communities; Regulatory requirements for stormwater, particularly with respect to nutrient management and pesticide safety; Clean cities and sustainable urban infrastructure that minimizes water use or improves or minimizes impacts to water quality; Prevention of contamination of water supplies, particularly by cryptosporidium or atrazine; Improved restoration science, focusing on monitoring and measurement of the benefits of ecosystem services over baseline conditions, particularly with respect to tidal areas; Improved understanding of natural resources economics; and Improved understanding of water and energy interrelationships and its impact on water use and water quality in the District.

The DC Water Resources Research Institute will continue to provide the District with inter-disciplinary research support to both identify and contribute to the solution of DC water resources problems. These research and educational projects provide students with essential practical skills required for future job opportunities and also allow faculty members access to new technologies and equipment that develop their expertise in water resource management. Reports for six projects funded are included in this technical report.

In summary, Dr. Seon Ho Kim's project, Development of a Fast Optimization Technique using Interactive Spatial Join for GIS Application in Water Resources, studied two statistical approaches for estimating spatial joins on quad-tree indexed raster data, namely, Probabilistic Joins (PJ) and Incremental Stratified

Research Program Introduction

Sampling Joins (ISSJ). We proposed a framework that combines two statistical approaches to allow fast interactive data explorations and the opportunity for the user to then drill down with full spatial joins if desired. Experimental evaluations on real and synthetic datasets showed that our proposed PJ method resulted in reasonably accurate results with near zero response time. Our ISSJ method, while not as fast as PJ, provides results with bounded confidence intervals up to an order of magnitude faster than full quad-tree join. Our framework can be used to build an end-user query visualization tool that allows true interactive exploration of large raster based GIS databases. This study used synthetic data for the evaluation of the proposed techniques. In the future we plan to evaluate the PJ method using real water resource dataset.

Dr. Pradeep Behera in phase II of his project *Development of Web-based Rainfall Statistical Analysis Tool for Urban Stormwater Management Analysis*, developed a rainfall statistical analysis tool extremely useful for urban stormwater modeling, especially for analytical probabilistic models. For analytical models, often analysis of rainfall data is cumbersome which is taken care in this project. Engineers and modeling professionals in any part of the nation can utilize the data to obtain the rainfall statistical parameters. Even though the application described throughout this paper is still in active development, there is already a short, but growing, list of feature requests. Chief among the requested features is the ability to upload files containing National Climatic Data Center (NCDC) data from multiple locations. The NCDC does provide consolidated files of this type already, so providing a mechanism by which these files can be utilized is of top priority. Incorporating this feature, and integrating a simple database of the locations of each station (latitude and longitude) stored by the NCDC, would allow other features, such as the ability to generate maps, to become realistic possibilities. As noted in previous sections, there are still some differences in functionality between the command-line version of the application and the web-based interface. In the near future, these disparities will be rectified, bridging the gap between the two interfaces. It would also be extremely advantageous to bring the full suite of plotting functionality provided by the matplotlib library into the user interface, allowing individuals to have much greater control of their graphical output.

Dr. Byunggu Yu's project *Application of Spatiotemporal Informatics to Water Quality (Phase II)*, investigated the problem of continuous monitoring of urban runoff at outfall points. This report presents the conceptual basis, technical details, and experimental results of a newly developed remote monitoring solution based on an advanced sensor platform. A prototype and accompanying algorithms were developed using the Sun SPOT as a sensor platform. Consequently, the collected accelerometer data were processed and analyzed in various ways to quantify the amount of water flow in the pipe. Our experimental results demonstrated that our approach has a great potential to measure the water quantity with any desirable precisions required in real applications of urban runoff monitoring. The platform is flexible and expandable and provides a possibility for monitoring the water quality details of the flow. The lab test results are promising and, based on this, we are planning to apply the proposed technology in various real sewer waterfalls in the field in near future. In addition to the SPOT, we will also study the application of other available platforms, such as iMote2, in a comparative manner. In the field, more challenges are expected in various areas including on-site power (solar panel charging the sensor platform), cellular communication or wide-area networking for automated data collection and run-time reprogramming of the platform (data and control transmission between central host computer and the field devices), memory and power optimization.

Dr. Valbona Bejleri project, *Modeling Model Uncertainty for Storm Water Quantity and Quality Analysis in DC Urban Area*, investigates statistical methodologies aiming to model the uncertainty associated with parameters of an environmental model. An empirical Bayesian approach combined with Regression analyses was used to estimate the uncertainty associated with hydraulic model parameters of river water quality model. Reliability of estimates obtained from different models is directly related to the management decisions in environmental or water resource management; i.e. for reducing combined sewer overflows (CSO) or improving surface water quality in the District of Columbia. Due to their prediction capacity and cost effectiveness, mathematical models have become an attractive tool. Model output is an estimate of the real measurement, and therefore its reliability depends partly on the relevance of model parameters and data

Research Program Introduction

gathered. We presented some preliminary results that help to better understand the behavior of an environmental model. A method for estimating the amount of uncertainty in the hydraulic model parameters of river water quality model was introduced. Analysis shows that beta does not differ much among two segments, while alpha appeared to be very different.

In the project Clam Active Biomonitoring and POM Passive Monitoring for Anacostia Watershed Contaminant Point Sources, Dr. Harriette L. Phelps, confirmed and extended knowledge of chlordane and PCB contaminant point sources in the Anacostia watershed. They explored POM as a new method for PCB congener analysis that could replace ABM. POM measured 86 PCB congeners while ABM measured 17 PCB congeners. POM detected mono congeners and ABM did not. POM plastic strips absorbed about twice the weight of PCBs as ABM clam tissue. However if PCB congeners 5, 8 and 28 are not included then ABM and POM results are statistically the same. PCB congeners 5, 8 and 28 were high for POM but not reported for ABM. POM and ABM detected a relatively similar pattern of di, tri, tetra, penta and hexa PCB congeners with an emphasis on the lower molecular weight congeners (di, tri, tetra). ABM did not detect mono chlorine PCB congeners. These heavier congeners are more toxic and found in fish. Publicity about these projects has generated increasing interest and involvement of citizen stream groups and the Maryland Department of the Environment in finding Anacostia watershed contaminant point sources. Point source contaminant remediation requires problem recognition by state environmental agencies leading to EPA involvement and this has been an important first step towards controlling Anacostia River's toxicity.

In Dr. Xueqing Song's project, Speciation of Some Triorganotin Compounds in Anacostia and Potomac River Sediments using NMR Spectroscopy, the speciation of Three tributyltin compounds (TBTs), tributyltin chloride (TBTCI), Bis(tributyltin) Oxide (TBTO) and tributyltin acetate (TBTOAc) under varying pH conditions (5, 7 and 9) was studied by NMR spectroscopy in both anaerobic and aerobic Anacostia River sediments. All TBT were found to first convert to a hydrated TBT species and then further decomposition depends on the speciation time and the nature of the sediments. The ^{119}Sn NMR chemical shifts of the spiked sediments indicated that changes in the pH did not affect the speciation of the tin compounds in either aerobic or anaerobic sediments. Dealkylation to mono/dibutyltin species was observed when speciation time is 4 weeks or longer. This dealkylation is very limited as the signal around -341 ppm is very weak for all sediment samples. This would suggest that the decomposition of toxic TBTs to low toxic DBT or MBT should take more than 8 weeks in sediment. A Comparison of the strength of signal of dealkylation species and undecomposed TBT species revealed that only less than 5% was decomposed to less toxic DBT or MBT.

Listed below are the eight grants awarded to researchers for FY 2010 104B grants:

Title: A Hierarchical Spatio-Temporal Dynamical Model for Predicting Precipitation Occurrence and Accumulation, Dr. Ali Arab, Assistant Professor, Department of Mathematics and Statistics, Georgetown University.

Title: Determining the Effectiveness of the Design-Build Method on Water Infrastructure Rehabilitation Projects in the District of Columbia, Dr. Kunhee Choi Assistant Professor of Construction Management Engineering, Architecture & Aerospace Technology, University of the District of Columbia

Title: DC Water Issues Forum and Water Research Faculty Professional Development Program, Dr. Tolessa Deksissa, Research Associate. Agricultural Experiment Station & Water Resources Research Institute, University of the District of Columbia

Title: Development of Analytical Tools to Evaluate the Performance of Low Impact Developments in the District of Columbia, Dr. Arash Massoudieh, Assistant Professor, Civil Engineering Department, The Catholic University of America

Research Program Introduction

Title: Determination of seasonal source variation of hydrocarbons, fatty acids, organics and nutrients in the Anacostia River: Stable isotope ratios of specific compounds, Dr. Stephen E. MacAvoy, Department of Environmental Science, American University

Title: Comparing Clam Active Biomonitoring and POM Passive Monitoring for DC Watershed Contaminant Point Sources (Phase II), Dr. Harriette Phelps, Professor Emeritus, Department of Biological and Environmental Sciences, University of the District of Columbia

Title: Speciation of Some Triorganotin Compounds in Anacostia and Potomac River Sediments using NMR Spectroscopy (Phase II), Dr. Xueqing Song, Associate Professor, Department of Chemistry, University of the District of Columbia

Title: The Application of Multiple-Antibiotic-Resistance (MAR) Profiles of Coliforms to Detect Sources of Bacterial Contamination of the Anacostia River, Dr. David W. Morris, Department of Biological Sciences, The George Washington University

Matching requirements were met with non federal in-kind contributions from the indirect cost waived by each university and cash match from the University of the District of Columbia. These research projects will provide water quality training for graduate and undergraduate students in the District of Columbia.

Modeling Model Uncertainty for Storm Water Quantity and Quality Analysis

Basic Information

Title:	Modeling Model Uncertainty for Storm Water Quantity and Quality Analysis
Project Number:	2009DC100B
Start Date:	3/1/2009
End Date:	2/28/2010
Funding Source:	104B
Congressional District:	District of Columbia
Research Category:	Water Quality
Focus Category:	Models, Methods, Wastewater
Descriptors:	None
Principal Investigators:	Valbona Bejleri, Tolessa Deksissa

Publications

There are no publications.

Modeling Model Uncertainty for Storm Water Quantity and Quality Analysis in DC Urban Area

Final Report



UNIVERSITY OF
THE DISTRICT
OF COLUMBIA



Valbona Bejleri, Ph.D.
Associate Professor
Department of Mathematics

Tolessa Deksissa, Ph.D.
Research Associate
Water Resource Research Institute &
Agriculture Experiment Station

University of the District of Columbia

May 2010

Modeling Model Uncertainty for Storm Water Quantity and Quality Analysis in DC Urban Area

ABSTRACT

Reliability of estimates obtained from different models is directly related to the management decisions in environmental or water resource management; i.e. for reducing combined sewer overflows (CSO) or improving surface water quality in the District of Columbia. Due to their prediction capacity and cost effectiveness, mathematical models have become an attractive tool. Model output is an estimate of the real measurement, and therefore its reliability depends partly on the relevance of model parameters and data gathered. In this study, we investigate statistical methodologies aiming to model the uncertainty associated with parameters of an environmental model. An empirical Bayesian approach combined with Regression analyses was used to estimate the uncertainty associated with hydraulic model parameters of river water quality model. We present some preliminary results that help to understand better the behavior of an environmental model.

INTRODUCTION

In the District of Columbia, for every inch of water there are combined sewer overflows (CSO). Low Impact Development (LID) has been proposed to address this problem (DCWASA, 2005). In order to evaluate the effectiveness of LID in reducing CSO or improving surface water quality in the District, application of appropriate mathematical model is required (Deksissa and Behera, 2008). In environmental or water resource management, mathematical models become more attractive tools than monitoring due to their prediction capacity and cost effectiveness.

The motivation for this research comes from the demand for reliable estimates, which has greatly increased in recent years due to their important role in decision making or in formulating policies. In case of model inputs, due to the cost or time frame, it is not always possible to collect/measure a large enough sample size yielding outcomes that fit within an acceptable uncertainty.

Model output is an estimate of the real measurement, and therefore its reliability depends partly on the relevance of model parameters and data gathered. The probabilistic measures of uncertainty associated with model outputs and model itself are crucial and can play an important role for administrators in management decisions regarding water resources.

While there exist several mathematical models in the literature applied to uncertainty analysis, there is still need for better approaches that can apply to complex urban wastewater management. Uncertainty on water-quality model outputs has been

investigated from (Bobba et al., 1996; Haan et al., 1998; Hession and Storm, 2000). Melching and Bauwens (2001) implemented a first error analysis and a Latin hypercubic sampling in coupled modeling system. Their work showed that uncertainty analysis was better method for identifying key sources of model uncertainty. Goethals and De Pauw (2001) used data driven modeling techniques and compared the outcome of the data driven models to expert rules from literature. Verdonck (2003) analyzed a range of statistical techniques comparing among parametric and nonparametric methods. Clement et al. (2004), worked on the development of spatio-temporal models for river assuming additive models and utilizing historical data.

The purpose of this research is to develop statistical tools that will help to measure the uncertainty in the input, output and model parameters, toward development of a model and statistical methods that will adjust for both model and parameter uncertainties. In this preliminary study, the empirical Bayesian approach combined with Regression analyses was applied to estimate the uncertainty associated with hydraulic model parameters.

METHODS

In this study, we utilized predictive inferences within the exponential family and construct prediction limits for two parameters of interest of the hydraulic model (Bejleri and White, 2005; and Bejleri, 2005). In non-tidal hydraulic model, the water balance in a given river stretch can be described by the following equations (Deksissa *et al.*, 2004):

$$\frac{dV}{dT} = Q_{in} - Q_e - (ET)A \quad (1)$$

$$Q_e = \alpha h^\beta \quad (2)$$

Where V is flow volume at a time T [m^3]; Q_{in} is inflow rate [m^3d^{-1}]; Q_e is effluent/outflow rate [m^3d^{-1}]; h is water depth at a time T [m], and α and β are stage flow relationship.

There is an uncertainty associated with the calculated coefficients alpha and beta of effluent flow rate process (2) which needs to be assessed. The exponential relationship in (2) was transformed into a linear relationship: $\ln(Q_e) = \ln(\alpha) + \beta \ln(h)$. Then, the uncertainty about α and β was estimated using empirical Bayesian approach (Carlin and Louis, 2000) and regression analysis. The assumption made, based on preliminary data analysis, was that the log transformed inflow rate and water depth data follow almost a normal distribution. Under this assumption, one can write the following equations: $x_h = \ln(h) \sim N(\theta_h, \sigma_h^2)$ and $y_Q = \ln(Q_e) \sim N(\theta_Q, \sigma_Q^2)$, where the symbol “ \sim ” stands for the words “is distributed”. A normal prior distribution was adopted for the mean

parameters $\underline{\theta}^T = (\theta_h, \theta_Q)$; $\theta_h \sim N(\mu_h, \tau_h^2)$ and $\theta_Q \sim N(\mu_Q, \tau_Q^2)$. Then, using posterior analyses we estimate the distribution of the parameter given the data: $\underline{\theta}^T | \text{data}, \ln(h)$ and $\ln(Q_e)$. The posterior means are given by the following equations:

$$E(\hat{\theta}_h | x_h) = \hat{B}_h \bar{x}_h + (1 - \hat{B}_h) x_h, \quad \text{Var}(\hat{\theta}_h | x_h) = (1 - \hat{B}_h) \sigma_h^2$$

$$E(\hat{\theta}_Q | y_Q) = \hat{B}_Q \bar{y}_Q + (1 - \hat{B}_Q) y_Q, \quad \text{Var}(\hat{\theta}_Q | y_Q) = (1 - \hat{B}_Q) \sigma_Q^2$$

$$\hat{B}_h = \left[\sigma^2 (K - 3) \right] / \sum_k (x_{hk} - \bar{x}_h)^2, \quad \text{and} \quad \hat{\mu}_h = \bar{x}_h,$$

$$\hat{B}_Q = \left[\sigma^2 (K - 3) \right] / \sum_k (y_{Qk} - \bar{y}_Q)^2, \quad \text{and} \quad \hat{\mu}_Q = \bar{y}_Q, \quad \text{where} \quad k = 1, 2, 3, \dots, K$$

The vector parameter consisting of these posterior means was used to generate a dataset of 100 observation ($\ln(h)$ and $\ln(Q_e)$). Then, α and β was calculated using regression analysis. The line of best fit for $\ln(Q_e)$ versa $\ln(h)$ is given by the equation $\ln(Q_e) = \ln(\alpha) + \beta \ln(h)$.

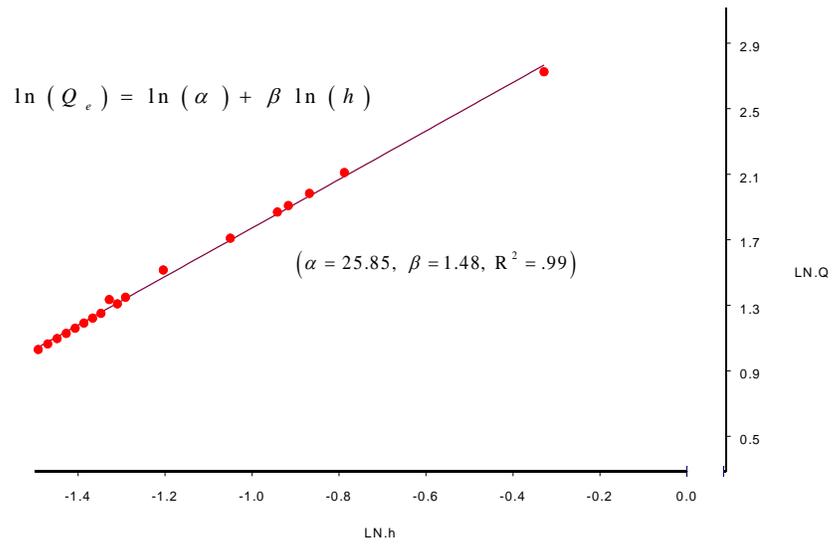
Data used for illustration of the proposed method are presented in Table1. To assure normality, of the data Log transformation was performed (Table 2). Figure 1 shows the line of best fit for the transformed inflow rate and water depth.

Table 1. Effluent flow rate (Q_e) and hydraulic depth (h) measured at two different segments

	Segment 1		Segment 2	
	h	Q_e	h	Q_e
Min	0.17	2	0.01	0.12
Mean	0.28	4.14	0.13	7.01
Median	0.25	3.19	0.12	5.14
Max	0.72	15.25	0.6	52
Std Dev	0.12	2.79	0.11	8.8

Table 2. Mean and standard deviation for the transformed measurements (for two different river segments)

		Mean	Std
Segment 1 K=27	Ln(h)	-1.33	0.33
	Ln(Q_e)	1.28	0.49
Segment 2 K=44	Ln(h)	-2.32	.84
	Ln(Q_e)	1.35	1.26



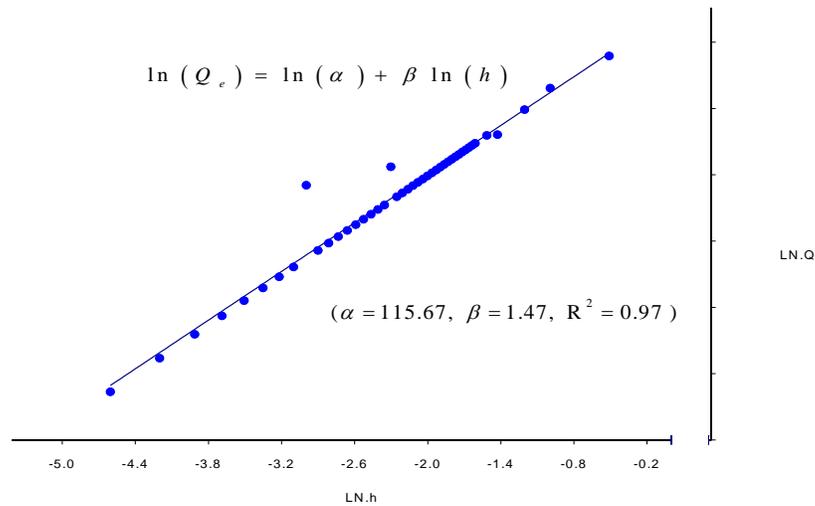


Figure 2. The line of best fit of effluent flow rate $\ln(Q_e)$ vs. hydraulic depth $\ln(h)$ for river segment 2.

RESULTS

Empirical Bayesian approach combined with Regression analyses (Fig.1 and 2) was used to estimate the variability in coefficients alpha and beta of effluent flow rate process in the hydraulic model. The procedure (as described in the methods section) was repeated 100 times resulting on samples of size 100 for α and β . Their respective means serve as point estimates for α and β . In table 3 are presented the posterior mean and standard deviations for the probability distributions of both alpha and beta for each segment considered. In figure 3 are graphed posterior distributions for both alpha and beta for each segment considered (a) segment 1 and (b) segment 2. Preliminary analysis shows that beta does not differ much among two segments, while alpha appeared to be very different.

The research on modeling uncertainty extends beyond this project. In this report, we presented some preliminary results that help to better understand the behavior of an environmental model. A method for estimating the amount of uncertainty in the hydraulic model parameters of river water quality model was introduced.

Table 3. Mean and standard deviation of the posterior probability distributions of alpha and beta for two river segments

	Segment 1		Segment 2	
	alpha1	beta1	alpha2	beta2
Mean	19.97	1.34	116.58	1.27
Std Dev.	6.14	0.23	9.51	0.22

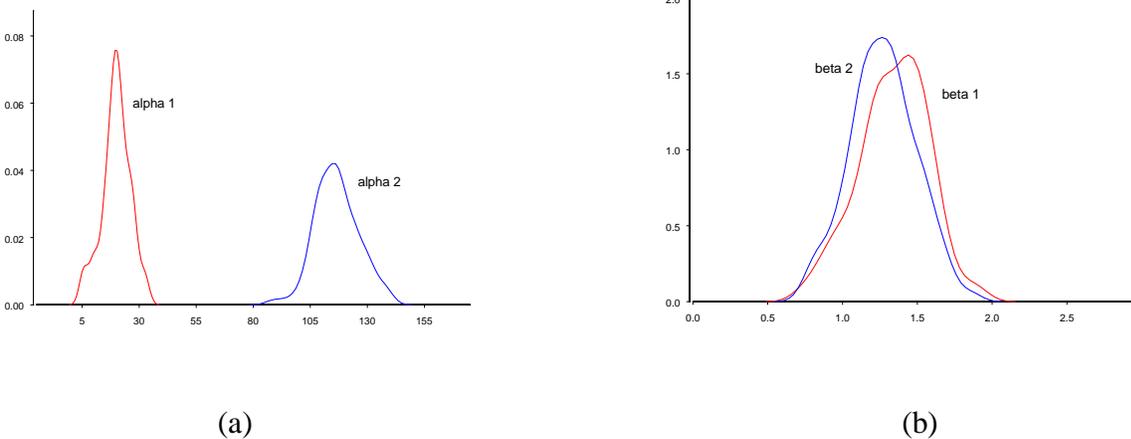


Figure 3. Posterior distributions of alpha (a) and beta(b) for two river segments.

KNOWLEDGE TRANSFER/PRESENTATION

The results of these study were presented at the 2009 Joint Statistical Meetings (Washington DC) and the First Annual UDC Research Mini-Symposium. The presentation is attached with this report.

ACKNOWLEDGEMENT

This work was financially supported by DC Water Resource Research Institute. Preliminary data were provided by Dr. Deksissa and his intern students.

REFERENCES

- Bejleri, V. and White, A., 2005. Bayesian prediction limits for Atlantic tropical storm occurrences *ASA Proceedings of the Joint Statistical Meetings*, 19-24
- Bejleri, V., 2005. Bayesian Prediction Intervals for the Poisson Model, Noninformative Priors, Ph.D. dissertation, America University, USA
- Beck, M.B. and Reda, A. (1994). "Identification and application of a dynamic model for operation management of water quality", *Wat. Sci. Tech.* 30(2), 31-41
- Bobba, A.G., Singh, V.P., and Bengtsson, L (1996). "Application of first-order and Monte Carlo analysis in watershed water quality models" *Water Resour. Manage.*, 10, 219-240.
- Carlin B. and Louis T. A. (2000). "Bayesian and Empirical Bayesian Methods, Second edition"
- Casella, G. and Berger, R. L., 2002. *Statistical Inference*, (Second Edition), Wadsworth Group, p 66-67.
- Clement, L., Thas, O., Vanrolleghem, P.A. and Ottoy, J.P., 2004. Spatiotemporal statistical models for river monitoring networks. In: *Proceedings of the 6th International Symposium on Systems Analysis and Integration Assessment (WATERMATEX 2004)*, Beijing, China, November 3-5, 2004.
- Coleman H. W. and Steele, W. G., 1999. *Experimentation and Uncertainty Analysis for Engineers*, (Second Edition), John Wiley & Sons, Inc.
- Cullen A. C. and Frey, H. C., 1999. *Probabilistic Techniques in Exposure Assessment*, Plenum Press, Plenum Publishing Corporation.
- Deksissa, T., 2004. *Dynamic Integrated Modeling of Basic Water Quality and fate and effect of organic contaminants in rivers*, Ph.D. dissertation, Ghent University, Belgium, ISBN 90-5989-015-9.
- Deksissa, T., Meirlaen, J., Ashton, P. J. and Vanrolleghem, P. A. (2004). "Simplifying Dynamic River Water Quality Modelling: A case study of inorganic Nitrogen Dynamics in the Crocodile River (South Africa) Water, Air and Soil Pollution 155: 303-320
- Deksissa T. and Behera P., 2008. Modeling Integrated Urban Wastewater System: Model Selection and Implementation. In: *proceedings of the 2008 Mid-Atlantic Regional Water*

Resource Research Conference: The Energy & Water Nexus: Water's Role in the Challenges of Energy Production in the 21st Century, November 17 –19, 2008, Shepherdstown, WV.

District of Columbia Water and Sewer Authority (DCWASA), 2005. Long Term Control Plan Consent Decree Status Report: Quarter No. 1 – 2005

Gelman, Andrew, Carlin, B. John, Stern, S. Hal, and Rubin, B. Donald (2004), *Bayesian Data Analysis*, (Second Edition), Chapman and Hall/CRC.

