5-2009

Hydrologic Information Systems: Advancing Cyberinfrastructure for Environmental Observatories

Jeffery S. Horsburgh
Utah State University

Recommended Citation

Horsburgh, Jeffery S., "Hydrologic Information Systems: Advancing Cyberinfrastructure for Environmental Observatories" (2009). *All Graduate Theses and Dissertations*. 219.
https://digitalcommons.usu.edu/etd/219

5-1-2009

Hydrologic Information Systems: Advancing Cyberinfrastructure for Environmental Observatories

Jeffery S. Horsburgh
Utah State University

HYDROLOGIC INFORMATION SYSTEMS: ADVANCING CYBERINFRASTRUCTURE FOR ENVIRONMENTAL OBSERVATORIES

by

Jeffery S. Horsburgh

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

Approved:

David G. Tarboton Major Professor	David K. Stevens Committee Member
David R. Maidment Committee Member	Mac McKee Committee Member
Ronald J. Ryel Committee Member	Byron R. Burnham Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2008

ABSTRACT

Hydrologic Information Systems: Advancing Cyberinfrastructure
for Environmental Observatories

by

Jeffery S. Horsburgh, Doctor of Philosophy

Utah State University, 2008

Major Professor: Dr. David G. Tarboton
Department: Civil and Environmental Engineering

Recently, community initiatives have emerged for the establishment of large-scale environmental observatories. Cyberinfrastructure is the backbone upon which these observatories will be built, and scientists' ability to access and use the data collected within observatories to address research questions will depend on the successful implementation of cyberinfrastructure. The research described in this dissertation advances the cyberinfrastructure available for supporting environmental observatories. This has been accomplished through both development of new cyberinfrastructure components as well as through the demonstration and application of existing tools, with a specific focus on point observations data. The cyberinfrastructure that was developed and deployed to support collection, management, analysis, and publication of data generated by an environmental sensor network in the Little Bear River environmental observatory test bed is described, as is the sensor network design and deployment. Results of several analyses that demonstrate how high-frequency data enable identification of trends and analysis of physical, chemical, and biological behavior that would be impossible using traditional, low-frequency monitoring data are presented. This dissertation also illustrates how the cyberinfrastructure components demonstrated in the Little Bear River test bed have been integrated into a data publication system that is now supporting a nationwide network of 11 environmental observatory test bed sites, as well as other research sites within and outside of the United States. Enhancements to the infrastructure for research and education that are enabled by this research are impacting a diverse community, including the national community of researchers involved with prospective Water and Environmental Research Systems (WATERS) Network environmental observatories as well as other observatory efforts, research watersheds, and test beds. The results of this research provide insight into and potential solutions for some of the bottlenecks associated with design and implementation of cyberinfrastructure for observatory support.

(223 pages)

ACKNOWLEDGMENTS

This work is dedicated to my children, Morgan, Kaitlyn, and Haven, who are my never-ending source of inspiration and light. I thank my girls and my wife, Amy, for reminding me every day that there is so much more to life than data and code, and for supporting me through many late nights and tired mornings. I thank my parents, Chuck and Kris Horsburgh, for teaching me the value of work and instilling within me the dedication to finish this degree.

I am indebted to my advisor, Dr. David Tarboton, for his guidance and for his dedication. I consider him a true friend and colleague, and will be forever grateful for his time, his high academic standards, and his commitment to his students. I would also like to thank my committee members, David Stevens and Mac McKee for talking me into doing this in the first place and David Maidment and Ron Ryel for their input and support.

I thank those who were not part of my committee, but who contributed as coauthors on chapters of this dissertation. I would also like to acknowledge the support of fellow graduate students Amber Spackman and Sandra Guerrero and all of those that assisted us in installing data collection infrastructure in the Little Bear River and collecting data. I will never forget our rousing conversations while wading through three feet of snow in below zero temperatures to access our monitoring sites, calibrate sensors, and collect samples. Special thanks also go to the group of computer programmers with whom I have worked over the past several years. Their contributions helped make this work successful.

This research was primarily funded by the National Science Foundation through grants that supported development of the Little Bear River environmental observatory test bed (CBET 0610075 – Tools for Environmental Observatory Design and Implementation: Sensor Networks, Dynamic Bayesian Nutrient Flux Modeling, and Cyberinfrastructure Advancement) and the Consortium of Universities for the Advancement of Hydrologic Science, Inc. Hydrologic Information System (EAR 0622374 – GeoInformatics: Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) Hydrologic Information Systems and EAR 0413265 – Collaborative Research: Development of Informatics Infrastructure for the Hydrologic Sciences, CUAHSI Hydrologic Information System).

Jeffery S. Horsburgh

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

How is current hydrological understanding constrained by the kinds of measurements that have heretofore been available and how can those constraints be loosened by new measurement technologies and new strategies for their deployment? These questions posed by *Kirchner* [2006] are focused on the fact that, despite the growing volume and sophistication of hydrological theorizing over the past several decades, the ultimate source of hydrologic information is field observations and measurements. Indeed, science and engineering research and education have recently become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared and analyzed [*National Science Foundation*, 2007].

Many researchers within the science and engineering research communities have suggested that new data networks, field observations, and field experiments that recognize the spatial and temporal heterogeneity of hydrologic processes will be needed to address complex and encompassing questions and advance the science of hydrology [*Woods et al.*, 2001; *Hart and Martinez*, 2006; *Kirchner*, 2006; *Montgomery et al.*, 2007]. This knowledge that current understanding is constrained by a lack of observations at appropriate spatial and temporal scales has motivated community initiatives (e.g., http://www.cuahsi.org, http://cleaner.ncsa.uiuc.edu, http://www.watersnet.org/) towards the establishment of large-scale environmental observatories, which aim to overcome this limitation through the collection of data at unprecedented spatial and temporal resolution.

To what extent is current understanding constrained by the tools and methods that have heretofore been used to organize, manage, publish, visualize, and analyze data? This question, which is a natural extension to those of Kirchner, is important because as the amount and complexity of data grows, it becomes increasingly difficult, if not impossible, for data analysts to identify trends and relationships in the data and to derive information that enhances understanding using simple query and reporting tools [*Connolly and Begg*, 2005]. Combining multiple lines of evidence (e.g., using data streams from multiple sensors or from multiple sites) into a single analysis becomes much more difficult when they consist of thousands or even tens or hundreds of thousands of observations. Thus, even if the data are available, without the tools to manage and manipulate the data their utility in fostering process understanding is limited.

Additionally, it is difficult for the broader scientific community beyond individuals who collected the data to use them for scientific analyses if they are never published or if semantic and syntactic differences among datasets preclude their use in common analyses. Recently, these questions of data availability, organization, publication, visualization, and analysis have come to the forefront within many scientific communities (e.g., hydrology, environmental engineering, etc.). With advances in observing, computing, and information technology, it is becoming increasingly important and feasible to develop systems and models that answer these questions. Hydrologic Information Systems are emerging as technology to address these questions in the area of Hydrology and Water Resources.

Observatory initiatives will require enormous investments in both capital and in information technology infrastructure to manage and enable the observing systems.

According to the *National Research Council* [2008], advanced information technology infrastructure will be required as a central component in the planning and design of observatories to help manage, understand, and use diverse datasets. Comprehensive infrastructure that is being used to capitalize on advances in information technology has been termed “cyberinfrastructure” and integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools [*National Science Foundation*, 2007].

The focus of the research described in this dissertation is on a single, yet very important, class of water resources data – observational data measured at a point (e.g., time series data collected at a stream monitoring site or weather station located at a fixed point in space). It is hypothesized that current hydrological understanding is constrained not only by the kinds of measurements that have heretofore been available, but also by the methods that have been used to organize, manage, analyze, and publish data. The overall purpose for this research, then, was to test this hypothesis in an environmental observatory setting with a goal of advancing the cyberinfrastructure available for supporting environmental observatories, experimental watersheds, and other observatory efforts.