Goethals, P. & De Pauw, N. (2001). Development of a concept for integrated river assessment in Flanders, Belgium. *Journal of Limnology* 60: 7-16

Hann, C. T., D.E., Al-Issa, T., Prabhu, S., Sabagh, G.J., and Edwards, D. R. (1998). "Effect of parameter distribution on uncertainty analysis of hydrologic models." *Trans ASAE*, 41(1), 65-70

Hession, W. C., and Storm, D. E. (2000). "Watershed – level uncertainty: Implications for phosphorous management and eutrophication" *J. Environ. Qual.*, 29, 172-179

Melching, C.S., and Bauwens, W. (2001). "Uncertainty in coupled non point source and stream water-quality models." *J. Water Resour. Plan. Manage.*, 127 (6), 403-412

Reichert, P., Borchard, D., Henze, M., Rauch, W., Shanahan, P., Somlyódy, L., and Vanrolleghem, P. A. (2001). *River Water Quality Model N0.1 (RWQM!)*, Scientific and Technology Report, IAWQ, London

Rossman, L. A., 2004. Storm water management model user's manual Version 5.0, United State Environmental Protection Agency.

Verdonck, F.A., 2003. Geo-Referenced probabilistic ecological risk assessment. Ph.D. Dissertation, Ghent University, Ghent, Belgium.

Clam Active Biomonitoring and POM Passive Monitoring for Anacostia Watershed Contaminant Point Sources

Basic Information

Title:	Clam Active Biomonitoring and POM Passive Monitoring for Anacostia Watershed Contaminant Point Sources
Project Number:	2009DC101B
Start Date:	3/1/2009
End Date:	2/28/2010
Funding Source:	104B
Congressional District:	District of Columbia
Research Category:	Water Quality
Focus Category:	Surface Water, Solute Transport, Toxic Substances
Descriptors:	None
Principal Investigators:	Harriette Phelps

Publications

There are no publications.

Clam Active Biomonitoring and POM Passive Monitoring for Anacostia Watershed Contaminant Point Sources

Final Report



UNIVERSITY OF
THE DISTRICT
OF COLUMBIA



Harriette Phelps, Ph.D.
Professor Emeritus

Department of Biological and Environmental Sciences
University of the District of Columbia

May 2010

Clam Active Biomonitoring and POM Passive Monitoring for Anacostia Watershed Contaminant Point Sources

Final Report to the DC Water Resources Research Center
Dr. Harriette L. Phelps

Abstract

In 2008 and 2009 active biomonitoring (ABM) with freshwater clams (*Corbicula fluminea*) was carried out at sites in the Anacostia River watershed to survey sources of EPA Priority Pollutants and focus on the polychlorinated biphenyls (PCBs) and chlordane responsible for the Anacostia fishing advisory. A total of seven projects were carried out: (1) in Still Creek which runs through Greenbelt National Park identify the source area of chlordane contamination, (2) survey Wells Run at University Park, MD for all EPA Priority Pollutants, (3) at the lower tidal Anacostia sediment ‘hot spot’ of Stickfoot Sewer look for increased EPA Priority Pollutants in clams. (4) examine the median stream of the Baltimore-Washington Parkway for relationship to the nearby Riverdale East stream chlordane contamination, (5) in Sligo Creek locate the source of the high chlordane contamination, (6) in upper Lower Beaverdam Creek identify the outlet source of polychlorobiphenyls (PCBs) and (7) in upper Lower Beaverdam Creek compare active PCB biomonitoring with clams (ABM) to passive PCB monitoring using polyoxymethylene plastic strips (POM). All projects were successful. In summary: (1) Still Creek chlordane contamination was found to originate outside the National Park, (2) Wells Run had polycyclic aromatic hydrocarbons (PAH) from combustion and chlordane exceeding reference, (3) clams placed near the Stickfoot Sewer outlet indicated the bioavailable contaminants did not exceed other tidal river sites except for 2X PAHs, (4) the Baltimore Washington Parkway median stream lacked the high chlordane of nearby Riverdale East so the Parkway may not be a source (5) The Sligo Creek chlordane source was in the upstream Main Branch, (6) The Lower Beaverdam Creek PCB outlet source was located above the previously considered outlet and (7) The Lower Beaverdam Creek ABM and POM PCB results at two and four weeks were not the same but both found high low-molecular-weight congeners upstream and were statistically the same when congeners 5, 8 and 28 were excluded.. This report includes results from previously unreported 2008 Anacostia watershed ABM studies. Projects were based on earlier ABM studies of contaminants in the Anacostia watershed and some results are being used for further investigation by the Maryland Department of the Environment.

Introduction

The tidal freshwater Anacostia River that flows from Maryland through DC to the Potomac River is one of three toxic Regions of Concern in the USEPA/NOAA Reference: Phelps, H.L. 2010. Clam Active Biomonitoring and POM Passive Monitoring for Anacostia Watershed Contaminant Point Sources. DC WRRI, Washington, DC. 11p.

Chesapeake Bay Program (Chesapeake Bay Program 1999) and listed among America's 10 worst rivers (<http://mapping2.orr.noaa.gov/portal/AnacostiaRiver/>). The Anacostia has a fishing advisory based on high polychlorinated biphenyls (PCBs) and chlordane in fish tissue. The sediment in the 10 km tidal river is toxic to benthic life (Phelps 1993) and the tumors in over 60% of resident fish are probably due to high sediment PAHs (Pinkney et al. 2000). The Anacostia tidal sediments were extensively studied from 1999 to 2002 by an EPA/NOAA partnership (Wade et al. 1994, Phelps 1995, Coffin et al. 1999, SRC 2000, AWT 2002, NOAA 2002) and the 2002 Anacostia toxics remediation plan was developed to cap tidal sediment "hot spots". The most recent toxics remediation plan added controlling stormwater runoff from tributaries (Gruessner et al. 1997, Warner et al. 1997, Washington Post 2004a, ARP 2010). However, active biomonitoring (ABM) studies were finding point sources of contaminants in the tributaries (Phelps 2002, Phelps 2003, Phelps 2004, Phelps 2005, Phelps 2008 Washington Post 2008a, Chesapeake Journal 2009). ABM used the locally available Asiatic clam (*Corbicula fluminea*) (Dressler and Cory 1980) known as a freshwater contaminant bioaccumulator (Dougherty 1990) that can detect low-level and variable levels of bioavailable contaminants at specific watershed locations (DeKock and Kramer 1994). *Corbicula* ABM for 62 EPA Priority Pollutants and seven metals in 13 major Anacostia subwatersheds found PCBs associated with an industrial park and 80% of pesticides as chlordane associated with legacy dump sites. ABM also determined that toxic metals were not an Anacostia problem and high PAHs were associated with industrial parks and parking lot runoff but not with coal-tar sealcoating (Phelps 2008).

The Anacostia has a Total Maximum Daily Load (TMDL) for PCBs and the Maryland Department of the Environment (MDE) is investigating a PCB source found by ABM in upper Lower Beaverdam Creek (Phelps 2003). ABM is limited to water temperatures over 50 deg. C. but year-round PCB congener monitoring using polyoxymethylene plastic strips (POM) is being developed by Dr. Ghosh of the University of Maryland Baltimore Campus (UMBC) (Sun and Ghosh 2008). Dr. Ghosh participated in the monitoring study comparing ABM and POM for PCB congener detection in Lower Beaverdam Creek.

Methods

In 2008 and 2009 active biomonitoring for contaminants in the Anacostia watershed was conducted using methods previously described (Phelps 2008). *Corbicula* clams (17 – 23 mm shell height) were collected by sieving the sandy Potomac River shoreline at the reference site of Fort Foote (MD), 5 km downstream from the mouth of the Anacostia. Clams were kept cool and dry and translocated in shellfish mesh bags within 6 hours to Anacostia watershed biomonitoring sites (Table 1, Figure 1). A Fort Foote Potomac (FF) clam reference sample was taken for analysis. ABM clams were deployed for two weeks except at Lower Beaverdam Creek where comparison was made among two and four week deployments for PCBs alongside polyoxymethylene (POM) plastic strips.

Results

Table 1. Corbicula Active Biomonitoring Anacostia Site Data 2009.

Sample		northing	westing
Dates/Site	Analysis		
7/31/09, 9/13/09			
FF (Forte Foote reference)	ALL	38 ⁰ 46'27.27"	76 ⁰ 01'45.50"
7/31-8/15/09			
A1 (2 week above Landover Metro site)	PCB	38 ⁰ 56'42.10"	76 ⁰ 52'15.66"
B1 (2 week below, at Landover Metro site)	PCB	38 ⁰ 55'56.38"	76 ⁰ 53'21.49"
UST (Upper Still Creek)	PEST	38 ⁰ 59'13.31"	76 ⁰ 52'05.97"
SCM (Sligo Creek Main Branch)	PEST	39 ⁰ 01'14.77"	77 ⁰ 01'58.03"
SCW (Sligo Creek Wheaton Branch)	PEST	39 ⁰ 01'14.75"	77 ⁰ 01'59.39"
BWP (BW Parkway Median)	PEST	38 ⁰ 59'15.84"	76 ⁰ 54'29.63"
7/31-8/28/09			
A2 (4 week above Landover Metro site)	PCB	38 ⁰ 56'42.10"	76 ⁰ 52'15.66"
B2 (4 week below, at Landover Metro site)	PCB	38 ⁰ 55'56.38"	76 ⁰ 53'21.49"
9/13-9/29/09			
PP (Poplar Point)	ALL	38 ⁰ 52'11.12"	79 ⁰ 59'52.65"
WRC (Wells Run Creek)	ALL	38 ⁰ 53'08.87"	76 ⁰ 56'32.13"

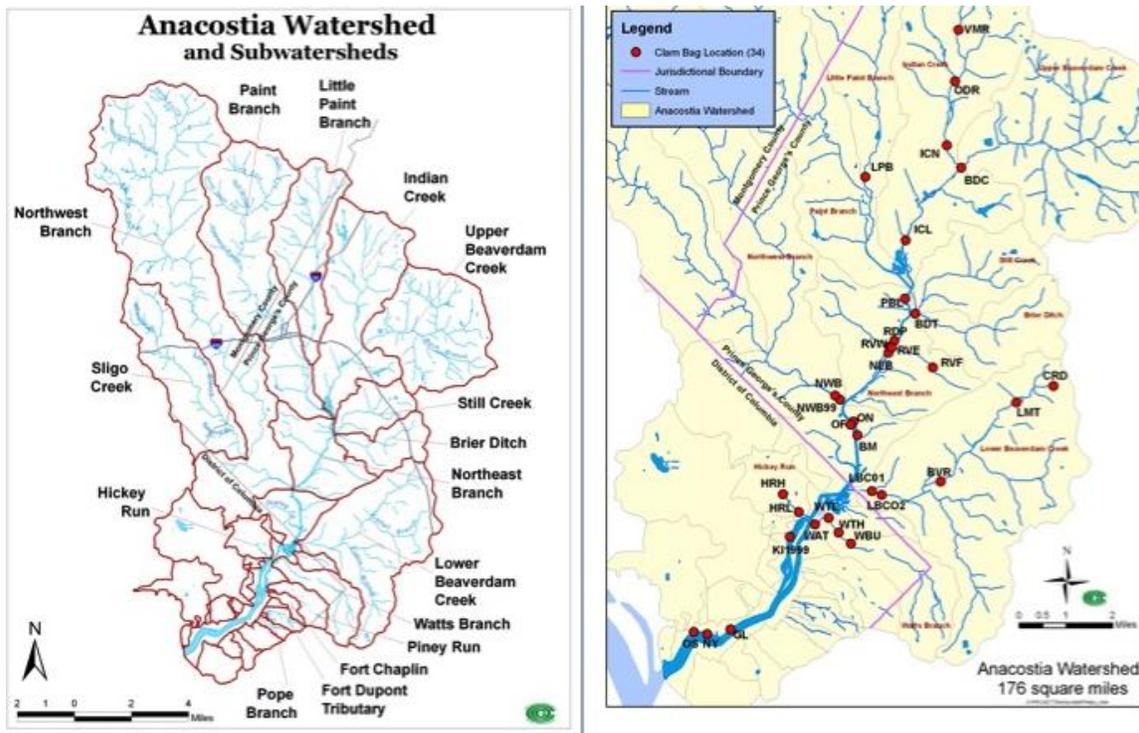


Figure 1. Maps of the Anacostia Watershed and several ABM monitoring sites.

Still Creek flows into to the Northeast Branch of the Anacostia (Fig. 1) and has a watershed of 4 square miles of which 43% is Greenbelt National Park (Fig. 2). In 2004 a complete ABM study at the mouth of Still Creek (Phelps 2005) found chlordane the only major EPA Priority Pollutant contaminant. In 2007 ABM at four first order streams within the Park found high chlordane only in the Upper Mainstem (Fig. 2). In 2009 active biomonitoring in Upper Mainstem outside the Park (site UST) found total chlordane (280 ppb) heptachlor epoxide (13 ppb), gamma chlordane (34 ppb) and alpha chlordane (35 ppb). The total chlordane of 280 ppb is close to the USFDA 300ppb UAFDA action level for human health. The Still Creek Upper Mainstem is a small first order tributary originating in a Greenbelt MD suburb with no known industry. Heptachlor epoxide is a chlordane breakdown product and the finding of 5% heptachlor epoxide suggested Still Creek contamination may originate from a legacy chlordane dump site created when chlordane use for termites was banned in 1983.

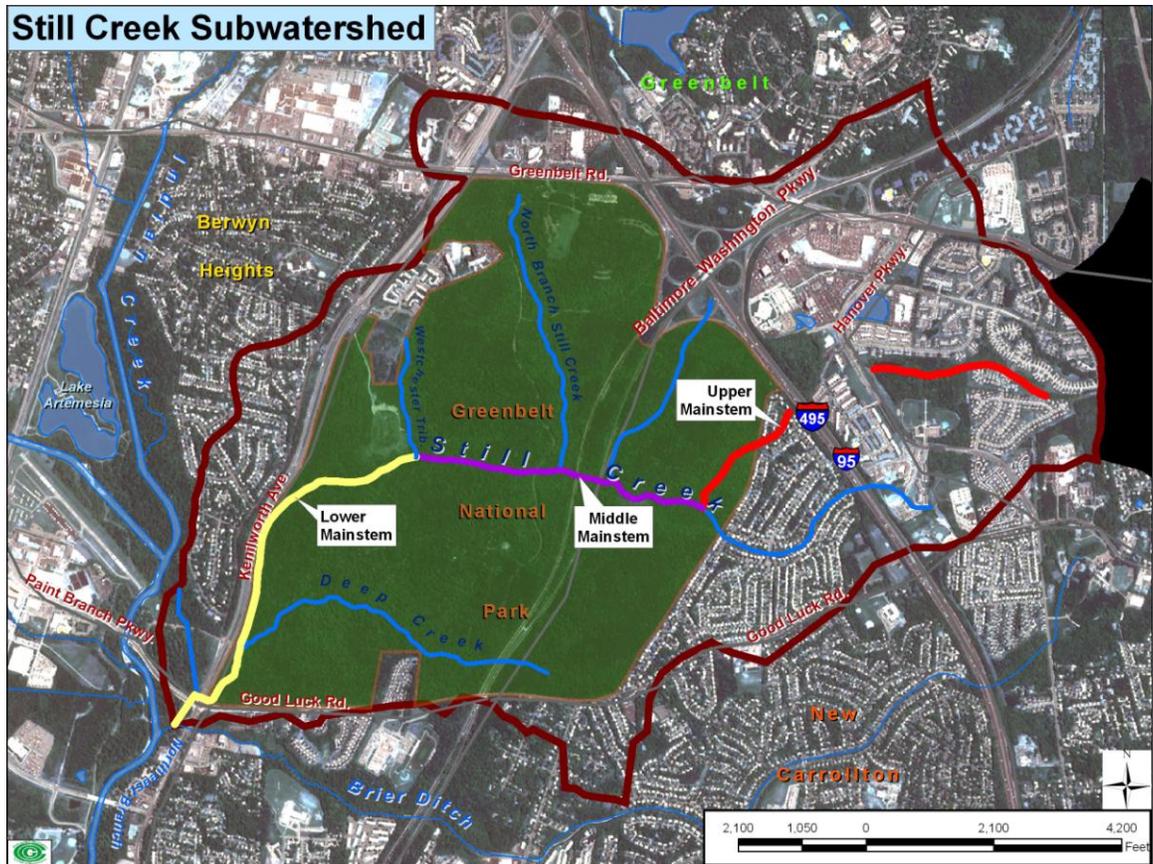


Figure 2. Still Creek watershed.

Wells Run is a small subtributary of the Northeast Branch that runs through University Park, MD. It has an active citizen group concerned about stream health that asked to have an ABM scan. One of the purposes in developing ABM was to encourage evaluation of stream contaminants in the entire Anacostia watershed. Clams were placed at site WRC in Wells Run for two weeks in September 2009. Wells Run clam contaminants compared with FF reference levels found no significant increase in PCB congeners, aroclors or metals. However total PAHs (436 ppm) statistically exceeded FF reference (104 ppm) and were similar to average high tPAHs in the Anacostia River (360 ppm). The majority were 4-5 ring PAHs typical of combustion such as auto exhaust and burning. Clams at site WRC also had significantly increased total chlordane (240 ppb) with gamma (17ppb) plus alpha (29ppb) chlordane as 19.2% of total chlordane. This could indicate an upstream chlordane source although heptachlor epoxide was not detected.

Poplar Point in the lower tidal Anacostia is an AWT toxic sediment ‘hot spot’ including the outlet site of Stickfoot Sewer (site PP) (NOAA 2002). Stickfoot Sewer is enclosed so clams were placed in the tidal river close to the outlet. ABM in 2001 and 2002 at five other tidal Anacostia River sites had found no significant contaminant differences, probably due to tidal mixing (Phelps 2002, Phelps 2003). Like ABM at three

other tidal river 'hot spots', tPCBs and tMetals accumulated by clams at site PP did not exceed reference (site FF). Total clam PAHS (681 ppm) exceeded reference (122 ppm) and were twice the Anacostia tidal site average (360 ppm) The only detected pesticide was high total chlordane (1200 ppb) which included gamma (82 ppb) and alpha (140 ppb) chlordane (18.3%) and exceeded reference (120 ppb) (site FF). There was no detectable heptachlor epoxide so a legacy source of chlordane was not established.

The small Riverdale East culverted stream entering the Northeast Branch (Fig. 1) had been found highly contaminated with chlordane (site RVE, 720 ppb chlordane) that increased up to the Baltimore Washington Parkway (site RVF, 1800 ppb chlordane) before ending in a suburb (Fig. 3) (Phelps 2003, Phelps 2004). The chlordane was accompanied by heptachlor epoxide, a chlordane degradation product, and it was suggested the origin could be legacy dump sites in the vicinity of the Baltimore-Washington Parkway. There is a small stream running down the forested median strip of the Baltimore-Washington Parkway near site RVF. Clams placed at that location (site BWP) for two weeks had twice the total chlordane (220 ppb) of reference (FF site) (100 ppb) but much lower than the nearby Riverdale East site RVF and did not have heptachlor epoxide. It appeared that chlordane in the Parkway median stream (site BWP) did not have the same origin as high chlordane in the nearby Riverdale East stream.

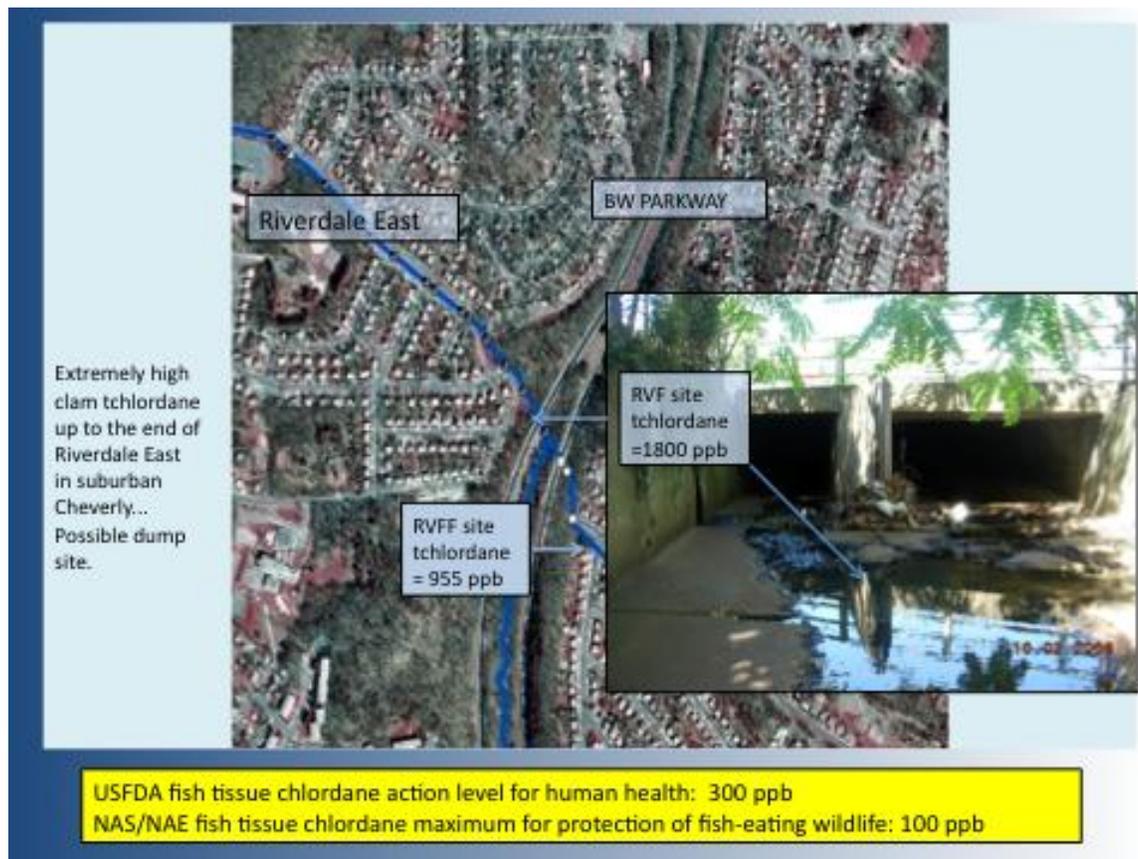
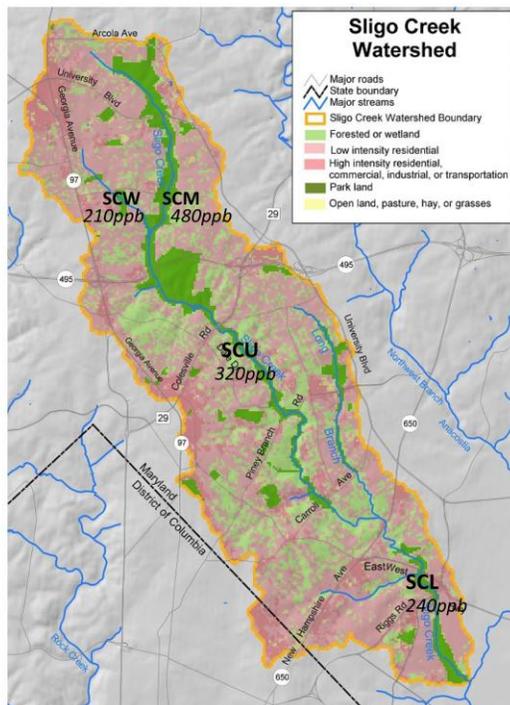


Figure 3. Riverdale East and the Baltimore Washington Parkway.

Sligo Creek is a large subtributary of the Northwest Branch with a watershed mostly in Montgomery County and parts in Prince Georges County and DC (Fig 1). Sligo Creek has a active citizens group with a number of restoration projects and was selected as a model for the Anacostia Restoration Program being developed by the Army Corp of Engineers (ACE) and the Metropolitan Washington Council of Governments (MWCOC) (Washington Examiner 2008; Washington Post 2008b). In 2008 students carried out complete ABM surveys at two sites in the lower Creek. PCB congeners did not exceed reference (site FF) and total PAHs were about twice reference but significant pesticides were detected: dieldrin, heptachlor epoxide and chlordane. Total chlordane was lowest (240 ppb) downstream (site SCL) and higher (320 ppb) upstream (site SCU) (Fig. 4). In 2009 ABM upstream found lower total chlordane (230 ppb) at Wheaten Branch (site SCW) than in the Main Branch (480 ppb) (site SCM). Alpha plus gamma chlordane was 16.2% at site SCW) and 19.0% at site SCM) but dieldrin (43 ppb) and heptachlor epoxide (32 ppb, 6.7%) were only found at site SCM which suggests a legacy chlordane dump site. It should be noted that the reference (site FF) total chlordane was high (120 ppb) with 20 ppb gamma chlordane and no alpha chlordane or heptachlor epoxide. It is hoped to further explore the upstream Sligo Main Stem to locate the chlordane source.



Sligo Creek is a model stream being studied for stormwater control (COE)

In 2008 student active biomonitoring used clams to study 72 EPA Priority Pollutants at upper (SCU) and lower (SCL) Sligo Creek sites. Only total chlordane (and PAHs) significantly exceeded the Potomac (reference) site. Highest total chlordane was upstream.

In 2009 upper Sligo Creek clam active biomonitoring for total chlordane found lowest levels in Wheaton Branch (SCW) and highest in Main Branch (SCM). Heptachlor epoxide suggested an old chlordane dump in SCM significantly contaminating the entire creek. This will be explored further in 2010.

Total clam chlordane: 100 ppb = NAS/NAE maximum for protection of fish-eating wildlife; 300 ppb = USFDA action level for human health

Figure 4. Chlordane ABM studies in Sligo Creek.

Lower Beaverdam Creek (LBC) has the greatest percent of land in industrial parks (Warner et al. 1997) and highest polychlorinated biphenyl (PCB) pollution (Phelps 2002) (Fig. 1). High pesticides (mostly chlordane), PCBs and Aroclors were found

starting at lower Lower Beaverdam Creek (sites LBC01, LBC02) proceeding upstream to Beaver Road (site BVR) near Tuxedo Industrial Park, then Landover Metro (site LMT) below the Ardwick Ardmore Industrial Center, followed by New Carrollton Metro Station (site NCM) and Corporate Drive (site CRD) just inside the Beltway, Route 495 (Fig. 1) (Phelps 2004, Phelps 2005, Phelps 2008). The upstream sites CRD and NCM had high total pesticides (mostly chlordane) but no elevated tPCB or tAroclor. Sites NCM, BVR and LBC downstream from the Ardwick-Ardmore Industrial Park had high tPCB with low-molecular-weight volatile PCB congeners, (mostly Aroclor 1242 and 1254) suggesting an ongoing source. ABM with stream walking at three sites between LMT and NCM (sites AA2, AA3, AA4) found a 3X tPCB increase at site AA4, mostly low molecular weight congeners. (Fig. 5) (Phelps 2008). In 2009 additional ABM placed using stream walking between LMT and AA4 were able to identify a short stream reach containing an outlet associated with high PCBs which is now being explored by the Maryland Department of the Environment.