The research described in this dissertation was accomplished through developing new cyberinfrastructure components as well as through the demonstration, application, and extension of existing tools. The following research objectives were chosen to test the above hypothesis with a particular focus on point observations data:

- *Objective 1: Establish a wireless sensor network for high frequency estimation of water quality constituent fluxes and investigation of the hydrologic and hydrochemical responses within an environmental observatory test bed*[1]. One focus of environmental observatories is creating a better understanding of the spatial and temporal variability in the fluxes and stores of water quality constituents through the use of sensor network technology. The use of water quality measures such as turbidity, which can be measured with high frequency, as surrogates for other water quality constituents that cannot economically be measured with high frequency (e.g., total suspended solids and phosphorus) has been proposed for creating high frequency estimates of constituent concentrations. Other water quality variables such as temperature, dissolved oxygen concentration, pH, and specific conductance measured using in-situ sensors can reveal a wealth of detail in short-term variability in water quantity and quality that is not well captured by conventional monthly, weekly, or even daily grab sampling programs. These high frequency measurements reveal detail that provides information on process physics heretofore inaccessible to measurements. Sensors, dataloggers, and telemetry systems, and the data streams that they produce are important components of the cyberinfrastructure required for establishing environmental observatories and information systems, and understanding how these systems work is important in developing infrastructure to support them.

[1] A test bed is a prototype or development environment used for testing methods prior to large-scale implementation.

- *Objective 2: Design a generic data model for point environmental observations.* Infrastructure will be required for managing the manipulation, storage, and retrieval of the large datasets generated by sensor networks within environmental observatories. A generic model of observational data from a range of water resources disciplines (hydrology, environmental engineering, meteorology, etc.) and accommodating a range of different variables (precipitation, streamflow, water quality) is needed to provide a standard data storage format that enables data discovery, analysis, visualization, and publication. Because observatory datasets will span investigators and domains, overcoming potential syntactic (i.e., differing file types and structures) and semantic heterogeneity (i.e., differing language used to describe data) is also of primary concern. A point observations data model provides a systematic way to store environmental observations and sufficient metadata to facilitate unambiguous interpretation and to promote effective data sharing.
- *Objective 3: Create an integrated observatory information system using cyberinfrastructure for environmental observatories.* Collectively, the components that make up an integrated observatory information system (including the sensor networks and observations databases) must provide the mechanisms for and the technology that enables the collection, storage, discovery, retrieval, visualization, and analysis of all of the observatory data. Additionally, an observatory information system should support the open and free publication

and exchange of the data in a way that achieves integration and interoperability across a network of environmental observatories.

These objectives were chosen to address three very high level categories of cyberinfrastructure functionality required to support environmental observatories: 1) data collection; 2) persistent data storage and management; and 3) data publication. They are focused on the challenges inherent in making the connection between sensors that collect environmental observations and the analysis and modeling applications that use these data to advance scientific understanding. Each of these objectives is addressed within one or more chapters of this dissertation as follows.

Chapter 2 addresses the first objective and presents the development of an environmental observatory test bed within the Little Bear River watershed of northern Utah, USA, which was designed with the overarching goal of improving the observing infrastructure and cyberinfrastructure available for the planning and implementation of environmental observatories. This paper describes the sensor network design, cyberinfrastructure components, and data collection procedures used and provides results from analyses related to creating high-frequency estimates of water quality constituent concentrations from surrogate measures and our investigations of the hydrologic and hydrochemical responses in the Little Bear River watershed using high-frequency data.

Chapter 3 addresses the second research objective and presents the Observations Data Model (ODM), which is a new and consistent format for the storage and retrieval of point environmental observations in a relational database. Within ODM, observations are stored with sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and to provide traceable heritage from raw measurements to useable information. Chapter 3 presents the design principles and features of ODM and illustrates how it can be used to enhance the organization, publication, and analysis of point observations data. ODM represents a new, systematic way to organize and share data that overcomes many of the syntactic and semantic differences between heterogeneous datasets, thereby facilitating an integrated understanding of water resources based on more extensive and fully specified information. ODM is part of the infrastructure required for managing the manipulation, storage, and retrieval of the large datasets generated by sensor networks within environmental observatories.

The third research objective is addressed by Chapters 4 and 5. Chapter 4 presents a new method for publishing research datasets consisting of point observations that employs a standard observations data model populated using controlled vocabularies for environmental and water resources data along with web services for transmitting data to consumers. This paper describes how these components have reduced the syntactic and semantic heterogeneity in the data assembled within a national network of environmental observatory test beds and how this data publication system has been used to create a federated network of consistent research data out of a set of geographically decentralized and autonomous environmental observatory test bed databases. Finally, in Chapter 5 we "put it all together" to present the components that have been created to form an integrated observatory information system for the Little Bear River environmental observatory test bed. The Little Bear River test bed information system demonstrates mechanisms for and technology that enables the storage and archival of all of the test bed data and the open and free distribution of the data via simple to use, Internet-based tools.

There is a fundamental need within the hydrologic and environmental engineering communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. The tools described in this dissertation represent new opportunities for many within the water resources community to approach the management, publication, and analysis of their data systematically, rather than relying on collections of ASCII text or spreadsheet files, thus removing the burden of learning and interpreting diverse data formats from data end users. Enhancements to the infrastructure for research and education that are described in this dissertation impact a diverse community and are valuable for those involved with prospective environmental observatories as well as other observatory efforts, research watersheds, and test beds because they provide insight into and potential solutions for some of the bottlenecks associated with design and implementation of cyberinfrastructure for observatory support.

References

Connolly, T., and C. Begg (2005), *Database Systems A Practical Approach to Design, Implementation, and Management*, 4th ed., 1374 pp., Addison-Wesley, Harlow, U. K.

Hart, J. K., and K. Martinez (2006), Environmental sensor networks: A revolution in earth system science?, *Earth-Science Reviews*, *78*, 177-191, doi:10.1016/j.earscirev.2006.05.001.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.

Montgomery, J. L., T. Harmon, W. Kaiser, A. Sanderson, C. N. Haas, R. Hooper, B. Minsker, J. Schnoor, N. L. Clesceri, W. Graham, and P. Brezonik (2007), The WATERS Network: an integrated environmental observatory network for water research, *Environ. Sci. and Technology*, *41*(19), 6642-6647. (Available at http://pubs.acs.org/subscribe/journals/esthag/41/i19/pdf/100107feature_waters.pdf)

National Research Council (2008), *Integrating Multiscale Observations of U.S. Waters*, Committee on Integrated Observations for Hydrologic and Related Sciences, Water Science and Technology Board, Division on Earth and Life Studies, The National Academies Press, Washington, D. C.

National Science Foundation (2007), Cyberinfrastructure vision for 21st century discovery, National Science Foundation Cyberinfrastructure Council, NSF 07-28. (Available at http://www.nsf.gov/pubs/2007/nsf0728/index.jsp)

Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental design and initial results from the Mahurangi River Variability Experiment: MARVEX, in *Observations and Modelling of Land Surface Hydrological Processes*, edited by V. Lakshmi, J. D. Albertson and J. Schaake, pp. 201-213, Water Resources Monographs, American Geophysical Union, Washington, D. C.