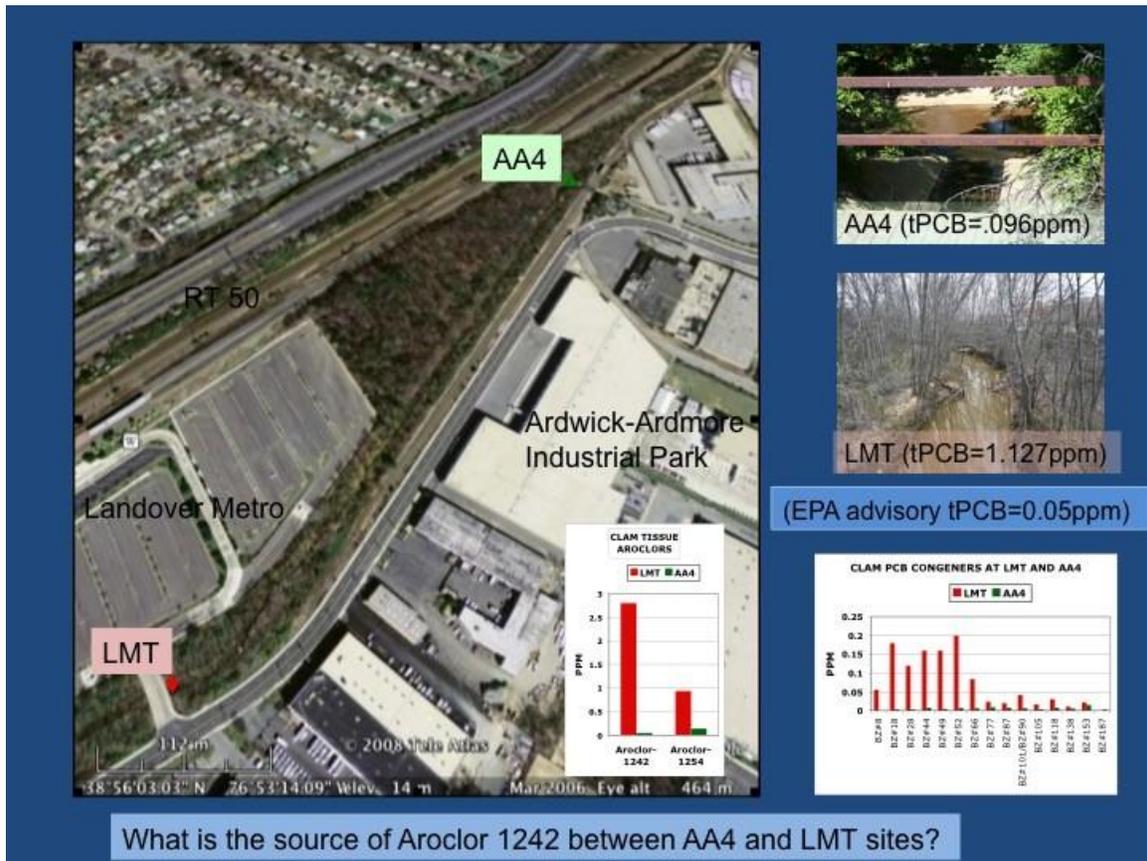


Figure 5. Lower Beaverdam Creek PCB congeners.

The second upper Lower Beaverdam Creek project compared PCB monitoring by clam ABM and polyoxymethylene strips (POM). Two monitoring sets including ABM clams and plastic polyoxymethylene strips (POM) were placed for two and four weeks at

Landover Metro (site LMT) called the B(elow) site and upstream at A(bove site close below the AA4 site where high PCBs were first detected (Phelps 2008). The POM strips were provided and analyzed by Dr. Upal Ghosh of UMBC. Samples UP A1 and DN B1 were collected at two weeks and samples UP A2 and DN B2 at four weeks. The dry POM strips were taken to Dr. Ghosh's laboratory for PCB congener extraction and analysis. The frozen ABM clam tissues were sent to the Philadelphia laboratory of TestAmerica for analysis. Dr. Ghosh's lab analyzed POM strips for 86 PCB congeners and TestAmerica analyzed the ABM clam tissues for 20 PCB congeners. POM PCBs were reported in units per dry weight while tissue PCBs are routinely reported in units per wet weight. Clam tissue is 80% water and if both results are reported in dry weight units, at two weeks the total PCBs by POM was greater (>2X) (Fig. 6). Both methods found much greater total PCB at site A (upstream) than site B (downstream, site LMT) and ABM at four weeks showed total PCB increase while POM total PCBs did not. Reference site (FF) ABM total PCB was 90 ppb. Both ABM and POM recorded peaks of low-molecular-weight PCB congeners. Three high PCB congeners (5, 8 and 28) were reported by POM analysis but not by ABM. If those congeners were excluded there was PCB congener statistical similarity detected by POM and ABM methods.

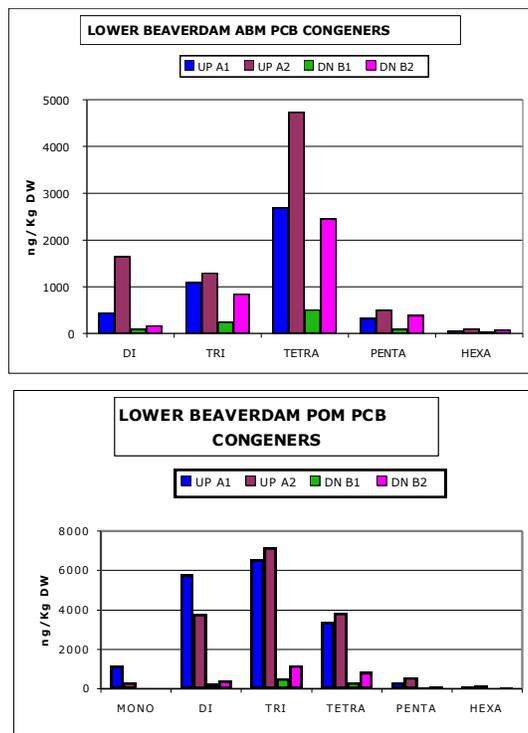


Figure 6. Lower Beaverdam Creek PCB congeners at sites A (upstream) and B (downstream) at two (A1, B1) and four (A2, B2) weeks using POM strips and active biomonitoring (ABM) with *Corbicula*.

These studies have confirmed and extended knowledge of chlordane and PCB contaminant point sources in the Anacostia watershed. They explored POM as a new method for PCB congener analysis that could replace ABM. The student projects were presented at school and results have appeared in articles (Chesapeake Journal 2009,

Washington Gazette 2010). Publicity about these projects has generated increasing interest and involvement of citizen stream groups and the Maryland Department of the Environment in finding Anacostia watershed contaminant point sources. Point source contaminant remediation requires problem recognition by state environmental agencies leading to EPA involvement and this has been an important first step towards controlling Anacostia River's toxicity.. A survey paper on the use of ABM in finding Anacostia watershed point sources of toxic contaminants is being prepared for publication.

Acknowledgements.

Grateful acknowledgement is made to the students who assisted in this research in 2008 and 2009: Gabriel Bell, Kelsey Corbitt, Jasmine Cook, Stephenie Eby, Molly Friedman, Alena Moran, Marika Nell, Neima Rahim, Kelsey van Bokkem, Caitlin Virta and Maia Werbos. Special acknowledgement goes to my reliable field assistant Earl Greenidge. Much thanks to the Eleanor Roosevelt High School Environment Club and the Friends of Lower Beaverdam Creek for funding in 2008.

References

- ARP. 2010. Anacostia Restoration Plan and Report.
http://www.anacostia.net/Restoration_Plan/download/Anacostia-Report-Web-Quality.pdf
- AWTA. 2002. Anacostia Watershed Toxics Alliance. Charting a Course Toward Restoration: A Contaminated Sediment Management Plan. Washington, DC
- Chesapeake Bay Program. 1999. Targeting toxics: A characterization report. A tool for directing management and monitoring actions in the Chesapeake Bay's tidal rivers: Annapolis, Maryland. Chesapeake Bay Program CBP/TRS 222/106 (EPA 903-R-99-010) 79p.
- Chesapeake Journal. 2009. Toxic hot spots in Anacostia, DC Creeks.
<http://www.bayjournal.com/article.cfm?article=3702>
- Coffin, RB, JW Pohlman and CS Mitchell. 1999. Fate and Transportation of PAH and Metal Contaminants in the Anacostia River Tidal Region. NRL/MR/6110-99-8327. Naval Research Laboratory, Wash., DC
- DeKock, WC and KJM Kramer. 1994. Active biomonitoring (ABM) by translocation of bivalve molluscs. In: Kramer, KJM (ed.) Biomonitoring of Coastal Waters and Estuaries. CRC Press, Boca Raton, FL.
- Dougherty F.S. 1990. The Asiatic Clam *Corbicula*-spp as a biological monitor in freshwater environments *Environ Mon Ass.* 15(2):143-188 51:269-313
- Dresler, P.V. and R.L. Cory. 1980. The Asiatic clam, *Corbicula fluminea* (Muller) in the tidal Potomac River, Maryland. *Estuaries* 3:150-151.
- Gruessner, B, DJ Velinsky, GD Foster, J Scudlark and R Mason. 1997. Final Report - Dissolved and Particulate Transport of Chemical Contaminants in the Northeast and Northwest Branches of the Anacostia River. Prepared DCRA, Environmental Regulation Administration, Washington, DC. ICPRB Report #97-10
- NOAA 2002. Watershed Database and Mapping Projects/Anacostia River.
<http://response.restoration.noaa.gov>
- Phelps, H.L. 1993. Sediment toxicity of the Anacostia River estuary Washington, DC. *Bull. Environ. Contam. Toxicol.* 51:582-587.

- Phelps, H.L. 1994. The Asiatic clam (*Corbicula fluminea*) invasion and system-level ecological change in the Potomac River estuary near Washington, DC. *Estuaries* 17(3):614-621.
- Phelps, H.L. 2000. DC's Contaminated Anacostia Estuary Sediments: A Biomonitoring Approach. DC Water Resources Research Center, Washington, DC. 10p.
- Phelps, H.L. 2001. PCB Congeners and Chlordane in Anacostia Estuary Sediments and Asiatic Clams (*Corbicula fluminea*): Possible Effects of Recent Dredging. 17p. DC Water Resources Research Center, Washington, DC. <http://www.his.com/~hphelps/>
- Phelps, H.L. 2002. Sources of Bioavailable Toxic Pollutants in the Anacostia (Part II). DC Water Resources Research Center, Washington, DC. <http://www.his.com/~hphelps/>
- Phelps, HL. 2003. *Corbicula* Biomonitoring in the Anacostia Watershed. DC Water Resources Research Center, Washington, DC. <http://www.his.com/~hphelps/>
- Phelps, HL. 2004. Sources of Bioavailable Toxic Pollutants in the Anacostia Watershed (Part III). DC Water Resources Research Center, Washington, DC. <http://www.his.com/~hphelps/>
- Phelps, HL. 2005. Identification of PCB, PAH and chlordane source areas in the Anacostia River watershed. DC Water Resources Research Center, Washington, DC. <http://www.his.com/~hphelps/>
- Phelps, HL. 2008. Active Biomonitoring for PCB, PAH and Chlordane Sources in the Anacostia Watershed. DC Water Resources Research Center, Washington, DC. <http://www.his.com/~hphelps/>
- Pinkney AE, JC Harshbarger, EB May and MJ Melancon. 2000. Tumor relevance and biomarkers of exposure in Brown Bullheads (*Ameiurus nebulosus*) from the Tidal Potomac River Watershed. *Environ Tox and Chem* 20:1196-1205.
- SRC (Syracuse Research Corporation) 2000. Interpretive Summary of Existing Data Relevant to Potential Contaminants of Concern within the Anacostia River Watershed. Report to the Anacostia Watershed Toxics Alliance. 218 pp + appendices.
- Sun, X, and U. Ghosh. 2008. The effect of activated carbon on partitioning, desorption, and biouptake of native PCBs in four freshwater sediments. *Environ. Toxicol. Chem.* 27, 2287-2295.
- Wade, T.L., D.J. Velinsky, E. Reinharz and C.E. Schlekat. 1994. Tidal river sediments in the Washington, D.C. Area. II. Distribution and sources of organic contaminants. *Estuaries* 17:304-320.
- Warner A, DL. Shepp, K. Corish and J. Galli. 1997. An Existing Source Assessment of Pollutants to the Anacostia Watershed. DCRA, Washington, DC.
- Washington Examiner 2008. Sligo Creek first step in reviving Anacostia River. 11/28/2008
- Washington Gazette. 2010. Study finds contaminants in creek, sparks debate. 4/28/2010. http://www.gazette.net/stories/04282010/bethnew214238_32567.php
- Washington Post. 2004. Anacostia River's Dirty Little Secret: Major Water Pollution Begins in Md, Not D.C. 1/29/04.
- Washington Post. 2008a. Aquatic Bloodhounds Unleashed in Anacostia Pollution Research. 8/4/08.
- Washington Post. 2008b. Anacostia Strategy Starts at Sligo. 11/29/2008.

Speciation of Some Trioganotin Compounds in Anacostia and Potomac River Sediments using NMR Spectroscopy

Basic Information

Title:	Speciation of Some Trioganotin Compounds in Anacostia and Potomac River Sediments using NMR Spectroscopy
Project Number:	2009DC102B
Start Date:	3/1/2009
End Date:	2/28/2010
Funding Source:	104B
Congressional District:	District of Columbia
Research Category:	Water Quality
Focus Category:	Sediments, Toxic Substances, Water Quality
Descriptors:	None
Principal Investigators:	Xueqing Song

Publications

There are no publications.

Speciation of Some Tributyltin Compounds in Anacostia and Potomac River Sediments using ^{119}Sn NMR Spectroscopy

Final Report



UNIVERSITY OF
THE DISTRICT
OF COLUMBIA



Dr. Xueqing Song
Associate Professor

Department of Chemistry
University of the District of Columbia

May 2010

Speciation of Some Tributyltin Compounds in Anacostia and Potomac River Sediments using ^{119}Sn NMR Spectroscopy

Abstract: The speciation of Three tributyltin compounds (TBTs), tributyltin chloride (TBTCl), Bis(tributyltin) Oxide (TBTO) and tributyltin acetate (TBTOAc) under varying pH conditions (5, 7 and 9) was studied by NMR spectroscopy in both anaerobic and aerobic Anacostia River sediments. All TBT were found to first convert to a hydrated TBT species and then further decomposition depends on the speciation time and the nature of the sediments. The ^{119}Sn NMR chemical shifts of the spiked sediments indicated that changes in the pH did not affect the speciation of the tin compounds in either aerobic or anaerobic sediments. Dealkylation to mono/dibutyltin species was observed when speciation time is 4 weeks or longer. This dealkylation is very limited as the signal around -341 ppm is very weak for all sediment samples. This would suggest that the decomposition of toxic TBTs to low toxic DBT or MBT should take more than 8 weeks in sediment. A Comparison of the strength of signal of dealkylation species and undecomposed TBT species revealed that only less than 5% was decomposed to less toxic DBT or MBT.

Key words: antifoulant NMR spectroscopy speciation Anacostia River sediments tributyltins triorganotins triphenyltins

Objectives

The overall objective of this research project was to investigate the environmental speciation of triorganotin compounds that are leached from antifouling paints into DC waterways, such as the Anacostia and Potomac Rivers, as a function of pH and to determine the transformation through interaction with the river sediments. Speciation of triorganotins is of major concern due to their species-specific toxicity. Tributyltins and triphenyltins were used in antifouling paints on ship hulls because of its strong biocidal effect, and triphenyltins can also be used as fungicides on a variety of crops. These applications are inevitably associated with triorganotin releases into the surrounding water, where it accumulates in suspended matter and in sediments. These compounds have been found to be toxic to other non-targeted marine organism, such as oysters and fish. The species that were produced as a result of these interactions were determined using NMR spectroscopy. Compared with other analytical methods, such as derivatization, pressurized liquid extraction, liquid chromatography –inductively coupled plasma mass spectroscopy, and Mossbauer spectroscopy, NMR spectroscopy offers an advantage in that it permits direct observation of the interaction between the triorganotins and the sediments.

Introduction

The Anacostia River and Potomac River are two major waterways located in the District of Columbia. Each year these rivers play host to extensive recreational activities for the residents of the metropolitan area. Two classes of pollutants that find their way into Anacostia and Potomac rivers, as well as other waterways that have high boat traffic, are tributyltins(TBTs) and triphenyltins(TPTs) since they are the toxic additives added to antifoulant marine paints. ¹ Marine paints are used to inhibit the attachment of barnacles, sea grass, hydroids and other marine organisms to the bottom of ships and other

submerged marine structures. Organotin marine paints contain as much as 20% by weight of antifoulant.¹ One mode of entry these triorganotins into the various waterways is through their release from vessels and underwater structures, such as harbors, estuaries, marinas and bays, than in open waters. The use of triorganotin compounds in the United States has been restricted by the Organotin Act which prohibits the use of organotin-based paints on vessels smaller than 25 meters.² However, vessels larger than 25 meters may still use marine paints containing organotins and a number of these larger vessels still travel these rivers, particularly, the Anacostia River where a naval shipyard is located.

Studies have shown that these organotin compounds still possess a major threat to the aquatic environment even after government regulations have restricted their use.^{3,4} In the aquatic environment, triorganotin compounds are known to have low aqueous solubility and mobility, and exhibit strong binding to sediments. These compounds are therefore easily absorbed by particular matter in water, which upon settling to the bottom, can be incorporated into the sediment.⁵ Any disturbance of the sediment will permit the direct and continuous re-introduction of the organotins back into the water column, where they can have adverse effect on non-targeted species such as crustaceans and fish.⁶

The presence of triorganotin in sediments has been regarded as long-term threat to marine and estuarine environments due to its persistence. Understanding its fate in the environment is therefore of primary importance to prevent its migration. TBT and TPT sorption were found to be reversible, indicating that contaminated sediment may release triorganotins to overlying waters following sediment disturbance.⁷ Hence the approach to understand the conditions affect the mobility of tin becomes a significant. While there have been numerous speciation studies of organotin compounds in various bodies of water around the world, there have been no similar extensive studies in DC waterways. While most investigators have focused on the determination of organotin species and their concentration in the environment, only a few studies has been initiated to study the interactions of the organotins with sediments. Thus, a study of the speciation of triorganotins in the sediments of Anacostia and Potomac rivers as a function of pH to evaluate their interaction with sediments would be essential for the understanding of the effects of triorganotins on the aquatic environment. The results from this study will alert those responsible for water quality to the long term impact of these hazardous chemicals and, therefore, allow them to plan accordingly. The results from this study will provide individuals and/or government agencies interested in water quality and planning of Anacostia and Potomac rivers with knowledge of the fate of these triorganotins once they are leached into these rivers. This information will enable those making decisions about the water quality to better assess the long term impact of these chemicals on the aquatic environment. In addition, understanding the long term environment effects of these compounds, particularly on the fish population in the Anacostia and Potomac rivers, is critical since many of the fish taken from these rivers are consumed. Consuming large amounts of these fish could have an adverse impact on the health of individuals since triorganotin are known to have mammalian toxicities.

Experiment

Triorganotin Compounds

Tributyltin chloride (TBTCI) and *bis*-tributyltin oxide (TBTO) were obtained from M & T Chemicals, Inc., Rahway, NJ, USA. Tributyltin acetate (TBTOAc) was

purchased from Gelest, Inc., Tullytown, PA. All the compounds contained the normal abundance of ^{119}Sn and were used as received without further purification to spike the sediment samples.

Collection of Sediments

Sediment samples were obtained as grab samples from the Anacostia River (Latitude: $38^{\circ} 52' 17''$ N; Longitude: $77^{\circ} 00' 18''$ W) in the DC metropolitan area. The samples were kept frozen until they were ready to be spiked. Aerobic sediments were prepared by allowing the anaerobic sediments to dry in air. The color of the sediments changed from black/greenish to black to brown.

Speciation Studies

The pH of the deionized water was adjusted to the desired values with either HCl or NaOH solutions prior to the addition of the triorganotin compounds and sediments. The anaerobic sediments were thawed in water to prevent oxidation. The following procedure was used in all experiments. Five g of aerobic or anaerobic sediment were spiked with 50 mg of the tributyltin compound. The mixture was then covered with 100 mL of deionized water. The mixture was shaken mechanically in a closed vessel in the dark for two weeks at room temperature and remained in the dark at room temperature for an additional week. The sediment samples will then be collected by gravity filtration and extracted with three portions of 15 mL of dichloromethane. The combined dichloromethane layer will be concentrated to about 5 mL using rotary evaporator and then sent for ^{119}Sn NMR analysis.

^{119}Sn NMR Analysis

All NMR measurements were made on a Varian Unity Inova 500 MHz spectrometer. Sample and instrument temperatures were controlled at 298 K. Proton-decoupled ^{13}C and ^{119}Sn spectra were acquired with WALTZ decoupling. ^{119}Sn chemical shifts were referenced to tetramethyltin externally. To identify the organotin species present, the experimental spectra were compared to spectra of known organotin compounds. Spectra of the pure compounds were recorded and used for comparison.

Results and Discussions

There are numerous analytical procedures in the literature for the determination of organotin compounds. Two recent reviews^{8,9} have indicated that the method most employed for the quantitative determination of organotin species in sediments involves some types of derivatization of the organotin species followed by species detection. For example, the determination of organotin by gas chromatography (GC) involves four steps: (1) extraction/concentration; (2) derivatization (hydridization or alkylation); (3) separation; and (4) detection.⁸ However, strong interaction between triorganotins and sediments can bias the results.⁹ Furthermore, the accuracy of butyl- and phenyltin determination is hampered by the lack of certified reference materials.⁹ It would be more advantageous to examine the original organotin species than to study their derivatized analogs, since metals and any organic species contained in the sediment can interfere with the derivatization of the organotin species.¹⁰⁻¹² Mossbauer spectroscopy has been

used in this lab to directly examine the original species in sediments.¹³⁻¹⁵ However, two unsolved problems in the speciation of organotin using Mossbauer spectroscopy make it difficult to get accurate information on the structure of the organotin species in sediments. First, due to low resolution of the Mossbauer spectrometer towards tin, to get a perfect Mossbauer spectrum, enough triorganotin compounds (0.1 g) have to be spiked with the sediment (100 g). In order to get a sediment sample close to nature, it usually will take at least 1 month to prepare a sample. The interaction between the unknown organic species contained in the sediments and the triorganotins will normally result in more than more organotin species in the sediments, it is not possible to differentiate these similar organotin species by using Mossbauer spectroscopy only.

A method that would eliminate this problem is NMR spectroscopy, since this method would allow direct examination of the organotin species in the sediments at a very low concentration. Lower concentration of tin in sediments would be environmentally closer to the natural sediment samples. The use of NMR spectroscopy for the elucidation of the molecular structure of the organotin compound is well documented in the literature.¹⁶ Specially, ^{117/119}Sn NMR provides a probe of the tin atom that is sensitive to oxidation number and the ligands around the tin atom. It has been established that the coordination number of the tin atom is related to the ¹¹⁹Sn Chemical shift. For trialkyltin complexes, four coordinate tin has ¹¹⁹Sn chemical shift ranging from about +200 ppm to -60 ppm, five coordinate tin from -90 to -190 ppm, and six coordinate tin from -200 to -400 ppm.¹⁶ For butyltin complexes, tributyltins with a coordination number 4 or 5 around tin atoms has ¹¹⁹Sn chemical shift in the rang 200ppm to 60 ppm, di butyl tin with a coordination number of 6 or even higher has ¹¹⁹Sn chemical shift in the rang -80ppm to -400 ppm, Small change of the coordinate environment to the tin atom will sensitively be reflected on the ¹¹⁹Sn NMR. Therefore, ¹¹⁹Sn NMR is an ideal analytical tool to record the complicate interaction between the triorganotin complexes and the sediments.

For preliminary studies, 8 sediment samples spiked with TBTs were prepared. The chemical shifts (δ) for the sediments spiked with TBT compounds at different pH are listed in Table 1. Typical spectra for the spiked aerobic and anaerobic sediments are shown in Fig. 1-8.

Based on the ¹¹⁹Sn NMR parameters, The preliminary data indicated that all TBTs, TBTOAc(δ : 118 ppm), TBTCl(δ : 141 ppm) and TBTO(δ : 89 ppm) were easily converted to other butyltin species in sediments. When compare with pure TBTs, no sediments samples have chemical shifts same as the pure one. This may be due to the formation of hydrated tributyltin complexes in sediment samples. Most of the hydrated TBT remained unchanged during the two weeks speciation. This is based on the observation that the major peaks around 158 ppm remain as the strongest in the ¹¹⁹Sn NMR spectra. Only very weak signals were observed which are ascribed to decomposition of the TBTs in sediments. This indicated that two weeks duration was not long enough to decompose TBTs in sediment. For this reason, samples with longer speciation time (4 weeks and 8 weekss) were also prepared for TBTOAc.

The preliminary data also indicated that changes in the pH values did not affect the decomposition of the tributyltin compounds in the same sediments. For example, TBTCl spiked with same Sediments at pH 5 (Fig 1) and 7 (Fig 2) shows very similar pattern in NMR spectra. However, different patterns were observed in NMR spectra for TBTs in

anaerobic and aerobic sediments samples. Compared with Fig 31 (aerobic sample at pH 7), anaerobic TBTOAc sample was decomposed more at same speciation time (2 weeks) Except the major signal from un-decomposed hydrated TBT, two signals at 105 (medium) and -341 (weak) ppm were also observed as shown in Fig 4. This would suggest that the organisms in the sediments are responsible for the decomposition of the TBTs. Since anaerobic and aerobic sediments have different organism composition, different pattern of decomposition are observed. This decomposition was also clearly shown in the ^1H NMR of TBTOAC samples. The typical acetate CH_3 proton with chemical shift around 2.1 ppm is missing in the ^1H NMR spectrum (Fig 5). The multiplets from 0.8-1.7-ppm are ascribed for butyl group in the sample. There are no other protons in the sample except typical protons from water around 1.6 ppm. It was also found that the tributyl hydroxide (TBTOH) in sediments will further decompose to more unknown species if enough time is given for the speciation. As shown in Fig 6 and 7, TBTOAC was converted to hydrated TBTs, then this hydrated TBT was converted to several unknown species with chemical shift from -11-109 ppm.

Chemical shifts around -340.9 ppm is an indication of dealkylation to di or monobutyltin species, though the amount of decomposition is low as the signals around -341 ppm are all very weak (Fig. 7). This would suggest that dealkylation of TBT takes a longer time than 8 weeks in sediment samples. A Comparison of the strength of signal of dealkylation species and undecomposed TBT species revealed that only less than 5% was decomposed to less toxic DBT or MBT. This is different from the conclusion we made in the studies on the speciation of triorganotin using Mossbauer Spectrometry when all the TBT were shown in Mossbauer spectra to convert to other hydrated TBT species. This would suggest that NMR spectroscopy is more sensitive spectrometer for detection of organotin species than Mossbauer spectrometer.

Reference

1. A. G. Davies and P. J. Smith, In *Comprehensive Organometallic Chemistry*, G. Wilkinson, F. G. A. Stone and E. W. Abel (eds.), Pergamon Press, Oxford, Great Britain, 1982, Vol. 2, p.613.
2. Federal Register, **53**, 39022 (1988).
3. Y. K. Chau, R. J. Maguire, M. brownFYang and S. P. Batchelor, *Water Qual. Res. J. Canada*, **32**, 453(1997).
4. W. R. Ernst, G. Julien, P. Henningar and I. Hanson, *Surveill. Rep. EPS (Environ. Can.; EPS-5-AR-99-1) i-iv*, 1 (1999).
5. S. J. Blunden, A. Hobbs and P. J. Smith, In: *Environmental Chemistry*, H. J. M. Bowen, (ed), the royal Society of Chemistry, London, 1984, p. 49.
6. S. Dodson, R. Cabridenc, M. Gilbert and P. G. Jenkins, *Tributyltin Compounds*, Environ. Health criteria 116, World health Organization, Geneva, 1990, p. 132 and reference therein.
7. P. H. Dowson, J. M. Bubb and J. N. Nester, *Appl. Organomet. Chem.*, **7**, 623(1993).
8. M. J. F. Leory, P. Quevauviller, O. F. X. Donard and M. Astruc, *Pure & Appl. Chem.*, **70**, 2051 (1998).
9. M. Abalos, J. M. Bayona, R. Compano, M. Granados, C. Leal and M. D. Prat, *J. Chromatogr. A.*, **788**, 1 (1997).
10. O. F. X. Donard, L. Randall, S. Rapsonmanikis and J. H. Weber, *Intern. J. Environ. Anal. Chem.*, **27**, 55 (1986).
11. P. J. Craig and S. Rapsonmanikis, *Inorg. Chim. Acta*, **80**, 119 (1983).
12. P. Quevauviller, F. M. Martin, C. Belin and O. F. X. DOnard, *Appl. Organomet. Chem.*, **7**, 149 (1993).
13. G. Eng, X. Song and L. May, *Hyperfine Inter.*, **170**, 117 (2006).
14. X. Song, A. Zapata, L. May and G. Eng, *Main Group Chem.*, **4**, 39 (2005).
15. G. Eng, D. Desta, E. Biba, X. Song and L. May, *Appl. Organometal. Chem.*, **16**, 67 (2002).
16. G. Davies and P. J. Smith, in *Comprehensive Organometallic Chemistry*, G. Wilkinson, F. G. A Stone and E. W. Abel (eds.), Vol. 2, Pergamon Press, New York, 1982, p. 519.

Table 1. ^{119}Sn NMR chemical shifts for TBTs spiked with sediment samples from the Anacostia River

TBTs	pH	Speciation Duration	Sample type	Chemical shifts	
				200 to 60 ppm	0 to -400 ppm
TBTCl	7	2 weeks	Anaerobic	157.0 (medium) 109.9 (medium) 107.3 (strong) 95.1 (weak) 76.9 (medium) 76.6 (medium) 64.3 (medium)	-11.4 (weak)
TBTCl	5	2 weeks	Anaerobic	158.0 (strong) 107.8 (medium) 110.7 (weak) 77.1 (weak)	
TBTCl	Pure			85	
TBTO	7	2 weeks	Anaerobic	156.0 (medium) 105.2 (strong)	-340.9 (weak)
TBTO	Pure			141	
TBTOAc	7	2 weeks	Aerobic	157.1 (strong)	
TBTOAc	7	2 weeks	Anaerobic	156.0 (medium) 105.2 (strong)	-340.9 (weak)
TBTOAc	7	4 weeks	Anaerobic	157.5 (strong)	-340.9 (weak)
TBTOAc	7	8 weeks	Anaerobic	156.1 (medium) 109.6 (medium) 106.8 (strong) 76.3 (medium) 63.8 (medium)	-11.4 (weak) -340.9 (weak)
TBTOAc	Pure			118	

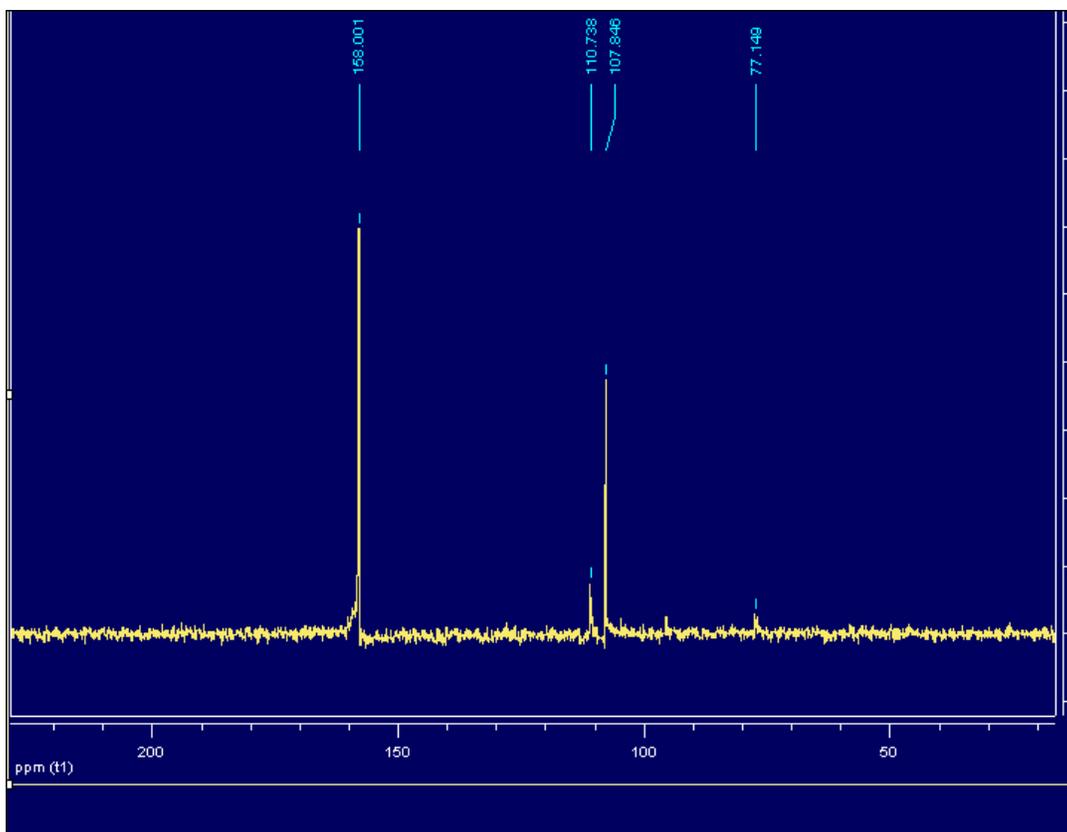


Fig. 1. ^{119}Sn NMR spectra of tributyltin chloride (TBTCI) in spiked anaerobic sediments from Anacostia River at pH 5. (Speciation time 2 weeks)

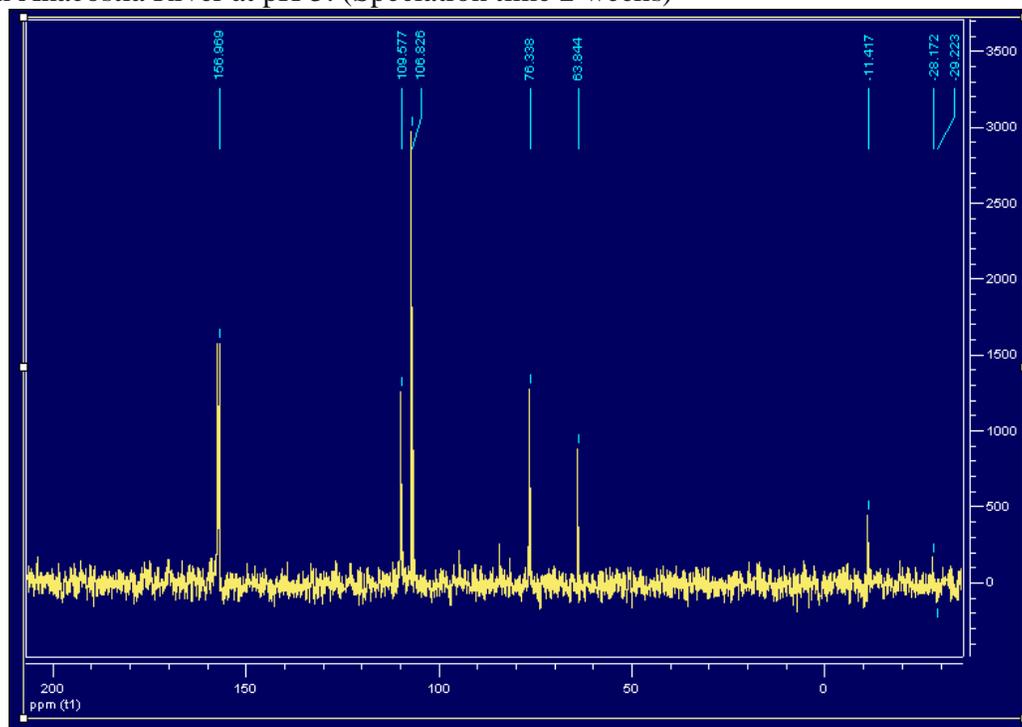


Fig. 2. ^{119}Sn NMR spectra of tributyltin chloride (TBTCI) in spiked anaerobic sediments from Anacostia River at pH 7. (Speciation time 2 weeks)

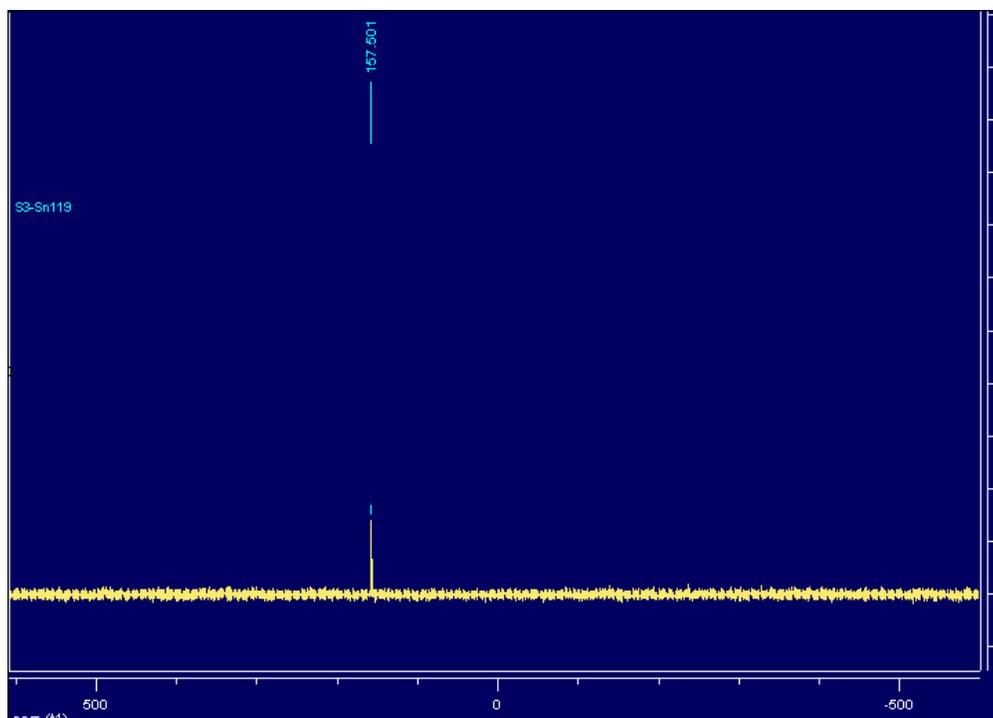


Fig. 3. ^{119}Sn NMR spectra of tributyltin acetate (TBTOAC) in spiked aerobic sediments from Anacostia River at pH 7 (speciation time 2 weeks).

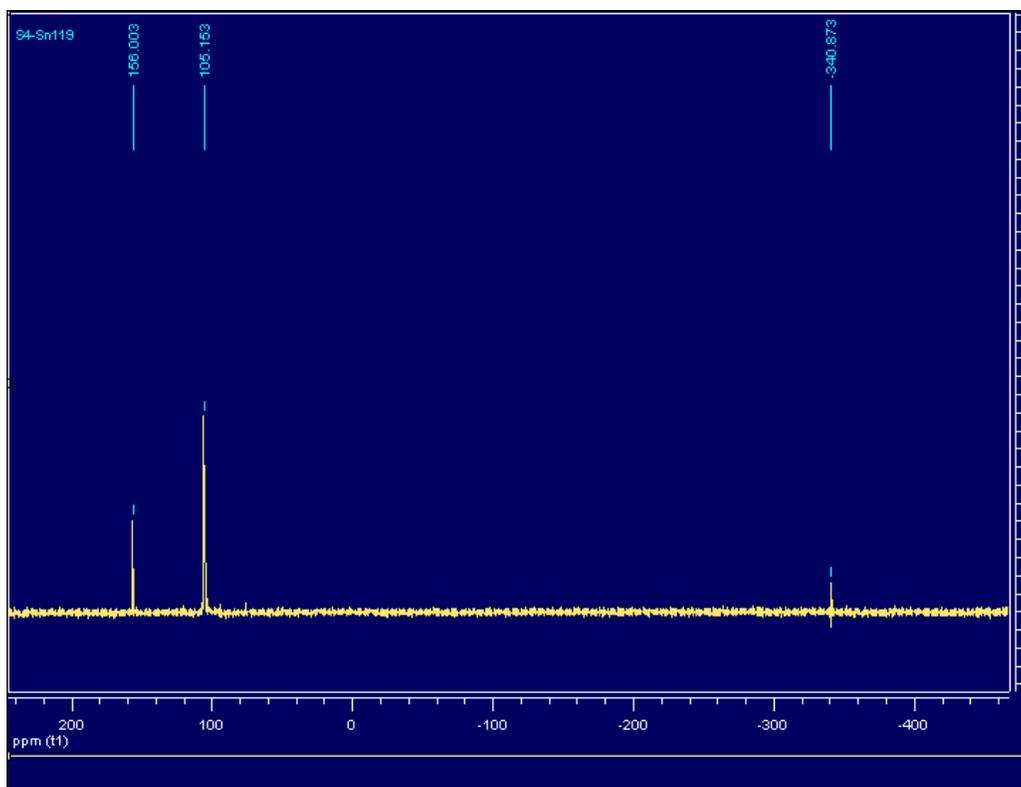


Fig. 4. ^{119}Sn NMR spectra of tributyltin acetate (TBTOAc) in spiked anaerobic sediments from Anacostia River at pH 7. (Speciation time 2 weeks)

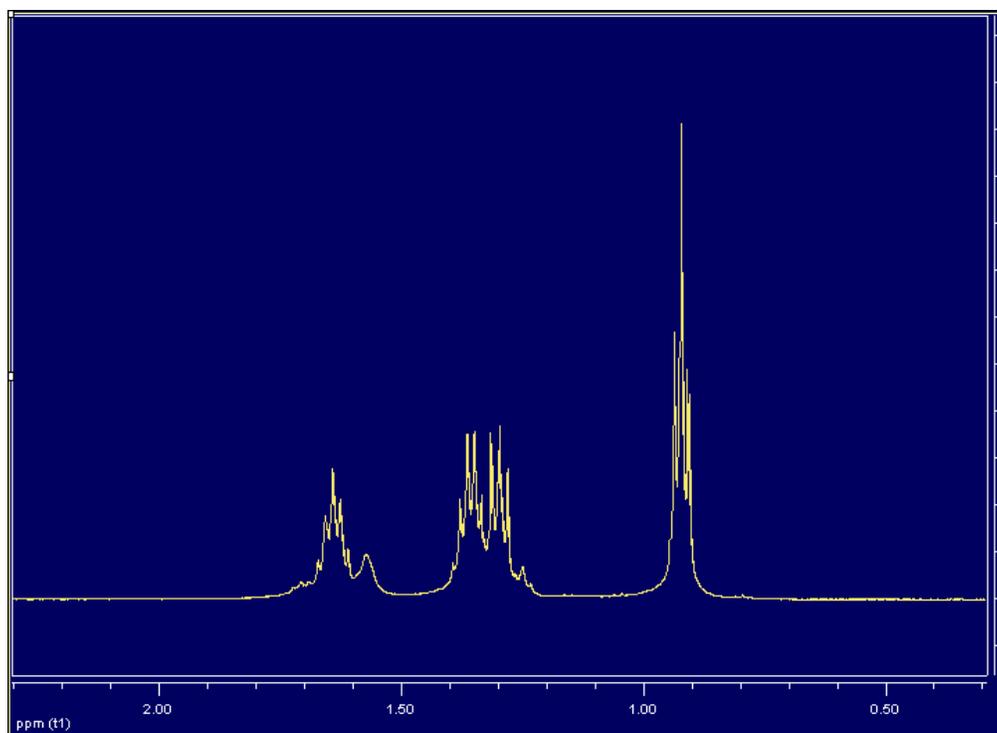


Fig. 5. ^1H NMR spectra of tributyltin acetate (TBTOAc) in spiked anaerobic sediments from Anacostia River at pH 7.

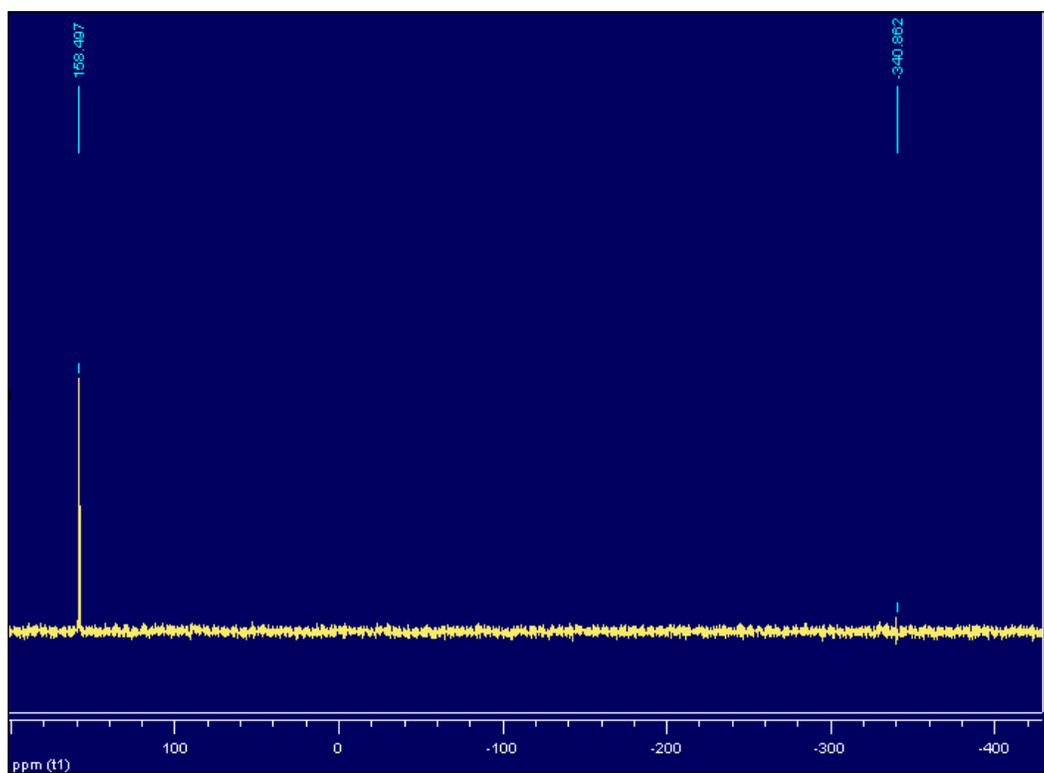


Fig. 6. ^{119}Sn NMR spectra of tributyltin acetate (TBTOAc) in spiked anaerobic sediments from Anacostia River at pH 7. (Speciation time: 4 weeks)

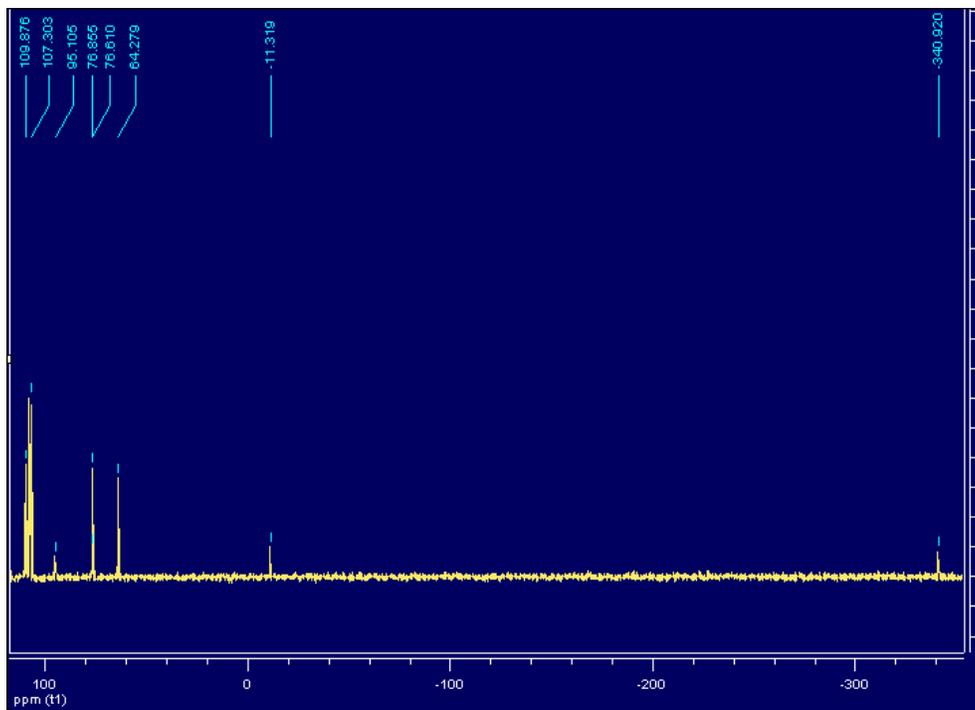


Fig. 7. ^{119}Sn NMR spectra of tributyltin actate(TBTOAc) in spiked anaerobic sediments from Anacostia River at pH 7. (Speciation time: 8 weeks)

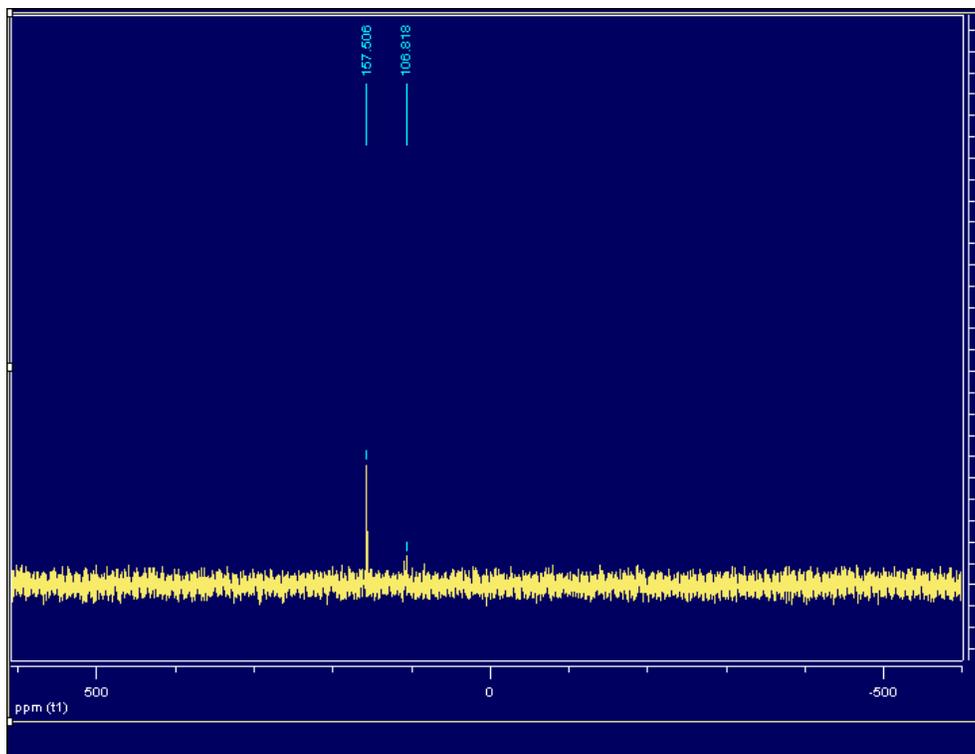


Fig. 8. ^{119}Sn NMR spectra of bistrbutyltin oxide (TBTO) in spiked anaerobic sediments from Anacostia River at pH 7. (Speciation time 2 weeks)

Development of a Fast Optimization Technique Using Interactive Spatial Join for GIS Application in Water Resources

Basic Information

Title:	Development of a Fast Optimization Technique Using Interactive Spatial Join for GIS Application in Water Resources
Project Number:	2009DC103B
Start Date:	3/1/2009
End Date:	2/28/2010
Funding Source:	104B
Congressional District:	District of Columbia
Research Category:	Engineering
Focus Category:	Water Quality, Management and Planning, Methods
Descriptors:	None
Principal Investigators:	Seon Ho Kim, Pradeep K. Behera, Byunggu Yu

Publications

There are no publications.