CHAPTER 2

A STUDY OF HIGH FREQUENCY WATER QUALITY OBSERVATIONS IN THE LITTLE BEAR RIVER UTAH, USA[1]

Abstract

Process-based understanding of short and longer-term behavior of catchments is becoming increasingly important as we work to increase our ability to predict hydrologic system response for use in managing limited water resources. The time scale of many important processes is on the order of minutes to hours, not weeks to months, and understanding the linkages between catchment hydrology and hydrochemistry requires measurements on a time scale that is consistent with these processes. These are motivating factors in the recent push toward establishment of large-scale environmental observatories within the hydrologic and environmental engineering communities that has seen the creation of a network of 11 observatory test beds. In this paper we present a study of high frequency water quality observations in the Little Bear River that have served as the basis for establishing the Little Bear River Test Bed (LBRTB) as one of these test beds. The LBRTB was established with the overarching goal of improving understanding of water quality fluxes and loads and the observing infrastructure and cyberinfrastructure needed to quantify these fluxes and loads in an environmental observatory network. We describe our sensor network design, cyberinfrastructure, and data collection procedures and provide results from four separate analyses that demonstrate how the scope and resolution of data generated by sensor networks enable

[1] Coauthored by Jeffery S. Horsburgh, Amber Spackman Jones, David K. Stevens, David G. Tarboton, and Nancy O. Mesner.

identification of trends and analysis of hydrologic and hydrochemical behavior that could not be observed by traditional water quality monitoring or short-term field campaigns. Using high-frequency data, we demonstrate the importance of early spring snowmelt in contributing to annual loads of total phosphorus and total suspended solids, the effect of sampling frequency on estimates of annual loading, the relative magnitudes and timing of baseflow versus quickflow as the dominant flow pathways, and the differences in ecological responses across sites.

2.1. Introduction

As water resource managers are faced with growing pressure on limited water resources, process-based understanding of short and longer-term behavior of catchments is becoming increasingly important. Our ability to predict hydrologic system response is dependent on our understanding of catchment behavior and the interacting processes that drive that response. In relatively small catchments, the time scale of many important hydrologic and hydrochemical processes is on the order of minutes to hours, not weeks to months, and understanding the process linkages between catchment hydrology and stream water chemistry, which is necessary for incorporating these processes into predictive models, requires measurements on a time scale that is consistent with these processes [*Kirchner et al.*, 2004].

Many believe that advancing the science of hydrology will require new measurements and hydrologic measurement techniques, and that data generated by coordinated, extensive field studies will be required to enable these advances [*Woods et al.*, 2001; *Kirchner*, 2006; *Hart and Martinez*, 2006]. This belief is primarily responsible for the recent push toward establishment of large-scale environmental observatories within the hydrologic and environmental engineering communities. The driver behind environmental observatories is that knowledge of the physical, chemical, and biological mechanisms controlling water quantity and quality is limited by lack of observations at the necessary spatial density and temporal frequency needed to infer the controlling processes [*Montgomery et al.*, 2007]. Within observatories, environmental sensor networks have been proposed as part of the cyberinfrastructure that will be required to generate data of both high spatial and temporal frequency and enable scientific discovery. Sensor network technologies offer several advantages over traditional monitoring techniques by streamlining the data collection process, reducing human errors and time delays, reducing overall cost of data collection, and increasing the quantity and quality of data on temporal and spatial scales [*Glasgow et al.*, 2004].

Continuous, high-frequency monitoring records generated using in-situ sensors can reveal detail in short-term variability in water quantity and quality that is not well captured by conventional monthly, weekly, or even daily grab sampling programs [*Jarvie et al.*, 2001; *Tomlinson and De Carlo*, 2003; *Kirchner et al.*, 2004; *Tetzlaff et al.*, 2007]. Continuous records can be critical in capturing and characterizing both regular and transient events and are becoming increasingly common as sensor technology improves. Observable short-term hydrologic and water quality signals include fluctuations in discharge related to precipitation, snowmelt, and agricultural diversions and return flows. Diurnal fluctuations in pH and dissolved oxygen concentration related to in-stream biological activity are evident in many systems [*Chapra*, 1997; *Wang et al.*, 2003; *Mulholland et al.*, 2005]. Spikes in turbidity related to sediment pulses occurring during spring snowmelt and storm events [*Uhrich and Bragg*, 2003; *Stubblefield et al.*, 2007], and changes in specific conductance related to variability in the sources of water that make up streamflow are also commonly observed [*Covino and McGlynn*, 2007; S*tewart et al.*, 2007]. In addition to characterizing short-term variability, high-frequency measurements made over long periods enable us to examine how short-term variability changes across hydrologic regimes and maximizes the chances for serendipitous discoveries [*Kirchner et al.*, 2004].