**Development of a Fast Optimization Technique Using
Interactive Spatial Join for GIS
Applications in Water Resources
(Phase I)**

Progress Report



UNIVERSITY OF
THE DISTRICT
OF COLUMBIA



Seon Ho Kim, Ph.D.
Byunggu Yu, Ph.D.
Pradeep K. Behera, Ph.D., P.E., D. WRE

School of Engineering and Applied Sciences
University of the District of Columbia

June 2010

Abstract

Many Geographic Information Systems (GIS) handle large geospatial datasets stored in raster representation. Spatial joins over raster data are important queries in GIS for data analysis and decision support. However, evaluating spatial joins can be very time intensive due to the size of these datasets. In this study, we propose a GIS tool support for interactive and real time spatial join, especially for the water resource research. We propose a new interactive framework that allows users to get approximate answers in near instantaneous time, thus allowing for truly interactive data exploration. Our method utilizes two proposed statistical approaches: probabilistic join and sampling based join. Our probabilistic join method provides speedup of two orders of magnitude with no correctness guarantee, while our sampling based method provides an order of magnitude improvement over the full quad-tree join and also provides running confidence intervals. We propose a framework that combines the two approaches to allow end users to trade-off speed versus bounded accuracy. The two approaches are evaluated empirically with real and synthetic datasets.

1 Introduction

Geographic Information Systems (GIS) are used for storage and retrieval of large spatial datasets. Each dataset is usually called a layer. Example layers may be roads, rivers, land elevation, etc. Layers are related if they have the same geographic coordinates. Spatial joins between two or more data sets are one of the most common GIS queries for data analysis. An example might be finding all roads within 100 feet of rivers located at 1000 feet altitude or less. GIS users often want to visualize query results and being able to do so in an interactive fashion would greatly increase the utility of the GIS. Unfortunately, the large dataset size makes interactive spatial joins difficult.

Geospatial data is usually stored in one of two alternative data formats: raster (grid cells) and vector (point, line and polygon). In this paper we focus on raster data, where little research has been done on optimization or approximation techniques for spatial joins. Currently, performing spatial joins on raster data requires layers to be compared on a cell-by-cell basis. This spatial join process, referred to as *map overlay*, requires intensive computation time. To enable interactive queries, more efficient methods for dealing with raster data are needed.

GIS systems are often used to visualize results for the end user to assist in decision making processes. In many applications, obtaining an approximate join result in a reasonably short time is far more important than evaluating an exact join over a long time period. Fast response times are especially important for user-driven data exploration used in GIS. We believe GIS users should be given the chance to see which are the “interesting” dataset join pairs without having to wait to compute the actual full joins. In this paper we propose an interactive query processing framework for spatial join that enables GIS users to obtain an approximate “big picture” visualization of an answer in two orders of magnitude faster time than the time required for obtaining the exact answer.

Our general interactive framework works as follows. Users specify queries and get near instantaneous visualizations of the answer using our proposed probabilistic join method. These

result visualizations are approximations with no guaranteed bound of correctness. For queries that had interesting results users can either use our proposed sampling algorithm to get a confidence bounded answer estimate, or, compute the full join. By allowing the user to get near instant approximate answers they are able to explore far greater numbers and sizes of datasets than previously possible. This increase ability does come at the cost of possibly making a mistake and hence may not be appropriate for systems used in critical life decisions.

Our approach is based on two techniques:

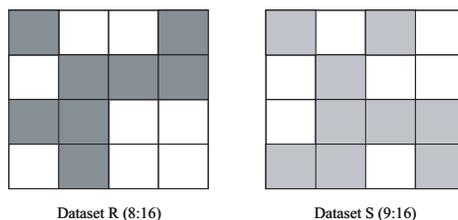


Figure 1: Raster cells of datasets R and S

- Probabilistic joins: The main idea is to calculate the join probability and the expected number of the joined cells of two raster datasets that have the same geographic coordinates. Data density (ratio of non-zero data cells to total data cells) in each node of a quad-tree of two datasets is used to calculate the join probability. Figure 1 shows an example of two datasets; R and S represented by a 4×4 raster grid. In this example, R has 8 non-zero data cells (density: $8/16$) while S has 9 non-zero data cells (density: $9/16$). Then R and S must intersect regardless of their shape and location. The ratio of non-zero data cells to the total number of data cells of each data set can be used in the calculation of the join probability and the expected number of joined cells.
- Sampling joins: Using quad-trees, overlapping blocks (sub-regions) are used to filter candidate pairs in order to speed up the joining process. Our sampling join approach is based on stratified random sampling from quad-trees and performing joins on the incremental samples to estimate the final answers of spatial joins with bounded confidence intervals.

Our proposed interactive framework combines the two proposed statistical approaches in order to speed up the process of obtaining estimations of the final join result in a reasonable time compared to the total time needed to perform a full join. Augmented quad-trees with non-zero data cells are used in the framework. We provide experimental results for both synthetic and real GIS datasets that demonstrate the efficacy of our approach comparing to full quad-tree joins. The speedup relative to a full quad-tree join increases as dataset size increases.

2 Related Work

One common raster data spatial join technique is map overlay [11]. Raster overlay is straightforward when the input rasters have the same cell boundaries. The resulting raster can be obtained cell by cell from the originals using the relevant operations on the cell values. However, little research work has been done on map overlay optimization techniques.

Since GIS data can reach gigabytes and possibly terabytes in size, full layer overlays could take hours and even days to complete. This necessitates a need for approximation techniques. Most of the work on relational database join approximations can not be directly applied to spatial databases. In [14, 1] the authors presented an approximation technique of vector-based spatial joins. First they converted vector data to raster format and filtered the possible joined pairs using the Four Color Raster Signature in [14] and the Three Color Raster Signature in [1]. They combined progressive and conservative approximations [4] in a single approximation to speed up the filtering step in identifying intersecting polygons. Their proposed techniques motivated us to obtain the join probability of two raster datasets.

The quad-tree is a very popular hierarchical data structure for the representation of binary images and maps and it is commonly used in spatial databases [9, 13], i.e., indexing for query processing, and optimizing decomposition. Our work assumes datasets are indexed by quad-trees. Quad-tree based sampling has been proposed in [8, 13]. In [13], the authors presented the analysis of four different sampling methods proposed by [8]. They applied sampling algorithms to specific quad-tree implementations to obtain approximate aggregate query results. They proposed two models in order to analyze sampling costs while our sampling

approach provides a faster approximation of the join result with a bounded confidence interval.

The idea of incremental sampling technique using R-trees to provide interactive spatial join processing was proposed in [2]. The authors proposed two R-tree based sampling methods that were used to incrementally refine the estimated join result while providing a bounded confidence interval. Their approach was applied for vector-based data rather than raster data. The proposed sampling method in this paper follows the same framework but using quad-trees instead of R-trees and with a more sophisticated sampling method.

Probabilistic query evaluation was studied for uncertain continuously changing data in relational databases [5]. In [6], the authors proposed probabilistic join over uncertain data. They provided techniques to answer queries that return results with probabilities exceeding a given threshold.

To the best of our knowledge, our work is the first attempt to apply probabilistic approaches to estimate raster-based spatial joins.

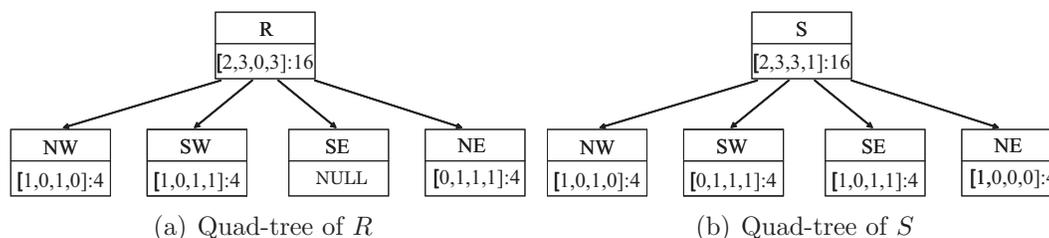


Figure 2: Examples of augmented quad-trees of datasets R and S

3 A Framework for Spatial Joins Over Raster Data

3.1 Augmented Quad-tree

Statistical methods are concerned with the estimations of parameters of the population in GIS. These approaches use information associated with the population, samples drawn from the population and distribution of the samples.

Augmented quad-tree data structure is used for *PJ* and *ISSJ*. Specifically, we augment nodes to include the total number of non-zero data cells of the subtree below. Our proposed statistical approaches use these augmented quad-trees for obtaining information associated with the population. Figure 2 (a) and (b) show two augmented quad-trees of the raster dataset examples in Figure 1. The nodes of the quad-trees are displayed in counter clock-wise order starting from the north-west. In our framework, all datasets are indexed by augmented quad-trees.

3.2 Probabilistic Joins vs. Random Sampling

In *PJ*, the augmented value (number of non-zero data cells) of each node of given two datasets is used to calculate the join probability and the expected number of joined data cells for each pair of subregions in the two joined datasets. *PJ* accesses nodes from the top to the bottom hence *PJ* is referred to as a top-down approach. *PJ* does not need to access all levels of quad-tree to calculate an estimate. It is mostly enough to access only a small number of top levels. Thus, it can greatly reduce time-consuming disk I/O operations in practice. The number of levels to be accessed is a system parameter. The greater number of levels is accessed, the more accurate the estimation can be. However, this would result in more I/Os. In the experiments, we set the number of levels to 4 resulting in only 64 nodes needed in memory, hence it is practical to store only required top level nodes of quad-trees for all joined data sets in memory. Although *PJ* provides no accuracy guarantee, our experimental results of synthetic and real data sets show the error bound is reasonably tight, e.g., a 9% error for 4th level join (Section 5).

In *ISSJ*, stratified random sampling is used to estimate the final answer of spatial joins. An accuracy guarantee is provided in the form of error bound confidence intervals. In contrast to *PJ*, *ISSJ* is performed on sampled leaf level data cells. Although far less number of I/Os are required in *ISSJ* comparing to a full quad-tree join, obtaining a reasonable confidence interval requires a significantly greater number of I/Os comparing to *PJ*.

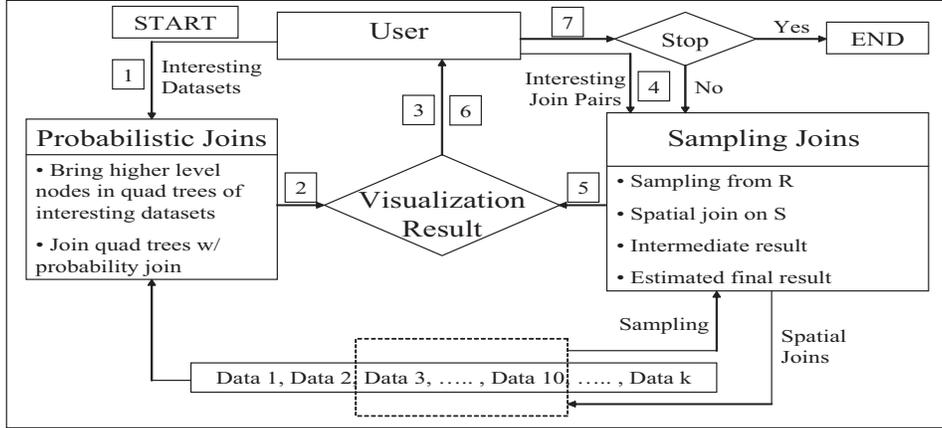


Figure 3: A framework for raster joins

3.3 Framework Overview

We proposed a query processing framework that combines the proposed statistical approaches. The framework consists of three main processes: probabilistic joins, result visualizations and sampling joins. The main idea is to use PJ and a visualization technique to allow users to discover “interesting” data set pairs and areas for further data exploration. Once the user identifies interesting data sets, he/she can have the system perform $ISSJ$ in order to produce tighter running estimates of join results, or the user can have the system use the full quad-tree join to obtain the exact answer.

Figure 3 shows the overview of the framework, where two relations R and S are joined. 1) Probabilistic Joins (PJ): Given the user’s interesting data sets, all higher level nodes (from level 0 to level 3 in our experiments) of the two data sets’ quad-trees are loaded in memory. Then the join probability of each pair of the corresponding nodes is obtained from a look-up table. Since join probability is defined on continuous space, the system can use a lookup table for discrete values of join probability. 2) Visualization and user interface: Based on a visualized result of probabilistic joins, the user can identify “interesting join pairs”. 3) Incremental Stratified Sampling Joins ($ISSJ$): $ISSJ$ starts incremental sampling process with the interesting pairs. Samples (non-zero cells) are randomly chosen from the outer relation R using stratified random sampling. Spatial joining on the corresponding cells of the

inner data set is performed. The number of joined cells in each step is used to calculate a running estimate and a confidence interval for the final result. Finally, the calculated running estimate and confidence interval are combined with the intermediate result into a query result through visualization process. Then the query result is reported to the user. The user can stop the query process if the given confidence interval is sufficient or if the user sees satisfying trends from the visualized actual join locations (intermediate result), otherwise, each step of the process is repeated in an incremental manner to calculate new estimates until a desired confidence interval is achieved. Thus, the time to get join estimates needs to be compared to the time required for the full quad-tree join.

4 Statistical Join Approaches

4.1 Probabilistic Join

Given a set X and two randomly chosen subsets A and B of X , what is the probability that $A \cap B \neq \emptyset$? Let us denote this probability by p . There is an easy answer in the finite case. Let $|X| = n$, $|A| = a$, $|B| = b$. Then $p = 1 - \frac{\binom{n-a}{b}}{\binom{n}{b}}$, since this is the probability that a randomly chosen b -element subset of X will not avoid a given a -element subset of X . But there is no reasonable answer in the infinite case, since we run into the well-known problems with (i) what is meant by “random” (the answer depends on how the experiment is conducted), (ii) measurability (how to determine the size of a set).

We therefore restrict our attention to subsets of special kind, and use the obtained answers as approximations to the (unsolvable) general case.

Theorem 4.1. (*Join Probability for intervals*)

Let $X = [0, 1]$, and let A, B be randomly chosen intervals in X of length a, b , respectively. Then, the probability p that $A \cap B \neq \emptyset$ depends only on a, b , and can be calculated by:

$$p(a, b)_1 = \frac{1}{1-b} \int_0^{1-b} \frac{\min\{x+b, 1-a\} - \max\{0, x-a\}}{1-a} dx$$

Proof. Let A and B be $[a_l, a_h]$ and $[b_l, b_h]$, respectively, such that $\overline{a_l a_h} = |A| = a$ and $\overline{b_l b_h} = |B| = b$. If A and B are picked at random, then $a_l \in [0, 1-a]$ and $b_l \in [0, 1-b]$ (see Figure 4). Assuming that x is a random variable for the value of b_l , we have $x \in [0, 1-b]$. Then

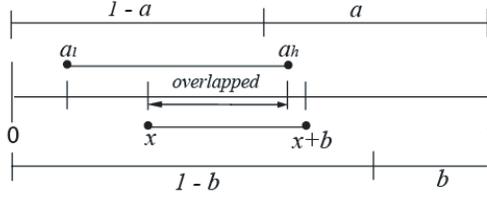


Figure 4: Join of two intervals

$p(a, b)$, the probability that $A \cap B \neq \emptyset$ (A intersects B), is as follows:

$$\begin{aligned}
 p(a, b)_1 &\equiv P((a_h \geq b_l) \wedge (a_l \leq b_h)) \\
 &\equiv P((a_h \geq x) \wedge (a_l \leq \min\{x + b, 1 - a\})) \\
 &\equiv P((a_l \geq x - a) \wedge (a_l \leq \min\{x + b, 1 - a\})) \\
 &\equiv P(\max\{x - a, 0\} \leq a_l \leq \min\{x + b, 1 - a\})
 \end{aligned}$$

In order to have $p(a, b) \neq 0$, we need to pick a_l between $\max\{x - a, 0\}$ and $\min\{x + b, 1 - a\}$ from the continuous space in which the range of x (b_l) is $[0, 1 - b]$ and the range of a_l is $[0, 1 - a]$. Then we have the following equation.

$$p(a, b)_1 = \frac{1}{(1-a)(1-b)} \int_0^{1-b} \min\{x + b, 1 - a\} - \max\{0, x - a\} dx$$

□

Theorem 4.1 can now be generalized to any number of dimensions. The 2-dimensional case is as follows:

Let $X = [0, 1]^2$, and let A, B be rectangles in X of area a, b , respectively. If the sides of A are of length $a_1, a_2 = a/a_1$ and the sides of B are of length $b_1, b_2 = b/b_1$, then we can use the 1-dimensional case to deduce that $P(A \cap B \neq \emptyset) = p(a_1, b_1) \cdot p(a_2, b_2)$. However, we do not know a_1 and b_1 . All we know is that $a_1 \in [a, 1]$ (since the length of each side of A has to be at least a) and $b_1 \in [b, 1]$. We therefore conclude that:

$$p(a, b)_2 = \frac{1}{(1-a)(1-b)} \int_a^1 \int_b^1 p(a_1, b_1) \cdot p\left(\frac{a}{a_1}, \frac{b}{b_1}\right) da_1 db_1$$

It is now easy to see the general formula for two n -d prisms A, B in $X = [0, 1]^n$ of volumes a, b , respectively. Let the lengths of sides of A and B be $(a_1, \dots, a_n), (b_1, \dots, b_n)$, respectively. Then

$$\begin{aligned}
p(a, b)_n &= \frac{1}{(1-a)(1-b)} \int_a^1 \int_{a_1}^1 \cdots \int_{a_1 \cdots a_{n-1}}^1 \int_b^1 \int_{b_1}^1 \cdots \\
&\quad \cdots \int_{b_1 \cdots b_{n-1}}^1 u db_{n-1} \cdots db_1 da_{n-1} \cdots da_1, \\
\text{where } u &= p(a_1, b_1) \cdots p(a_{n-1}, b_{n-1}) p\left(\frac{a}{a_1 \cdots a_{n-1}}, \frac{b}{b_1 \cdots b_{n-1}}\right)
\end{aligned}$$

The expected overlapped length (area, volume) of A and B can be calculated using the conditional probability, since it is assumed that the two datasets are chosen independently:

$$P(A \cap B) = P(A) \cdot P(B)$$

The formulae for the join probability and the expected join numbers can be extended to more than two datasets joins.

4.2 Incremental Stratified Sampling Join

Sampling methods are used to estimate the final result from a subset (samples) of the data and to provide a bounded confidence interval. Query estimations and confidence intervals are statistically meaningful only if samples are retrieved at random. A weighted random sampling method, *Acceptance/Rejection* [8], is used in the *ISSJ* algorithm. We study stratified random sampling without replacement for raster data spatial joins. Each sampling is conducted in an incremental manner and the performance is evaluated with varying data sets and buffer sizes.

4.2.1 Stratified Random Sampling

Stratified random sampling is chosen because its property matches the property of quad-trees that provides systematic decomposition of a space with no overlaps between subregions. In stratified random sampling, the given region (population of all data cells) is divided into a number of non-overlapping subregions called strata. Then each stratum contains a set of raster data cells. Stratified random sampling can result in smaller error bounds on an estimation and can reduce the sampling cost [10].

In our algorithm, stratification is based on non-overlapping geometric forms such as rectangles (nodes at each level). We define the internal nodes of the quad-tree for a given level

as strata, i.e., the second level nodes of quad-tree are used as strata in our experiments. We assume that the strata is pre-defined in our experiments. Algorithm 1 describes the *ISSJ* algorithm.

Samples (non-zero cells) are then randomly chosen from each stratum by conducting simple random sampling. The sample size of each stratum $n_i, i=1, ..k$, is calculated for every sampling step, and it is proportional to the total number of non-zero cells within that stratum. Then the sampling size for a sampling step is $n_s = \sum_{i=1}^k n_i$. If the value of the chosen data cell is 1, searching the corresponding joined cell of the inner data set is performed in the quad-tree of the inner data set (line 15 of Algorithm 1). If the value of the corresponding cell is 1, then two data cell join. For each stratum, we obtain the number of joined cells, and this number is used to calculate the estimate and confidence interval for the corresponding stratum. The sum of the joined cells of each stratum is the current intermediate result, and the estimates and confidence intervals of all strata are combined for an estimate and a confidence interval of the final answer. The user can stop the query process if the given confidence interval is sufficient, otherwise the process repeats.

4.2.2 Estimates for Stratified Random Sampling

To provide bounds on the accuracy of our result, we incrementally calculate the current estimate with a confidence interval. The estimates and confidence intervals of *ISSJ* are based on population proportion and *the Central Limit Theorem* (CLT) [7, 10]. We use the binomial probability distribution [10] for statistics of *ISSJ*. In *ISSJ*, the population is the non-zero cells of the outer relation R and \hat{p} is the fraction of the elements in the sample that possess the characteristic of interest (“join” in our algorithm). Hence \hat{p} is the fraction of cells in the sample that joins with the corresponding cell of the inner relation S . Confidence intervals depend on the size of samples and the distribution of the sample space (i.e., *Student t-distribution*).

Let N be the size of population (total number of non-zero cells of the outer datasets) and n_s be the sample size for a sampling step. If N_i is the number of non-zero cells in stratum i ,

Algorithm 1 *ISSJ*(R, S, ST)

```
1:  $ST = \{ST_1, \dots, ST_k\}$ ;  $ST$  is a set of strata
2:  $I_1, \dots, I_k \leftarrow 0$ ;  $C_I \leftarrow 0$  {the current joined cells for stratum  $i$ ; confidence interval}
3:  $n_s \leftarrow 0$ ;  $n_{init} \leftarrow 30$  {the sample size for a sampling step; the initial incremental sample
   size for a sampling}
4:  $n_1, \dots, n_k \leftarrow 0$ ;  $s_1, \dots, s_k \leftarrow 0$  {the sample size for stratum  $i$ ; the incremental sample size
   for stratum  $i$ }
5: repeat
6:   compute  $s_1, s_2, \dots, s_k$  for  $ST_1, ST_2, \dots, ST_k$  using  $n_{init}$ 
7:    $S \leftarrow \sum_{i=0}^k s_i$ ;  $n_s \leftarrow n_s + S$ 
8:   for  $i = 1$  to  $k$  do
9:      $n_i \leftarrow n_i + s_i$ 
10:    for  $j = 1$  to  $s_i$  do
11:       $L \leftarrow$  choose a leaf from  $ST_i$  at random
12:       $c_r \leftarrow$  choose a non-zero cell from  $L$  at random
13:      if cell  $c_r$ 's value is 1 then
14:         $P_r \leftarrow$  the center point of the chosen cell  $c_r$ 
15:         $c_s \leftarrow$  findJoinedCell( $S, P_r$ )
16:        if cell  $c_s$ 's value is 1 then
17:           $I_i \leftarrow$  add 1
18:        end if
19:      end if
20:      remove  $c_r$  from  $L$ 
21:    end for
22:    remove  $L$  from  $ST_i$  if  $L$  is empty
23:  end for
24:   $I \leftarrow \sum_{i=0}^k I_i$ 
25:   $C_I \leftarrow$  Compute a confidence interval w/all  $I_i$  and  $n_i$ 
26:   $EV \leftarrow$  Compute an estimate w/all  $I_i$  and  $n_i$ 
27:  report  $EV, C_I$ , and  $I$ 
28: until  $C_I$  is sufficient to the user or all  $ST_i$  are empty
```

and n_i is the sample size for stratum i , then $N = \sum_{i=1}^k N_i$, and $n_s = \sum_{i=1}^k n_i$, where k is the number of strata. Let I_i be the total number of cells that join the corresponding cells of S in stratum i . The following equations are used for a sampling step for *ISSJ*:

Estimator of the population proportion, where $\hat{p}_i = \frac{I_i}{n_i}$:

$$\hat{p} = \frac{1}{N}(N_1\hat{p}_1 + N_2\hat{p}_2 + \dots + N_k\hat{p}_k) = \frac{1}{N} \sum_{i=1}^k N_i\hat{p}_i \quad (1)$$

Estimate variance of \hat{p} :

$$\hat{V}(\hat{p}) = \frac{1}{N^2} \sum_{i=1}^k N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1} \right) \quad (2)$$

Confidence interval:

$$E = t_c \sqrt{\hat{V}(\hat{p})} \quad (3)$$

where t_c is the critical value for confidence level c taken from a Student t-distribution. Equations (1), (2) and (3) are valid for the incremental stratified sampling process. The proof of incremental equations can be found in our technical report [3].

5 Experiments

In this section, we present experimental results of *PJ* and *ISSJ* with both synthetic and real GIS datasets. The performances of two are compared with each other as well as with the full quad-tree join.

5.1 Data Sets and Experimental Methodology

	synthetic datasets								real datasets			
	uni1	uni2	uni3	uni4	exp1	exp2	exp3	exp4	AZ	CO	OR	WY
# total cells	65536	65536	262144	262144	65536	65536	262144	262144	65536	65536	65536	65536
# N.E. cells	17325	28365	39120	48298	14256	24736	36290	45231	6 datasets Mineral Resources from USGS			
density	0.26	0.43	0.15	0.18	0.22	0.38	0.14	0.17				
description	uniformly distributed data				exponentially distributed data							

Table 1: Synthetic and real datasets

In our experiments, we consider both synthetic and real data sets shown in Table 1. We generated four sets of uniformly distributed raster data and four sets of exponentially distributed (a mean of 0.3 and a standard deviation of 0.3) raster data. Our real data sets are from the 2001 and 2005 U.S. Geological Survey [12]: six datasets are chosen from Arizona, Colorado, Oregon and Wyoming in the US. These datasets are about minerals, stream sediments, water sediments, rocks, pluto sediments and unconsolidated sediments. Each dataset was converted into raster format. In Table 1, we present the total number of data cells (pixels), the total number of non-zero data cells and the data density for the synthetic and real datasets.

It is necessary that both the outer and inner datasets are indexed by augmented quad-trees and they have the same number of data cells as well as the same size of cells. Our experiments were conducted using the following parameters: Augmented quad-trees are implemented for *PJ* and *ISSJ* while nonaugmented quad-trees are used for the full quad-tree join. The page size of the quad-tree was set to 4Kbytes, resulting in 100 nodes and 64 nodes for the non-augmented tree and augmented tree, respectively. Assuming an LRU replacement policy, we vary the buffer size: 5%, 10% and 20% of the size of one of the two relations. For all presented results, the estimates and the corresponding confidence intervals are shown with a 95% confidence level.

5.2 Experimental Evaluation

First we present the accuracy of join probability using the 1-dimensional formula (p_1) and 2-dimensional formula (p_2) discussed in Section 4. The total number of joins obtained by the 1-d and 2-d join probability were compared with the total number of actual joins. For discrete values of join probability, we created two lookup tables (20×20). Table 2 illustrates a portion of 2-d join probabilities from the lookup table used in the experiments. We randomly selected two corresponding nodes from the quad-trees of two real datasets. We checked the occupancy rates (non-zero data cells/total data cells) in the two chosen nodes and obtained the 1-d and 2-d join probabilities from the lookup tables. Then the expected numbers of joins were calculated. We repeated this process for varying size of sample pairs: 5%, 10%, 20% and 50%

of the total quad-tree nodes. We ran the experiment 10000 times with each of the sample sizes and presented the average. In Table 3 we show the results of a join, unconsolidated sediments \bowtie minerals in CO. The table entries are actual error values, thus, for example, an error of 0.1060 is a 10.60% error. Clearly, the 2-d join probability provides better approximation of the actual join.

To evaluate the quality of the “big picture” visualization obtained by PJ , we calculated the expected number of joins using the 4th level tree nodes. When using the 4th levels of two quad-trees, only 64 subregions are joined. As a result, users can obtain the approximate result visualization in near instantaneous time with a truly interactive manner. We present results showing the difference between the PJ method and the full quad-tree join method (see Table 4). For the real datasets we compared PJ and $ISSJ$ for all 15 possible pairwise joins of the 6 datasets. We grouped the synthetic datasets into two: group 1 (uni1, uni2, exp1, exp2) and group 2 (uni3, uni4, exp3, exp4). We computed all possible 6 pairwise joins of each of the two groups. In Table 4, we present the average differences in the join density. The minimum and maximum of maximum difference, and the average maximum difference are also presented. Finally we calculated the average error in the expected number of joins of all the pairwise joins. As can be seen, PJ is reasonably accurate in all the cases of both real and synthetic datasets. With real data sets, PJ resulted in less accuracy due to the scattered clusters found in the datasets. As shown later in Figure 6, for the data we explored, these modest inaccuracies have little effect on the overall visual join-result appearance.

P	0.2	0.4	0.6	0.8	1.0
0.2	0.7683	0.9277	0.9903	1.0	1.0
0.4	0.9277	0.9937	1.0	1.0	1.0
0.6	0.9903	1.0	1.0	1.0	1.0
0.8	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0

Table 2: Example of a 2-d lookup table

Next, we present the performance of $ISSJ$ compared to the augmented full quad-tree join. Figure 5 shows the result using the real datasets (minerals \bowtie unconsolidated sediments from

sample size	actual join	2-d (error)	1-d (error)
5 %	54	48 (0.1060)	39 (0.2778)
10 %	109	99 (0.0917)	78 (0.2844)
20 %	218	197 (0.0963)	155 (0.2889)
50 %	545	494 (0.0936)	389 (0.2862)

Table 3: Join probability

join datasets	real datasets				synthetic datasets	
	AZ	CO	OR	WY	group1	group2
average diff.	0.0060	0.0087	0.0049	0.0058	0.0032	0.0024
minimum of max. diff.	0.0047	0.0038	0.0045	0.0014	0.0018	0.0015
maximum of max. diff.	0.1208	0.0973	0.0849	0.1143	0.0410	0.0312
average max. diff.	0.0329	0.0237	0.0214	0.0199	0.0201	0.0182
average error of estimates	0.1105	0.0729	0.0629	0.0904	0.0324	0.0229

Table 4: Join density differences of probabilistic joins from actual joins (4^{th} level)

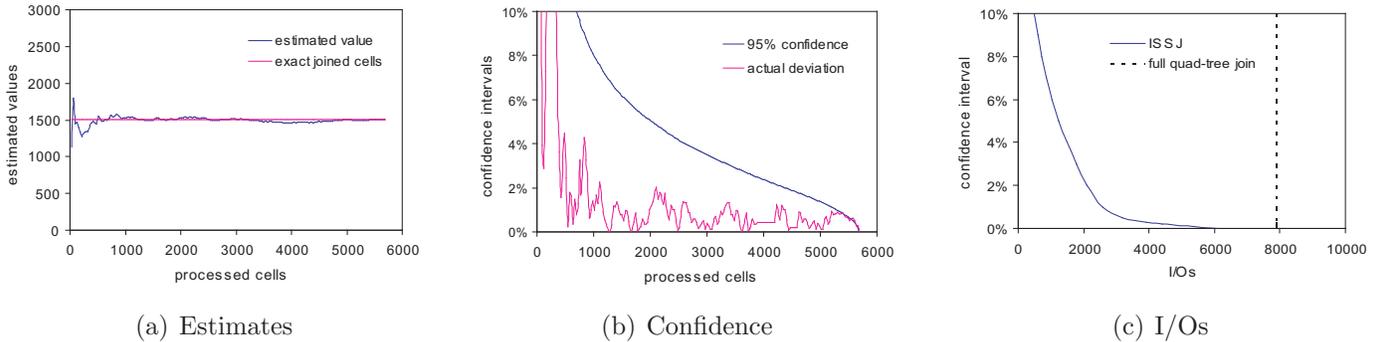


Figure 5: Estimates, confidence intervals and I/Os of *ISSJ*: unconsolidated sediment \bowtie mineral in CO

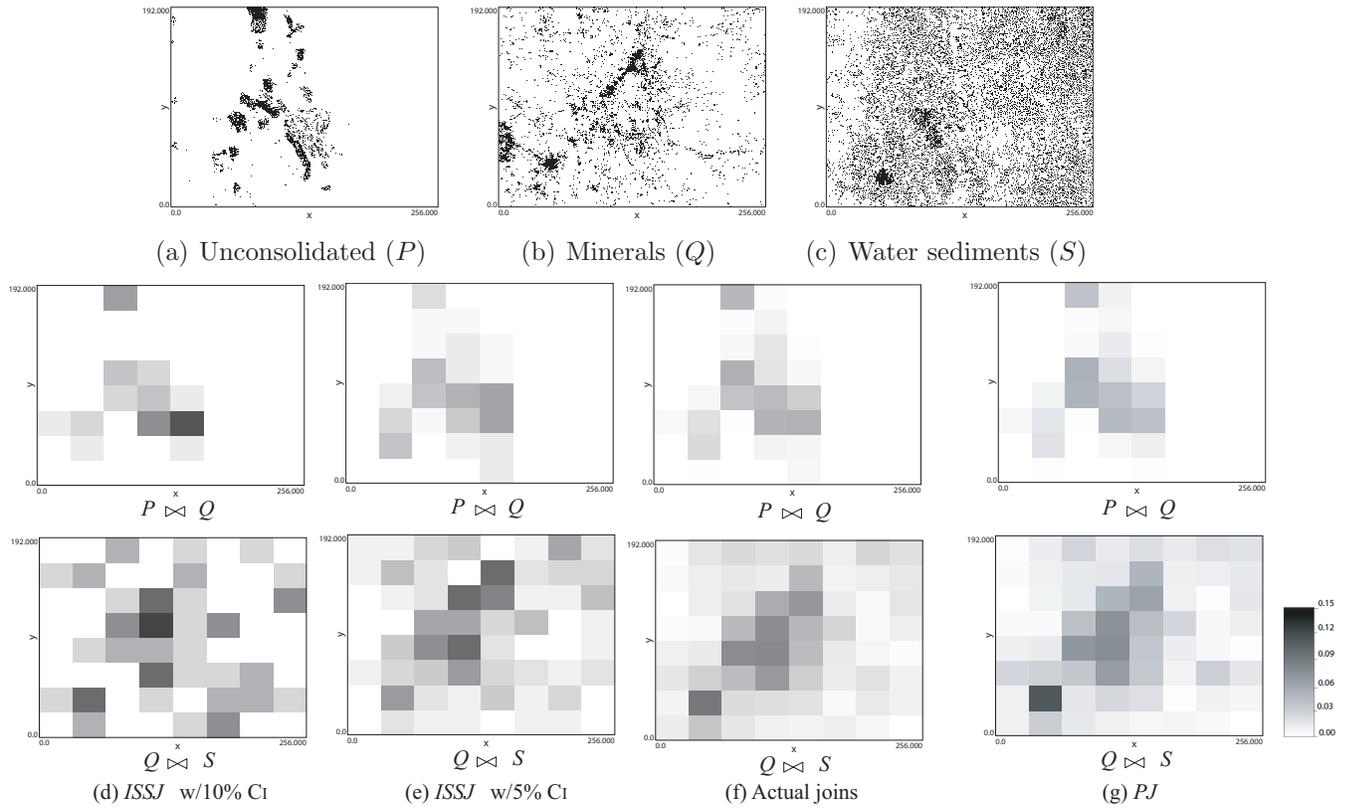


Figure 6: Expected number of joins: $ISSJ$ vs. PJ vs. Actual joins for real datasets in CO

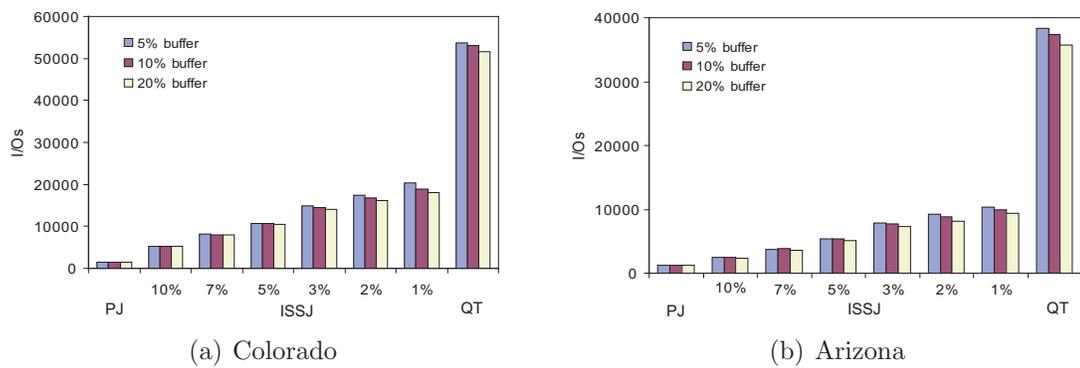


Figure 7: Number of I/Os of PJ , $ISSJ$ and the full quad-tree join

Colorado). The estimates and confidence intervals are plotted versus the number of samples (non-zero data cells) processed as well as the exact answer. Figure 5 (a) shows the estimated values of the final joins calculated by *ISSJ*. Figure 5 (b) shows how fast the confidence intervals converge. By showing the deviations from the actual joins, we demonstrate that *ISSJ* provides good estimates of the final answer. In Figure 5 (c), we showed how fast an accurate estimation could be calculated compared to the time required for the full quad-tree join. For example, it took about 1900 I/Os to reach an estimate with a 5% confidence interval while 8,000 I/Os were required for the exact answer obtained by the full quad-tree join.

We next show how accurately the proposed approaches provide a “big picture” of the actual join. Figure 6 (a), (b) and (c) show three datasets for the state of Colorado: unconsolidated sediments (P), minerals (Q) and water sediments (S). The results of PJ and *ISSJ* for $P \bowtie Q$ and $Q \bowtie S$ are presented as well as that of the actual join. The result from left to right corresponds to: *ISSJ* with a 10% confidence interval (d), *ISSJ* with a 5% confidence interval (e), actual joins (f) and finally PJ of the 4th level nodes (g). PJ and *ISSJ* with a 5% confidence interval provided a reasonably accurate approximation of the actual join.

In Figure 7 we present I/O comparisons between PJ and *ISSJ* with varying the confidence intervals, as well as with the full nonaugmented quad-tree join (QT). All possible pairwise joins from the six datasets of CO and AZ were run and the number of I/Os were plotted for buffer sizes of 5%, 10% and 20% of the size of one dataset quad-tree. We plot the average total number of I/Os of each method averaged over all 15 pairwise joins. The results for PJ are on the left, then *ISSJ* for confidence interval bounds of 10, 7, 5, 3, 2 and 1%, and finally the results for the full quad-tree join on the right. Note that the performance difference varying buffer sizes is very small since there is few re-visiting of the leaf nodes hence little opportunity to benefit from buffer caching.

The PJ method resulted in up to two orders of magnitude less I/Os than QT for both datasets. The *ISSJ* algorithm obtained a very reasonable confidence interval (e.g. 5%) with far less I/Os compared to QT . PJ is significantly faster than the *ISSJ* algorithm, but does not provide correctness bounds. However, PJ does provide a good overall picture for the data explored even though there is no guarantee of the quality of the estimate.

6 Conclusions and Future Work

Due to the large dataset size, spatial joins of GIS data may take unreasonably long time to complete. The traditional map overlay joining method does not provide any idea of how the final result will look like until the join is completed. Hence, to enable an interactive data exploration, it is essential to allow a user to get a fast estimation, ideally a “big picture” visualization, of the join result. User comfort in using approximations can be increased by a method that also provides a confidence interval bound on the estimate.

In this report, we studied two statistical approaches for estimating spatial joins on quad-tree indexed raster data, namely, Probabilistic Joins (*PJ*) and Incremental Stratified Sampling Joins (*ISSJ*). We proposed a framework that combines two statistical approaches to allow fast interactive data explorations and the opportunity for the user to then drill down with full spatial joins if desired. Experimental evaluations on real and synthetic datasets showed that our proposed *PJ* method resulted in reasonably accurate results with near zero response time. Our *ISSJ* method, while not as fast as *PJ*, provides results with bounded confidence intervals up to an order of magnitude faster than full quad-tree join. Our framework can be used to build an end-user query visualization tool that allows true interactive exploration of large raster based GIS databases.

This study used synthetic data for the evaluation of the proposed techniques. In the future we plan to evaluate the *PJ* method using real water resource dataset.

References

- [1] L. G. Azevedo, R. H. Gting, R. B. Rodrigues, G. Zimbardo, and J. M. de Souza. Filtering with raster signatures. In *Proceedings of ACM GIS*, pages 187–194, 2006.
- [2] W. D. Bae, S. Alkobaisi, and S. T. Leutenegger. An incremental refining spatial join algorithm for estimating query results in GIS. In *Proceedings of DEXA*, pages 935–944, 2006.

- [3] W. D. Bae, S. Alkobaisi, and S. T. Leutenegger. *irsj*: Incremental refining spatial joins for interactive queries in gis. In *Technical Report DU-CS-07-10*. University of Denver, 2007.
- [4] T. Brinkhoff, H. P. Kriegel, and R. Schneider. Comparison of approximations of complex objects used for approximation-based query processing in spatial database systems. In *Proceedings of ICDE*, pages 40–49, 1993.
- [5] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *Proceedings of ACM SIGMOD*, pages 551–562, 2003.
- [6] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter. Efficient join processing over uncertain data. In *Proceedings of CIKM*, pages 738–747, 2006.
- [7] P. J. Hass. Large-sample and deterministic confidence intervals for online aggregation. In *Proceedings of SSDM*, pages 51–63, 1997.
- [8] F. Olken. *Random Sampling from Databases*. PhD thesis, University of California at Berkeley, 1993.
- [9] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, MA, 1990.
- [10] R. J. Serfling. *Basic Statistics for Business and Economics*. McGraw-Hill, 2002.
- [11] H. Tveite. *Data Modeling and Database Requirements for Geographical Data*. PhD thesis, University of Norway, 1997.
- [12] USGS. Mineral resources on-line spatial data: <http://tin.er.usgs.gov/>, 2001,2005.
- [13] M. Vassilakopoulos and Y. Manolopoulos. On sampling regional data. *Data and Knowledge Engineering*, 22:309–318, 1997.
- [14] G. Zimbro and J. M. de Souza. A raster approximation for the processing of spatial joins. In *Proceedings of VLDB*, pages 558–569, 1998.

Application of Spatiotemporal Informatics to Water Quality (Phase II)

Basic Information

Title:	Application of Spatiotemporal Informatics to Water Quality (Phase II)
Project Number:	2009DC105B
Start Date:	3/1/2009
End Date:	2/28/2010
Funding Source:	104B
Congressional District:	District of Columbia
Research Category:	Engineering
Focus Category:	Methods, Water Quality, Non Point Pollution
Descriptors:	None
Principal Investigators:	Byunggu Yu, Pradeep K. Behera

Publications

There are no publications.

Application of Spatiotemporal Informatics to Water Quality (Phase II)

Final Report



UNIVERSITY OF
THE DISTRICT
OF COLUMBIA



Byunggu Yu, Ph.D.
Pradeep K. Behera, Ph.D., P.E., D. WRE

School of Engineering and Applied Sciences
University of the District of Columbia

May 2010

Environmental urban runoff monitoring

Byunggu Yu*, Pradeep K. Behera, Seon Ho Kim, Paul Cotae, Juan F. Ramirez Rochac,
Travis Branham

University of the District of Columbia, 4200 Connecticut Ave. NW, Washington, DC, USA 20008

ABSTRACT

Urban stormwater runoff has been a critical and chronic problem in the quantity and quality of receiving waters, resulting in a major environmental concern. To address this problem engineers and professionals have developed a number of solutions which include various monitoring and modeling techniques. The most fundamental issue in these solutions is accurate monitoring of the quantity and quality of the runoff from both combined and separated sewer systems. This study proposes a new water quantity monitoring system, based on recent developments in sensor technology. Rather than using a single independent sensor, we harness an intelligent sensor platform that integrates various sensors, a wireless communication module, data storage, a battery, and processing power such that more comprehensive, efficient, and scalable data acquisition becomes possible. Our experimental results show the feasibility and applicability of such a sensor platform in the laboratory test setting.

Keywords: sensor, water, flow measurement, stormwater, runoff, monitoring

1. INTRODUCTION

Over the last couple of decades, it has been recognized that urban stormwater runoff is a large contributor to water quantity and quality problems of many receiving waters (e.g., lakes and rivers⁴). During wet-weather events (e.g., rainfall and snowmelt) the uncontrolled runoff from urban watersheds may result in excessive flooding, stream-bank erosion and transportation of a wide spectrum of pollutants to local receiving waters¹⁴. The pollutants in urban runoff include suspended solids, oxygen demanding materials, nutrients, pathogenic micro-organisms, and toxicants, such as heavy metals, pesticides, and hydrocarbons. These pollutants impose considerable physical, chemical, and biological stresses on the receiving water that affect aquatic life and human health⁸, impairing the designated uses of water resources. Typical runoff problems include the degradation of aquatic habitats, degradation in water quality during and after wet weather events, beach closures, accelerated eutrophication in lakes and estuaries, and thermal pollution¹⁴. These problems have been prevalent in most receiving water systems in the vicinity of urban or urbanizing areas.

In response to growing public concerns and Clean Water Act requirements for attaining the minimum water quality standards in the nations receiving waters, which are affected by Combined Sewer Overflows (CSOs) and stormwater discharges, the U.S. Environmental Protection Agency (EPA) has promulgated several regulations. One of them includes a new CSO Control Policy requiring municipalities develop and implement CSO system monitoring programs for planning, compliance, and reporting purposes. For example, under Section 303 (d) of Clean Water Act, municipalities with separate sewer systems (MS4) contributing pollutants of concern to an impaired water body are required to develop quantitative limits to meet the state and federal National Pollutant Discharge Elimination System (NPDES) regulations (U.S. EPA, 2008⁷). The new regulatory focus on long-term CSO control increases the need for accurate flow rate monitoring and measurement systems that may be deployed at many locations within the city-wide sewer networks and receiving water outfalls.

*byu@udc.edu; phone 1 202 274-6289; csit.udc.edu

Nationally, controlling urban stormwater pollution has cost the US several billion to hundreds of billions of dollars¹⁰. To develop efficient urban stormwater management solutions, engineers follow three key steps – 1) assessment of pollution through short-term monitoring (collection of rainfall, runoff flow data, water quality data and other hydrologic characteristics), 2) development of a continuous simulation model of the system using the collected field data (calibration and verification of model), and 3) development of alternative solutions based on different CSO and stormwater management scenarios.

Monitoring and measuring actual CSOs and stormwater discharges are complex and very expensive due to its intensive resource requirements (e.g., costly equipments and personal deployment). Consequently limited field data is collected at only key locations and a simulation model, such as Stormwater Management Model (SWMM), is used for predicting the runoff quantity and quality. However, achieving high accuracy with flow measurements is critical to meaningful system modeling and analysis, since the propagation of uncertainties (errors) through flow networks can rapidly grow to unmanageable proportions².

Although continuous runoff monitoring is an ideal solution, often it is understated or conducted just on a short-term basis due to the exorbitant cost of equipment installation, operation, and maintenance. Continuous and real-time monitoring and measuring of runoffs over a large geospatial region is highly cumbersome. Nevertheless, such real data is crucial for the civil infrastructure maintenance and design and for emergency decision making and planning. As the civil infrastructure expands and ages, the demand for a cost-efficient instrumentation for measuring the runoff quantity and quality is ever more pronounced.

Such real data is also beneficial for several other key operations including planning and evaluation of new CSO and stormwater control alternatives, regulatory reporting and compliance documentation, operation alternation to CSO system malfunctions, optimization of the operation of the treatment facilities, and the allocation of the user cost and billing for the operation and maintenance of the system.

This paper targets a flow monitoring system for runoffs at sewer outfalls with the following specific design goals: 1) cost efficiency – the development of a cost-efficient (in terms of installation, operation, maintenance, and geospatial scalability) sensor technology for continuous monitoring of stormwater runoffs and CSOs, 2) scalability – the development of a cost-efficient and geo-spatially scalable wireless communication and information system that can utilize such sensor technology and that can continuously monitor, record, and report real-time flow data with minimal human support and that can actively alert human as necessary, and 3) Reliability – to provide a reliable measurement of runoff quantity (i.e., flow rate and cumulative volume), which will provide reliable estimations of pollutant loads for the maximum pollution limits (Total Maximum Daily Loads) analysis.

2. RELATED WORK

Flow monitoring for runoffs at sewer outfalls (discharge pipes along the receiving water) is complex because of the intermittent nature of flow. The applicable solutions vary greatly in complexity, cost, and accuracy. They can be classified in a broad sense as either direct or indirect. Direct methods involve measurements of the quantity (volume or weight) of the flow for a given time interval for closed conduits flow or pressurized flow. For gravity flow, the principle of open channel flow is typically used. The flow rate is computed based on the velocity and area of the flow. Hydraulic engineers have developed equations that describe the relationship between the depth of an open channel/closed conduit and the velocity of gravity flow. For a given depth of flow, there is a predictable flow velocity in the pipe which can be derived using Manning's Equation⁵ or a curve fitting method like the Colebrook-White pipe curve³.

Indirect methods involve the measurement of a pressure change for closed conduits flow (or some other variable), which is directly related to the rate of flow. The examples include venturi meters, orifices, etc. Weirs are devices that employ indirect means to obtain partial flow rates. Another type of indirect devices includes electromagnetic flow meter, which operates on the principles that a voltage is generated when a conductor moves in a magnetic field.

Among the simplest methods are the conventional manual methods that include direct measurements of runoff quantity and quality (using various measuring tools, such as poles, bottle board and chalking, and dye testing). However, the manual methods rely extensively on labor-intensive field efforts during storm events and do not provide an accurate, continuous flow record⁶.

Recently developed techniques employ more advanced flow sensor devices:

1. Ultrasonic Flow Meter: This device employs a technique to measure the difference in travel time for a sound wave traveling upstream and downstream between two measuring stations. The difference in travel time is proportional to flow velocity. The other type of operation uses the Doppler effect, which is based on capturing the differential frequency (Doppler shift) during the projection of ultrasonic beam into a inhomogeneous fluid. The measured frequency difference is related to the flow velocity.
2. Ultrasonic Sensors – typically mounted above the flow in the pipes: The depth is computed based on the time the reflected signal returns to the sensor. The depth measurements can be affected by the suspended solids of the runoff.
3. Pressure Sensors – sense the pressure of water above them: They are used along with a flow monitor that converts pressure value to a depth measurement. Hard to use at an open end of a pipe.
4. Bubble Sensors – emits a continuous stream of fine bubbles: A pressure transducer senses resistance to bubble formation, converting it to a depth measurement value. The bubble tubes may become clogged, requiring frequent calibration.
5. Float Sensor – uses a mechanical float, often encapsulated, designed to damp out surface waves: Floats can clog with grease and solid materials.

Typically flow meters are used to monitor the discharge in combined sewer and storm sewers. These flow meters are required to operate under both free flowing (open channel conditions) and surcharged (pressurized) conditions. Ultrasonic transit-time flow meters are most commonly used for CSO monitoring and compliance because of their accurate and continuous flow rate monitoring capability during dry and wet weather conditions. Transit-time flow meters include bi-directional (reverse flow) measurement capability and can be configured for multiple acoustic paths, making them highly accurate over a wide range of changing water level and flow conditions, as well as in locations where other measurements methods cannot reliably function.

3. PROPOSED APPROACH

3.1 Wireless Sensor

Our proposed approach is based on a newly developed Sun Small Programmable Object Technology (SPOT¹³) computer (Figure 1). SPOT is a small, sensor-rich, wireless, battery-powered device developed at Sun Labs to explore the next frontier of network computing. This tiny computer-sensor platform consists of three stacked layers¹³: Li-Po battery, sensor board, and tiny computer with a CPU (Java programmable), timer (AT91 timer for measuring time elapses), USB, power switch, and memory. The sensor board includes the following: 3-dimensional accelerometer (LIS3L02AQ), temperature sensor, light sensor, eight LEDs, two switches, five general-purpose I/O pins, and four high current output pins. The advantage of SPOT is that one can easily attach extra sensors using USB and I/O. Thus, after one monitoring system for a specific sensor has been developed, it can be extended in a simple and straightforward way for various sensor applications.

This platform is convenient to create any kind of program using Java, a high-level and general-purpose programming language. Programmers can use any standard Java development tools to write codes for SPOT applications. The package came with a variety of demos which can give a quick, hands-on educational practice before getting into advanced programming.

The integrated accelerometer LIS3L02AQ is a low-power three-axis linear type⁹. We focus on the acceleration output from SPOT while it is moving. The output consists of the 3-dimensional accelerations in gravity units and the corresponding time in milliseconds. These raw values represent the X, Y, and Z accelerations of the sensor in any given point in time. The data is captured by the sensor and transmitted over a wireless network to the base station which records the data into a secondary storage (limited amount of in-device logging is possible).

(a)

(b)

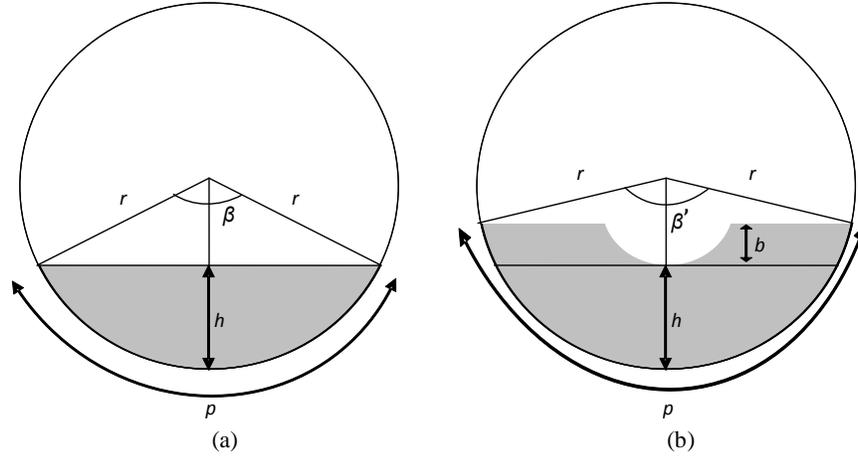


Figure 3. Calibration of the area of water flow

When there is no water in the pipe, the sphere is at the bottom and the tilt in the Z direction is zero ($\theta=0^\circ$). While water is flowing, the water level or height (h) can be continuously calculated using the tilt. For each h , the area A of the flow can be calculated using Equation 2 as shown in Figure 3.a, i.e., the area is a function of height $A=f(h)$ (Equation 2). However, the sphere with SPOT inside has a considerable weight which can make it submerged into the water, which may affect the accuracy of the water quantity. Hence, to calibrate the height h for accurate calculation of the water quantity, we introduce a buoyancy factor b which measures how deep the sphere is submerged. A preliminary buoyancy test was performed to quantify the buoyancy factor b . This resulted in a modified flow area A as shown in Figure 3.b. Hence, $A=f(h+b)-s$, where f is the function in Equation 2 and s is the white cut-off part of A in Figure 3.b representing the circular segment of the submerged portion of the SPOT sphere. In this case, $s=r_s^2/2\times(\theta-\sin \theta)$, where r_s is the radius of the SPOT sphere and $\theta=2\arccosine((r_s-b)/r_s)$. Note that the buoyancy factor b becomes critical when the ratio between the diameter of the pipe and the diameter of the sphere gets small, which is the case of our laboratory test setting. Equation 4 shows the basic flow area function that our lab developed for the proposed sensor sphere solution.

$$A = \frac{r^2}{2} (\beta' - \sin(\beta')) - \frac{r_s^2}{2} (\beta_s - \sin(\beta_s)), \text{ where}$$

r is the radius of the pipe, r_s is the radius of the sensor sphere, (4)

$$\beta' = 2\arccosine\left(\frac{r-h'}{r}\right), \beta_s = 2\arccosine\left(\frac{r_s-b}{r_s}\right),$$

$h' = h + b$, and b is how deep the sensor sphere is submerged.

Regarding the flow velocity, Equation 3 can be applied to the steady gravity-driven flows of the actual sewer pipes in the field. Considering a short small-diameter pipe in a laboratory setup as shown in Figure 5, however, it is hard to simulate such flow. The flow enters the pipe with some initial velocity, which makes it hard to straightforwardly use the Manning's equation to calculate V . Thus, in our experiments, the water velocity was separately measured and used in the calculation of water quantity.

3.3 Noise Reduction: Kalman Filtering

Like any sensor outputs, SPOT's acceleration outputs are vulnerable to errors, i.e., noise. Moreover, flowing water creates ripples which make sensor outputs fluctuate. This can affect the accuracy of sensor outputs. Signal filtering is a well-known approach for acquiring cleaner and more accurate sensor outputs. We adopted a linear Kalman filter algorithm to minimize the noise. The Kalman filter is an optimal linear estimator first introduced in 1960¹¹. This filter is recursive and it can estimate the state of a linear dynamic system from a series of noisy measurements.

The Kalman filter has two distinct phases: Prediction and Correction. The first phase uses the state estimate from the previous step to calculate an estimate of the state at the current step generating the *a priori* state estimate which does not include any observation from the current step. In the second phase, the correction, the current *a priori* estimation is combined with current observation information to filter the state estimate generating the corrected state estimate, which is called the *a posteriori* state estimate. We used the simple linear Kalman filter¹¹ with the following parameters: $H=1$,

$Q=0.00001$, and $R=0.1$.

In summary, this study proposes a smart threefold approach to flowing water monitoring using wireless sensors with 3D accelerometers for continuous measuring of flowing water quantity in a pipe. The sensor platform provides the enhanced sensing and communication resources and is expandable for future water quality monitoring. The real-time measurement is done by our custom-made functions enhanced by a linear Kalman filter. Then common hydraulic principles and formula are used for the higher-level applications. The following section describes a comprehensive set of our lab experiments and discusses how the presented approach is implemented and resulted.

4. EXPERIMENTS

4.1 Experimental Setting

The major part of this study is focused on the testing of proposed instrumentation in the laboratory environment. The experiment was conducted with a pipe that has a diameter of 0.2 meters as shown in Figure 5. The polyvinyl chloride pipe is 1.4m long and its Manning coefficient “ n ” (level of roughness of the inner wall)^{1,12} is 0.011. As introduced earlier, the diameter of our lab-made sensor sphere prototype was 5 inches, which is 0.127m. Another parameter that is important to note is the buoyancy factor, which in the case of our sensor package results in a $b=0.025\text{m}$ (i.e., how deep the sphere submerges in water). Summary: $n=0.011$, $r=0.1\text{m}$, $r_s=0.0635\text{m}$, $b=0.025\text{m}$.

The acceleration data is captured by the SPOT which is encased in the water-tight package attached to the inside of the pipe. As introduced earlier, a run-about hamster ball from a local pet store was used. We secured the SPOT in the middle of the ball using some styrofoam-like material. Then we used fishing line to attach the sphere in a stable V-shape (two connecting lines) inside the pipe in such a way that the sphere can barely reach the bottom of the inside of the pipe and swing freely. The data is wirelessly transmitted to the base station. The laptop provides the human interface to store, query and retrieve this data to allow further offline analysis. The entire process is illustrated in Figure 6.

The experiments implemented the following steps:

1. Instrumentation: Encase a SPOT computer inside a sealed plastic sphere and tether it inside the pipe. A lab-made small pipe and artificial water input were tested.
2. Runoff logging: The 3D acceleration data stream (swing motions sequence) from the sensor represents the simulated runoff flow over time.
3. Filtering and Aggregation: The programmable SPOT filters out insignificant events and noise for optimal use of storage and communication bandwidth and can provide representative summary data (pre-processing).
4. All logged data is automatically transmitted to the laptop’s hard disk through the wireless communication channel.

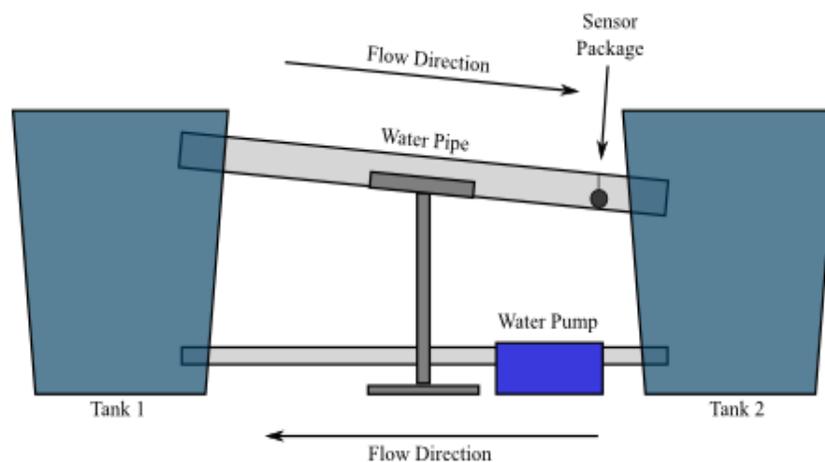


Figure 5. Runoff monitoring setting

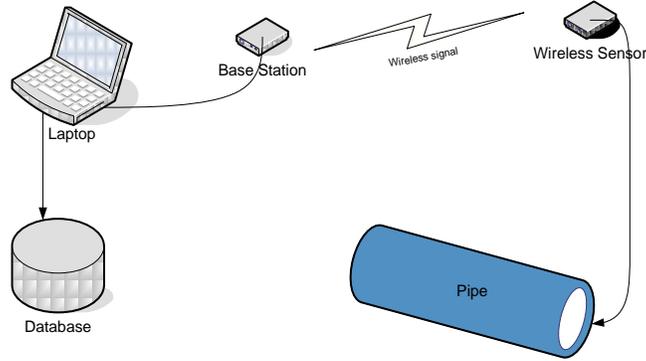


Figure 6. Civil-informatics overview

4.2 Results

We performed two sets of tests using the setup described in Section 4.1. For the first set, the distance between the tanks was the 1.29 meters (51 in) and the difference in heights was 0.076 meters (3in). Thus the pipe had a slope of 3/51. In the second set of tests, we changed the height to 0.11 meters (4.5in), resulting in a slope of 4.5/51. In each test, we poured a bucket of water (measured amount) through the higher end of the pipe and collected accelerometer outputs from the SPOT sphere while the water was flowing. We repeated this ten times for each of the two sets.

Figure 7 shows one example of the raw data taken by the sunspot in an actual test. As expected, there were little changes in A_x and A_y values while obvious changes were observed in A_z values. At the same time, we can see some spikes which demonstrates the existence of noise in the signals. Figure 8 illustrated the results of applying the Kalman filter on a test data set. It clearly shows the smoothing effect on signals, which results in a clearer signal.

Tables 1 and 2 summarize the results of water quantity monitoring using SPOT outputs. The second column, *Expected value*, is the actual amount of water that actually flowed in the test. And the fourth column, *Calculated result*, is the estimated quantity calculated using the SPOT solution without the Kalman filter. All values are in liters. On average, 13% of error in the first series of tests and 12% in the second series of tests were observed, respectively. When we applied the Kalman filtering and recalculated the estimated results, the average errors reduced to 6% and 5%, respectively. We believe that this range of error margin is quite satisfactory and our approach can be applied and tested in real sewer outfalls without any problems. We reserve this as our future work.

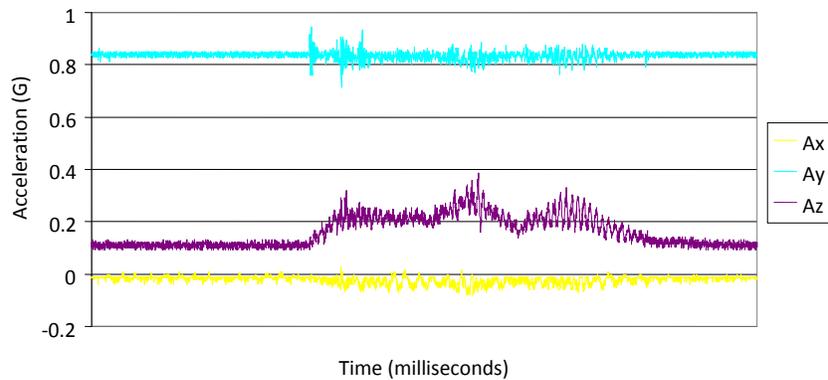


Figure 7. Example of raw output from the Sun SPOT accelerometer.

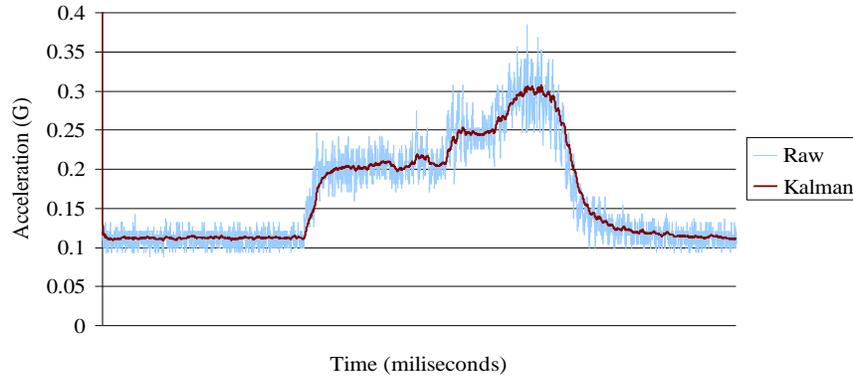


Figure 8. Effect of Kalman filter over the noisy acceleration input (A_z).

We also recognized that the calibration factor b is vital for the accurate calculation of water quantity. In some preliminary experiments without the calibration factor, the relative error was quite significant (underestimation by more than a factor of two). After measuring and incorporating the actual buoyancy b into the equation (Equation 4), we achieved far more consistent results.

Table 1. Summary of experiments with slope $S = 3/51$.

Test No.	Expected value (L)	Duration (s)	Calculated result (L)	Relative error %	Calculated result Kalman filter (L)	Relative error %
1	5.306	25.337	5.9292	12%	5.2883	0%
2	5.266	23.730	5.6516	7%	5.2846	0%
3	5.421	24.108	5.6198	4%	5.4252	0%
4	4.801	24.449	5.3884	12%	4.9453	3%
5	5.039	30.293	7.2599	44%	6.7135	33%
6	4.993	23.665	5.3799	8%	5.0067	0%
7	4.711	26.266	4.8168	2%	4.7209	0%
8	5.193	21.310	5.6165	8%	5.3978	4%
9	4.946	23.877	6.0078	21%	5.5267	12%
10	6.958	22.883	6.4912	7%	6.3930	8%
Average	5.791	24.592	6.4112	13%	5.4831	6%

Table 2. Summary of experiments with slope $S = 4.5/51$.

Test No.	Expected value (L)	Duration (s)	Calculated result (L)	Relative error %	Calculated result Kalman filter (L)	Relative error %
1	5.381	22.637	5.4587	1%	5.889042	9%
2	4.961	23.271	5.2450	6%	5.625359	13%
3	5.459	23.611	5.3863	1%	5.437048	0%
4	4.986	30.301	6.8815	38%	4.998436	0%
5	5.802	30.338	8.5076	47%	6.5046	12%
6	5.194	27.461	5.6707	9%	5.4442	5%
7	5.437	27.180	5.8900	8%	5.4895	1%
8	5.228	25.730	5.2993	1%	5.3318	2%
9	5.531	26.254	5.9421	7%	5.7884	5%
10	5.491	26.679	5.7190	4%	5.3180	3%
Average	5.347	26.346	6.0000	12%	5.3877	5%

4.3 Performance Enhancement

In remote sensor applications, memory usage and power consumption are of primary concerns. The maximum sampling rate of SPOT is 1 sample per 0.006 seconds. Since the on-board memory storage is very limited, we experimented with the possibility of reducing the sampling rate and its impact on the accuracy of the monitoring. This experiment was conducted in order to find a knee point in the accuracy – sampling rate space for more efficient use of limited power and memory on board the remote sensor platform. We derived the water flow using different sampling rates: 1/20T samples, 1/50T samples, 1/100T samples, and 1/500T samples. A sampling rate of 1/20T denotes that we only use 1 out of each 20 consecutive measurements in a non-replacement manner. In other words, instead of using one sample every six

milliseconds, we used one every 120 milliseconds. Some of the results of this analysis are summary in Table 3, which is based on the results of Calculated Results of Tests 1-4 and Calculated Results Kalman Filter of Tests 5-10 in Table 1. The knee point that we see in the result is 1/100T. In our other tests, different knee points were found, showing that there is room for improvement. A smart sampling rate decision module that can dynamically adapt to the continuously changing inputs (water flow) and power supply (varying amounts of sun lights for the field solar panel powering the sensor platform) will be valuable, especially in the field setting.

Table 3 Impact of reducing the sampling rate on the relative error.

Test No.	1	2	3	4	5	6	7	8	9	10	AVG
Time in seconds	25.337	23.73	24.108	24.449	30.293	23.665	26.266	21.31	23.877	22.883	-
Actual value (L)	5.306	5.266	5.421	4.801	5.039	4.993	4.711	5.193	4.946	6.958	-
Max Sampling Rate 1/1	5.9292 12%	5.6516 7%	5.6198 4%	5.3884 12%	6.7135 33%	5.0067 0%	4.7209 0%	5.3978 4%	5.5267 12%	6.393 8%	9%
Sampling Rate 1/20	5.7352 8%	5.5229 5%	5.7525 6%	5.1712 8%	6.8246 35%	5.1118 2%	4.6232 2%	5.4736 5%	5.4712 11%	6.288 10%	9%
Sampling Rate 1/50	5.4484 3%	5.9976 14%	5.7814 7%	4.9635 3%	6.1574 22%	5.161 3%	4.9906 6%	5.801 12%	5.6329 14%	6.4084 8%	9%
Sampling Rate 1/100	5.3095 0%	5.7887 10%	5.1542 5%	4.8101 0%	6.7302 34%	5.4653 9%	5.1981 10%	5.5171 6%	5.8982 19%	6.3006 9%	10%
Sampling Rate 1/166	6.1521 16%	6.8205 30%	6.5188 20%	5.1306 7%	7.6049 51%	4.233 15%	5.2376 11%	5.2077 0%	5.2998 7%	6.6296 5%	16%
Sampling Rate 1/500	5.5916 5%	5.1229 3%	4.6592 14%	5.2643 10%	7.9503 58%	4.1699 16%	5.2499 11%	5.0419 3%	4.0894 17%	5.9738 14%	15%

5. CONCLUSIONS AND FUTURE DIRECTIONS

We investigated the problem of continuous monitoring of urban runoff at outfall points. The paper presented the conceptual basis, technical details, and experimental results of a newly developed remote monitoring solution based on an advanced sensor platform. A prototype and accompanying algorithms were developed using the Sun SPOT as a sensor platform. Consequently, the collected accelerometer data were processed and analyzed in various ways to quantify the amount of water flow in the pipe. Our experimental results demonstrated that our approach has a great potential to measure the water quantity with any desirable precisions required in real applications of urban runoff monitoring. The platform is flexible and expandable and provides a possibility for monitoring the water quality details of the flow. The lab test results are promising and, based on this, we are planning to apply the proposed technology in various real sewer waterfalls in the field in near future. In addition to the SPOT, we will also study the application of other available platforms, such as iMote2, in a comparative manner. In the field, more challenges are expected various areas including on-site power (solar panel charging the sensor platform), cellular communication or wide-area networking for automated data collection and run-time reprogramming of the platform (data and control transmission between central host computer and the field devices), memory and power optimization, to list a few.

6. ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation (NSF) Grant CMMI-0940393 and DC Water Resources Research Institute.

7. REFERENCES

- [1] Bishop, R.R. and Jeppson, R.W., "Hydraulic characteristics of PVC sewer pipe in sanitary sewers", Utah State University. Logan, Utah. September (1975).

- [2] Burch, T.L., and Phillips, J.M., 1995, "High-accuracy CSO and stormwater flow monitoring" in Torno, HC, ed., Proceedings of the Engineering Foundation Conference, *Stormwater NPDES related monitoring needs*, Crested Butte, Colo., American Society of Civil Engineers, August 7–12, 609–616 (1994).
- [3] Crowe, C.T., Elger D. F., and Roberson, J.A., "Engineering Fluid Mechanics", New Jersey, U.S.A.: John Wiley & Sons, New Jersey (2005).
- [4] DC Water and sewer authority (DC WASA), "A District of Columbia Water and Sewer Authority Biannual Report", October (2005).
- [5] Edwards, K., "Manning's Equation", <http://www.lmnoeng.com/manning.htm>, LMNO Engineering, Research, and Software, Ltd. (Retrieved Jan 2010).
- [6] EPA, "Combined Sewer Overflows, Guidance for Monitoring and Modeling", EPA Report 832-B-99-002 (1999).
- [7] EPA, "Understanding Impaired Waters and Total Maximum Daily Load (TMDL) Requirements for Municipal Stormwater Programs", EPA Report 833-F-07-009, January (2008).
- [8] Field, R., Borst, M., O'Connor, T.P., Stinson, M.K., Fan, C., Perdek, J.M., and Sullivan, D., "Urban wet-weather flow management: research directions", *Journal of Water Resources Planning and Management*, 124(3), 168-180 (1998).
- [9] Goldman, R., "A Sun SPOT Application Note: Using the LIS3L02AQ Accelerometer", <http://www.sunspotworld.com/docs/AppNotes/AccelerometerAppNote.pdf>, Sun Microsystems, Inc (Retrieved May 2009).
- [10] Heaney, J.P., Wright, L., and Sample, D., "Research Needs in Urban Wet Weather Flows", *Water Environmental Research*, Vol. 71, No. 2, 241-250 (1999).
- [11] Kalman, R.E., "A New Approach to Linear Filtering and Prediction Problems", *Journal of Basic Engineering*, 82 (Series D): 35–45 (1960). <http://www.elo.utfsm.cl/~ipd481/Papers%20varios/kalman1960.pdf> (Retrieved Oct 2009)
- [12] Neale, L.C. and Price, R.E., "Flow characteristics of PVC sewer pipe", *Journal of the Sanitary Engineering Division*, Div. Proc 90SA3, ASCE. 109-129 (1964).
- [13] Sun Microsystems, "Sun Small Programmable Object Technology (Sun SPOT) Theory of Operation", Sun Microsystems, Inc. (2007).
- [14] World Environment Federation (WEF), "Urban Runoff Quality Management", *Manual of Practice No.23*, ASCE, Alexandria, VA (1998).

Development of Web-based Rainfall Statistical Analysis Tool for Urban Stormwater Management Analysis (Phase II)

Basic Information

Title:	Development of Web-based Rainfall Statistical Analysis Tool for Urban Stormwater Management Analysis (Phase II)
Project Number:	2009DC106B
Start Date:	3/1/2009
End Date:	2/28/2010
Funding Source:	104B
Congressional District:	
Research Category:	Engineering
Focus Category:	Models, Wastewater, Methods
Descriptors:	None
Principal Investigators:	Pradeep K. Behera

Publications

There are no publications.

Development of Web-based Rainfall Statistical Analysis Tool for Urban Stormwater Management Analysis (Phase II)

Final Report



UNIVERSITY OF
THE DISTRICT
OF COLUMBIA



Pradeep K. Behera, Ph.D., P.E., D. WRE

School of Engineering and Applied Sciences
University of the District of Columbia

May 2010

Development of a Rainfall Statistical Analysis Tool for Analytical Probabilistic Models for Urban Stormwater management Analysis

Travis Branham, CS Major

Department of Computer Science
University of the District of Columbia, Washington DC, 20008, USA
Email: tbranham@gmail.com

Pradeep K. Behera, Ph.D., P.E.

Department of Engineering, Architecture & Aerospace Technology,
University of the District of Columbia, Washington DC, 20008, USA
Tel: 202 274 6186, Fax: 202 274 6232, Email: pbehera@udc.edu

ABSTRACT

In this research paper, the frequency analysis of point rainfall data is examined. The emphasis has been placed on the statistical analysis of storm events. The design of urban stormwater management systems based on analytical probabilistic modeling approach depends on the statistical analysis of input meteorology, rainfall. The long-term rainfall record is discretized into independent storm events by defining an inter event time definition (IETD) and each event is characterized by four event characteristics (e.g., rainfall event volume, duration, intensity, interevent time). The time series of discretized storm events representing each of the four characteristics are fitted with probability density functions. The parameters of the PDFs of rainfall characteristics constitute the input to the analytical probabilistic models. As a complement to continuous simulation models (e.g., US EPA SWMM), the computational efficient analytical probabilistic models can use parameters of PDFs as the input to the model for urban stormwater analysis.

The objective of this research is to design and build a software utility for engineers and professionals which will perform the aforementioned statistical rainfall analysis. The research methodology is as follows: (1) various sources of freely accessible rainfall records were explored such as (NOAA and NCDC web sites); (2) Class and Relationship diagrams were developed for the overall system architecture (note that the system is separated into a back-end parsing engine, and a series of front-end applications); (3) the Python programming language was selected for development, and the back-end system architecture was implemented; (4) a series of tests were performed to assess the proper functionality of the system; (5) the front-end systems, including a plotting application and a web-based interface, were developed. This tool can be useful for any location in the United States that has a viable rainfall record, and could be used to generate a comprehensive atlas.

1.0 Introduction

In order to protect society and environment adequately from the stormwater impacts such as flooding, erosion and receiving water problems engineers and professionals use stormwater management models for planning and design level analysis. These models require adequate representations of both hydrologic and hydraulic behavior of the drainage system (both sewer system and watershed) in order to size and configure control system elements. There are three major methods used to model urban drainage needs: event-based models, continuous simulation

models, and analytical probabilistic models. However, for a comprehensive analysis of stormwater quantity and quality problems management models are mathematical models which use continuous simulation approach rather than single design event approach. Continuous simulation models are physically based use a long-term rainfall records to simulate catchment hydrology and hydraulics, and pollutant processes with long-term meteorological records as model input. However, the use of continuous simulation during screening-level analysis is relatively cumbersome, time consuming and very expensive.

Alternative approaches to continuous simulation, a set of analytical probabilistic models have been developed for screening-level analysis (Adams and Papa, 2000). The basic premise for both the continuous simulation and analytical modeling approaches remains same – long-term meteorology is the input to the model. In continuous modeling approach, the time series of rainfall records is used as input; however, in the analytical modeling approach the same rainfall record is pre-processed to determine the probability density functions (PDF) of rainfall characteristics (e.g., rainfall volume, duration, intensity and inter-event time) which are used as input to the model. The analytical models are derived based on the Derived Distribution Theory. In this technique, the PDF of a dependent variable is derived from the PDF of independent variables using the functional relationship between dependent and independent variables. For the stormwater management models, the PDFs of rainfall characteristics constitute input to the model and simpler hydrologic and pollutant processes that are similar to continuous simulation models constitute functional relationship. Since these analytical models are often closed form algebraic equations, they are more computationally efficient compared to simulation models. These models can be used to analyze the runoff pollution condition for each of the sub-catchments within a large watershed.

The analytical models are also developed based on different hydrologic models such as US Army Corps STORM hydrology and U.S. EPA's SWMM hydrology (Adams and Papa, 2000, Adams and Bontje, 1984, Guo and Adams, 1998 and Behera et. al. 2006). In order to facilitate the use of analytical probabilistic models, a statistical analysis of long-term rainfall record is necessary. The proposed research developed a software utility tool to conduct the statistical analysis of rainfall records and estimate the parameters of fitted probability distribution functions.

2.0 Development of the Tool

In order to develop the rainfall statistical analysis tool, the input, an isolated meteorological event, at one point in space as described as a hyetograph must first be defined. The storm event can have both internal and external characteristics (Adams et al., 1986). The external characteristics are the total storm event volume, the duration of the storm, the average intensity of the storm and the interevent time or duration since the last storm. The internal characteristics are both numerous and complicated such as number of peaks and time to peak etc. stormwater model typically use the external characteristics of rainfall events which are used for this analysis. The following sections will describe the details derived from the rainfall records which will be analyzed by the tool.

2.1 The Rainfall Event

A chronological rainfall record may be split up into two distinct groups of time periods: rainfall “events”, and the intervening times between rainfall “events”. Here, a rainfall event is characterized by some measurable precipitation. The available continuous chronological rainfall record is first discretized into individual rainfall events separated by a minimum period without rainfall – termed the interevent time definition (IETD). If the time interval between two consecutive rainfalls is greater than the IETD, the rainfall events are considered as two separate events. Once this criterion is established, the rainfall record is transformed into a time series of individual rainfall events and each rainfall event can be characterized by its volume (v), duration (t), interevent time (b) and average intensity (i). Next, a frequency analysis is conducted on the magnitudes of the time series of rainfall event characteristics, from which histograms are developed. Probability density functions are then fitted to these histograms. The intensity parameter is a calculated value given by: $i = v/t$

2.2 The Rainfall Record

The first, and most important, design consideration when writing the Rain Event Parser was to determine the likely source material for rainfall records. A decision was made to build the utility to parse the hourly precipitation data files produced by the National Climatic Data Center (NCDC, see: <http://www.ncdc.noaa.gov/oa/ncdc.html>). These files catalog hourly precipitation information, and present the records in a comma-delimited ASCII text file. The advantage of supporting files from the NCDC is that there is a great deal of coverage, both geographically and historically, in their database; a disadvantage is that the actual formatting of the data within each file is somewhat cumbersome to parse effectively.

The NCDC hourly precipitation data is transmitted as a collection of several files: 1) a general document which describes the format of the data file, 2) a file that contains the characteristics (including latitude and longitude) of the collection locations, 3) an inventory file which describes the date ranges that have been provided for each location, and 4) a file containing hourly precipitation values recorded at each location.

The actual format of an NCDC hourly precipitation data file is as follows: first, the files have been downloaded in a comma (,) delimited format; this is somewhat misleading, however, because the files are actually in a fixed width format, but also contain commas between fields. This becomes an issue because one of the flags that are added to certain data items (described later) is a comma character, and consequently some pre-processing of the data is required in order to preserve the column order.

Second, each line item following the header row represents a single day's worth of precipitation information. It is important to note that only days that have recorded precipitation (or days that contain information related to error flags) are present in the file; days with no recorded precipitation are not listed. There are 107 total columns in each data file. The first 7 columns contain descriptive information including the Cooperative Station Number (or Station ID, which can be cross-referenced with the Station information file to find the location of the site), the date, and the accuracy of the measurement (we are using files that have an accuracy of 1/100 inch). The remaining 100 columns contain a measurement for each hour in the day, several flag fields

for each measurement, and a total daily precipitation field.

Third, there are two flag values for each hour of data collected. The first flag field contains error or anomaly codes related to data collection (such as whether a trace amount of rainfall was recorded, for example) and the second field contains codes related to data quality (whether or not evaporation could have occurred, for example). Decisions have been made on how to handle each of the error cases; since the sites chosen for statistical analysis generally have long historical rainfall records, many of the error values are simply treated as zero precipitation, and are, therefore, effectively removed from the overall record.

Overall, parsing the data file is the most computationally intensive portion of the entire application, due to the up-front processing that needs to be done on each line item in order to account for the multitude of flag scenarios that may be encountered. Work is ongoing to reduce the number and type of operations that need to be performed on each record in order to improve the overall performance of the application.

It is important, however, that the design of the tool remain flexible enough that it will not require a significant overhaul of the software in order to process different source material, should that need arise. It is therefore quite important to create a robust, modular design which will allow for small interface portions of code, here called *parsing engines*, to be tailored to different input file descriptions. The next section describes the requirements of the software, and finally its design.

2.3 Requirements of the Application

The Rain Event Parser was designed from the beginning to be an easily accessible utility. First and foremost, the application must correctly parse the rainfall events from the provided rainfall record; this requirement is crucial, because it is the basis by which all of the other functions will perform their duties. The ultimate result of the application is to produce the necessary input parameters for the Exponential and Gamma PDF functions, which will allow researchers to use these values to help design urban stormwater infrastructure elements.

The Rain Event Parser has four main elements: A series of classes which describe rainfall Events, an engine for parsing rain events from supplied rainfall records, a front-end application interface, and a series of utilities to aid in performing statistical analysis on the parsed events. In order to accomplish these design requirements, all of the necessary classes and parsing elements have been developed using the Python programming language (see: <http://www.python.org/>). Python was chosen for its efficient text processing capabilities, and because it is a mature, fully Object Oriented language. The former consideration is obvious, given the nature of the input medium; the latter consideration was made in order to use high-level abstractions to treat all of the input parameters.

3.1 Class Descriptions

The Event class is the fundamental building block by which the rest of the application is built. An Event object captures the hourly rainfall found for a rainfall event from the NCDC record, along with the date and time that the event began. Every entry in the list stored within the Event

object represents one hour of rainfall data. Hours during the event which had no recorded rainfall (but were below the IETD event delimiter) are stored as zero values, to preserve the average intensity calculations. The Event class contains methods which compute the sum of the hourly rainfall volumes for an event, compute the total duration of an event, and calculate the intensity of an event.

The EventContainer class is used to store all of the events which were found within the input file for the desired IETD delimiter. Objects of this type are also used to compute all of the necessary statistics related to the sample. The EventContainer is not, however, responsible for the actual parsing of the input file; this design feature allows the class to be used potentially to capture events recorded in other formats at a later date.

3.2 The Parsing Engine

The RainDataParser class has the task of parsing the actual events from within the NCDC data record. The method for extracting the events from an input file is complicated by several factors: first, dates for which no rainfall was measured are not present in the file; second, records where a “trace” amount of rainfall was recorded are present in the file; and third, events may span across several days, where each day is represented by a different record. None of these issues are, in and of themselves, difficult to overcome, yet it should be noted that extreme caution was taken to ensure that these issues were addressed to preserve proper event recognition within the application.

3.3 The Statistical Analysis

As previously stated, the analytical probabilistic model which is being implemented by this tool needs to deliver the Exponential and Gamma PDF's. The following two sections illustrate the method by which the application generates these functions.

3.3.1 The Exponential PDF

Parameters of the exponential PDF of rainfall volume, duration, average intensity and interevent time are denoted by λ , ζ , β , and ψ *respectively* and the values of these parameters can be obtained by taking the inverse of the average event volume, average duration, average rainfall intensity and average interevent time, respectively. The Exponential PDF is given by:

$$\begin{aligned}
 f_x(x) &= \gamma e^{-\gamma x}, \quad x \geq 0 & f_v(v) &= \zeta e^{-\zeta v}, \quad v \geq 0, \text{ where } \zeta = \frac{1}{v} \quad (\text{mm}^{-1}) \\
 & & f_T(t) &= \lambda e^{-\lambda t}, \quad t \geq 0, \text{ where } \lambda = \frac{1}{t} \quad (h^{-1}) \\
 & \text{and} & & \text{and,} \\
 & & f_I(i) &= \beta e^{-\beta i}, \quad i \geq 0, \text{ where } \beta = \frac{1}{i} \quad (h/\text{mm}) \\
 \gamma &= \frac{1}{\mu_x} = \frac{1}{\sigma_x} & f_B(b) &= \psi e^{-\psi b}, \quad b \geq 0, \text{ where } \psi = \frac{1}{b} \quad (h^{-1})
 \end{aligned}$$

where μ_x is the mean, and σ_x is the standard deviation. In order to return the proper value for the

input parameters, one must simply return the inverse of the average for each parameter to the researcher.

3.3.2 The Gamma PDF

The Gamma PDF is given by:

$$f_x(x) = \frac{x^{\rho-1} e^{-x/\tau}}{\tau^\rho \Gamma(\rho)}$$

where,

$$\rho = \mu_x^2 / \sigma_x^2, \quad \text{and} \quad \tau = \sigma_x^2 / \mu_x$$

and,

$$\Gamma(\rho) = \int_0^{\infty} z^{\rho-1} e^{-z} dz$$

is the Gamma function.

In contrast to the Exponential PDF, the Gamma PDF requires more computation to produce the result. The SciPy Python package (see: <http://scipy.org/>) is used to provide useful statistical support for this project; specifically, the Gamma function defined in the package is used to generate the curve for the graphical representations (described in a subsequent section).

3.4 The User Interface

There are currently two different methods for accessing the utility: a web-based front-end application, and a command-line driven application. Each interface will be described in the following two sections.

3.4.1 A Web-based Interface

Since the goal of the utility is to help relieve some of the difficulty in modeling stormwater management scenarios, it was important to account for the user-friendliness of the application. The initial design considerations for the Rain Event Parser were to either build a Graphical User Interface (GUI) or a web-based interface; the web-based interface eventually took precedence, as it appeals to a broader audience, and requires no additional software to be installed on the researcher's computer. There are two possible avenues one can choose from when designing a web-based application of the type required for the Rain Event Parser: A Common Gateway Interface (CGI) script, or a Web Application Framework. CGI scripts provide a simple interface which communicates directly with a user's web browser. These scripts require very little in the way of special software, yet they make more complex tasks (such as database access) somewhat more complicated. In order to provide a scalable application, it was decided to use a Web Application Framework to build the Rain Event Parser. This application allows the engine of the Rain Event Parser to be run in a browsing session; it handles user input/output routines, builds customized HTML pages, and allows for simple database access (note: a database feature was not part of the original specification, but the chosen framework enables one to be used in the future with a minimum impact).

As a web-based application, the Rain Event Parser is extremely simple to use. Once a researcher acquires a properly formatted file from the NCDC (there is a tutorial document, in Portable Document Format, to aid an individual in requesting the correct information) they may upload the file into the interface, choose the options that they want, and hit the 'Submit Query' button.

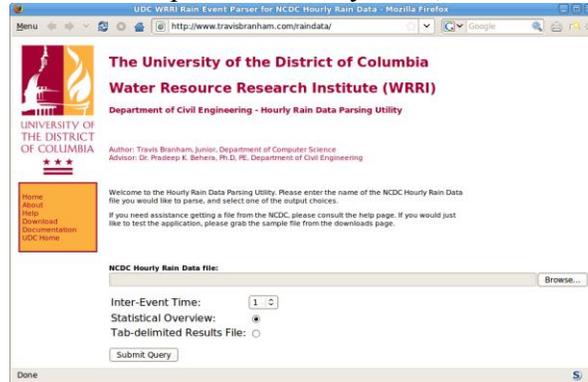


Figure 1: A screenshot of the web-based interface

3.4.2 A Command-line Interface

Prior to developing the web-based front-end, a simple command-line utility wrapper was created around the core parsing engine to enable for testing and debugging of the application. This version has an advantage over the web-based front-end in that it is capable of running directly on the researcher's machine, provided they have the Python Programming Language and the SciPy library installed. Both Python and SciPy are free, open-source software packages, which mean that there are no additional monetary costs associated with running the application in this manner.

In its current incarnation, the command-line interface utility is quite simple: it accepts the name of the NCDC rainfall record file to be parsed at the command line, along with several flags which enable a fine degree of control over the output of the program. Here is an example of the output of the help feature:

```
Options:
  -h, --help                show this help message and exit
  -d DATE_RANGE, --date-range=DATE_RANGE
                            Restrict parsing to this date range:
                            'YYYYMMDD,YYYYMMDD'
  -s, --stats                Print statistical output [default True]
  -q, --no-stats            Do not print statistical output
  -t, --table                Print table output [default False]
  -g, --graph                Print graphs [default False]
  -i IETD, --ietd=IETD      Restrict parsing to this IETD [overrides --all]
  -a, --all                  Parse input file for ALL IETD's
  -r, --header              Input file contains header rows [default True]
  -x, --no-header           Input file does not contain header rows
  -w WINTER, --winter=WINTER
                            Skips parsing for the months provided: 'x,y,z,etc'
```

3.5 Output

Both the command-line and web-based utilities are capable of generating several types of output,

in order to best serve the needs of researchers. Currently, due to limitations on the testing server which the web-based application is running, the graphs which are generated for the parsed events, and the associated plots for the PDF functions, are not being provided. Researchers using the web-based interface will, however, receive the input parameters for the Exponential and Gamma PDF's so that they may plot these functions in a spreadsheet application themselves. The command-line interface utility, however, does generate full plots, as well as text-based output. Both output types are described in detail in the following sections.

3.5.1 Text-based Output

There are two primary text-based output types associated with the web-based Rain Event Parser: the Statistical Overview and the Tab-delimited Results File (suitable for opening in a spreadsheet application). If one chooses the statistical overview, they will be presented with the relevant statistics related to the input file, such as the number of events, the averages and standard deviations for the four main parameters. The tab-delimited results file provides a comprehensive list of all of the events, and associated statistics, which were parsed from the file. The former output method is only beneficial for performing quick calculations on the aggregate: calculations which only require knowing, say, the number of events in the period, or the average amount of rainfall; the latter output method is useful for importing into a spreadsheet application so that the researcher may scrutinize the parsed events, or for building a comprehensive atlas of rainfall parameters.

3.5.2 Graphical Output

As was mentioned previously, the web-based interface does not currently support the generating of graphs, though the feature is forthcoming. The command-line utility, however, is capable of generating sophisticated, print-ready graphs of the rainfall parameter histograms, in addition to plots of the Exponential and Gamma PDF functions. Samples of the typical graph style are shown in the Results section.

The plots are generated with the help of the *Matplotlib* Python library (see: <http://matplotlib.sourceforge.net/>). Further, since the core parsing engine is written in Python, a researcher familiar with mathematical and statistical packages such as Matlab will be able to use the core parsing engine as a module in the iPython interface, which allows for interactive graph manipulation. For example, one can dynamically change the output parameters on the generated graph to zoom into an area of interest, or generate isohyetal maps of the summary statistics using the *Basemap* module addition to the matplotlib library.

4.0 Testing

The core engine of the complete application has gone through several phases of testing to ensure accuracy at all levels. Initially, the program was given extremely short duration data files (2 to 3 months worth of data) which were also parsed by hand in a spreadsheet application. Each test file was parsed (in both ways) for IETD's ranging from 1 hour to 24 hours, and the summary statistics were calculated. The parsing engine proved effective at accurately finding the appropriate number of events from the samples when compared to human parsing.

The application was also tested against output generated by a program written by a group at McMaster University for a specific site in the Washington DC area (Ronald Reagan National Airport). This test was also successful, showing very little difference between the statistics generated from the two applications.

It should be noted, however, that there are still several outstanding issues in generating accurate statistical events. Specifically, there are some compelling reasons to remove certain intervals of time from the historical record; dry summer seasons, or winter snowfall seasons can be problematic when trying to determine the likely rainfall parameters for a given location. The parsing utility allows an individual to select specific months to be removed from a given rainfall record. This functionality comes at a price, however, when analyzing the interevent time parameter. As months are removed from the historical record, the first event immediately following a removal must have a null value stored for the interevent time in order to prevent errors in calculation, since it is clearly impossible to know the actual duration since the true previous event. Care has been taken to ensure that the summary statistics do not factor in these values as zero, which would skew the results; instead, they have been removed from the calculations entirely by being treated as null values.

5.0 Results & Conclusions

Below are some of the statistics generated for two locations in Virginia which display typical characteristics for files processed from the same geographical region:

Station ID	448906	Location	Reagan Aprt.	IETD	2 Hours	Total Events	8347
Duration	(Hours)	Volume	(Hundredths of inches)	Intensity	(Hundredths of inches per hour)	Interevent Time	(Hours)
Average	4.391	Average	28.809	Average	6.362	Average	59.581
Standard Deviation	4.732	Standard Deviation	47.035	Standard Deviation	9.693	Standard Deviation	77.713
Coef. of Variation	1.076	Coef. of Variation	1.633	Coef. of Variation	1.524	Coef. of Variation	1.304
Lambda	0.228	Zeta	0.035	Beta	0.157	Psi	0.017
Rho	0.861	Rho	0.375	Rho	0.431	Rho	0.588
Tau	5.099	Tau	76.792	Tau	14.769	Tau	101.362

Station ID	447201	Location	Richmond Aprt.	IETD	2 Hours	Total Events	8495
Duration	(Hours)	Volume	(Hundredths of inches)	Intensity	(Hundredths of inches per hour)	Interevent Time	(Hours)
Average	4.351	Average	30.922	Average	6.841	Average	58.240
Standard Deviation:	4.788	Standard Deviation:	51.289	Standard Deviation:	10.628	Standard Deviation:	78.024
Coef. of Variation:	1.100	Coef. of Variation:	1.659	Coef. of Variation:	1.554	Coef. of Variation:	1.340
Lambda:	0.230	Zeta:	0.032	Beta:	0.146	Psi:	0.017
Rho:	0.826	Rho:	0.363	Rho:	0.414	Rho:	0.557

Tau:	5.269	Tau:	85.069	Tau:	16.513	Tau:	104.529
-------------	-------	-------------	--------	-------------	--------	-------------	---------

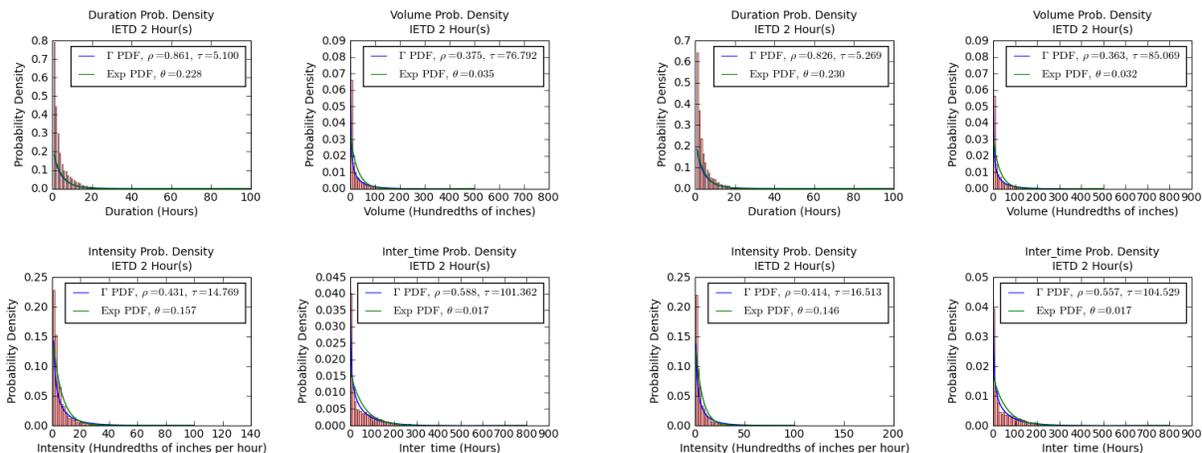


Figure 3: PDF parameters for Ronald Reagan Airport

Figure 2: PDF parameters for Richmond Airport

The results shown above were parsed from files consisting of more than 61 years worth of rainfall data each. The length of the rainfall records used should provide for a very reasonable statistical model of the four major rainfall event parameters. It will be possible to use the output of this application to build a comprehensive atlas of statistical rainfall events which may be used by civil engineers in designing elements of an urban stormwater infrastructure. This type of atlas will be of tremendous benefit to engineers and professionals because it is far less time and labor intensive to use than a more complex continuous modeling method, especially during the early design phase of a project.

6.0 Conclusions and Future Work

The rainfall statistical analysis tool developed in this research will be extremely useful for urban stormwater modeling, especially for analytical probabilistic models. For analytical models, often analysis of rainfall data is cumbersome which is taken care in this project. The engineers and professionals in any part of the nation can utilize the NCDC data to obtain the rainfall statistical parameters. Even though the application described throughout this paper is still in active development, there is already a short, but growing, list of feature requests. Chief among the requested features is the ability to upload files containing data from multiple locations. The NCDC does provide consolidated files of this type already, so providing a mechanism by which these files can be utilized is of top priority. Incorporating this feature, and integrating a simple database of the locations of each station (latitude and longitude) stored by the NCDC, would allow other features, such as the ability to generate maps, to become realistic possibilities. As noted in previous sections, there are still some differences in functionality between the command-line version of the application and the web-based interface. In the near future, these disparities will be rectified, bridging the gap between the two interfaces. It would also be extremely advantageous to bring the full suite of plotting functionality provided by the matplotlib library into the user interface, allowing individuals to have much greater control of their graphical output.

7.0 References

Adams, B. J., and J. B. Bontje, Microcomputer applications of analytical models for urban stormwater management, in *Emerging Computer Techniques in Stormwater and Flood Management*, edited by W. James, 138-162, Am. Soc. Civ. Eng., New York, NY, 1984.

Adams, B. J., H. G. Fraser, C. D. D. Howard, and M. S. Hanafy, Meteorological data analysis for urban drainage system design, *J. Environ. Eng.*, 112 (5), 827-848, 1986.

Adams, B. J., and F. Papa, *Urban Stormwater Management Planning with Analytical Probabilistic Models*, John Wiley and Sons, New York, NY, 2000.

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

Guo Y., and B. J. Adams, Hydrologic analysis of urban catchments with event-based probabilistic models, 1, Runoff volume, *Water Resour. Res.*, 34(12), 3421-3431, 1998.

“Matplotlib v0.99.1.1 documentation,” Accessed: March 7, 2009. <<http://matplotlib.sourceforge.net/>>.

“NCDC: * National Climatic Data Center (NCDC) *,” Accessed: 10, September, 2008. <<http://www.ncdc.noaa.gov/oa/ncdc.html>>.

“Python Documentation Index,” Accessed: November 13, 2008. <<http://python.org/doc/>>.

“SciPy: Scientific Tools for Python,” Accessed: February 19, 2009. <<http://www.scipy.org/>>.

Information Transfer Program Introduction

Our partnership with the Cooperative Extension Service, the outreach unit of our land-grant university has been our major outlet for information transfer. A Water Quality Education Program provides community and school-based workshops on water quality and quantity issues. Our Marketing Specialist promotes land grant programs through website and public relations development. The Institute benefitted immensely with increased visibility of our Water Highlights Newsletter and technical reports, thereby enhancing trust. Our limited resources were spent to design and print the several issues of the new and improved Water Highlights Newsletter. With the very limited resources available, this program has compensated for the information transfer projects which were not proposed nor funded by the Institute.

For the first time in 2010, we received and approved an information transfer project. The proposed information transfer project including the student seminar series, faculty brownbag lunch series, and planning for the Semi-Annual Forums will support the development of a more cohesive water research community for the District of Columbia, bringing together the different university populations and disciplines. These various programs will also provide an opportunity for the DC water stakeholders, including agency personnel and professionals in private industry and consulting practices, to exchange information on issues and research needs, interests, and capabilities, as well as funding programs and hiring opportunities. The programs will be developed concurrently and in a collaborated manner through the DCWRRI, with input through the current Stakeholder Advisory Committee, National Capital Region Section of AWRA, and university water leaders (to be organized into a Universities Water Advisory Council).

The Institute website, <http://www.udc.edu/wrri/>, provides updated information about current activities. The Institute also completes bi-seasonal issues of the Water Highlights Newsletter. These documents are very informative and highlight current research and educational projects sponsored by the Institute along with interactions among faculty members and their student interns on projects and conferences. Please visit the Institute's website at <http://www.udc.edu/wrri/publication.htm#poster> for copies of all newsletters and publications.

An electronic mailing list of over 150 Water Resources faculty and experts in the consortium of universities in Washington DC is maintained and regularly updated and disseminated via email to report updates on local, regional, and national water issues received by the Institute. This line of information transfer has enhanced the visibility and credibility of the Institute among its stakeholders.

USGS Summer Intern Program

None.

Student Support					
Category	Section 104 Base Grant	Section 104 NCGP Award	NIWR-USGS Internship	Supplemental Awards	Total
Undergraduate	6	0	0	0	6
Masters	1	0	0	0	1
Ph.D.	0	0	0	0	0
Post-Doc.	0	0	0	0	0
Total	7	0	0	0	7

Notable Awards and Achievements

For the first time, a follow on project to one of our projects by Dr. Yu and Dr. Behera, Application of Spatiotemporal Informatics to Water Quality, received an NSF follow-on award of \$100,000.

NSF CMMI Grant 0940393: Environmental Urban Runoff Monitoring, received an award amount of \$100,000, PI and Co PIs, Drs. Byunggu Yu, Pradeep Behera, Seon Kim and Paul Cotae Conference presentation by Byunggu, Yu, P. K. Behera, Seon, H. Kim, J.F.R. Rochac, Travis Branham, Environmental Urban Runoff Monitoring Proceedings, SPIE Smart Structures NDE, March 7-11, 2010, San Diego, CA

Conference presentation by Byunggu, Yu, P. K. Behera, Seon, H. Kim, J.F.R. Rochac, Travis Branham, Environmental Urban Runoff Monitoring Paper and poster presentaiton, SPIE Smart Structures NDE, March 7-11, 2010, San Diego, CA.

This is the very first NSF award to perform a research project from UDC. We are very proud of this accomplishment as we continue to strengthen our research capabilities.

Publications from Prior Years