Despite advances in technology, however, and in some cases because of them, many challenges associated with establishing sensor networks for scientific research remain. Developing and deploying sensor networks can be an onerous task that requires a great deal of expertise, and domain scientists must step outside of their primary knowledge area to gain the skills necessary for designing and deploying field experiments that employ sensor networks [*Szlavecz et al.*, 2006; *Burns et al.*, 2006]. The sheer volume of data generated by sensor networks presents challenges associated with data processing, quality control, archiving, and analysis that are much different than those encountered with more traditional data. Additionally, logistical challenges, such as obtaining site access, hardening deployments against environmental conditions, and overcoming communication limitations, are inherent in sensor network design and deployment [*Lundquist et al.*, 2003]. In many cases, sensor technology does not yet exist to measure important variables, which has driven research into new sensor technologies and the use of existing sensor measurements as surrogates for variables that cannot be measured continuously [*Christensen et al.*, 2002; *Uhrich and Bragg*, 2003; *Stubblefield et al.*, 2007]. If sensor networks are to reach their potential as standard research tools, there is a need to simplify and standardize aspects of the design, setup, configuration, programming, deployment, and maintenance of sensor network components.

In 2006, recognizing the challenges associated with establishing sensor networks, and, on a broader scale, the entire infrastructure to support large-scale environmental observatories, a network of 11 environmental observatory test bed projects was created across the United States. These test beds are part of the WATERS (WATer and Environmental Research Systems) network (http://www.watersnet.org/), and each was selected to demonstrate techniques and technologies that could be used in the design and implementation of a national network of large-scale environmental observatories. Technologies investigated within the test beds range from innovative application of environmental sensors to achieve a better understanding of the stores and fluxes of environmental constituents to development of software components for publishing observations data in common formats that can be accessed by investigators throughout the scientific community [*Minsker et al.*, 2006; *Moore et al.*, 2007; *Welty et al.*, 2007; *Stevens et al.*, 2007; *Fisher et al.*, 2007].

The Little Bear River test bed (LBRTB) was established primarily to test the hypothesis that high-frequency sensor data collected at multiple sites can improve hydrologic and hydrochemical process understanding. We are examining turbidity as a surrogate for concentrations of total suspended solids (TSS) and total phosphorus (TP) to provide a means for better quantifying patterns in constituent fluxes within the watershed. Turbidity can be measured with high-frequency relatively inexpensively, whereas there are currently no reliable continuous in-situ sensors for TP and TSS. We are also examining specific conductance as a tracer that can be measured with high-frequency for investigating flow pathways and dissolved oxygen as an indicator of ecosystem function and dynamic diurnal processes. Secondary research goals within the LBRTB include investigating the effects of sampling frequency on estimates of annual TP and TSS loads and advancing available cyberinfrastructure for storing, archiving, accessing, visualizing, and analyzing observatory data.

In this paper we present findings from our analyses of high-frequency data collected using in-situ sensors to date that include: 1) high-frequency synthetic time series of TSS and TP generated from surrogate turbidity data that reveal concentrated periods of high TSS and TP loading that dominate the annual load and occur primarily during early spring snowmelt; 2) annual TP and TSS load estimates calculated from daily, weekly, and monthly subsets of the high-frequency data that show how annual loads calculated from infrequent samples are only order of magnitude estimates that tend to underestimate the true annual loading in the majority of cases; 3) a two-component hydrograph separation based on specific conductance that shows quickflow (i.e., new water) dominating the spring snowmelt hydrograph and baseflow (i.e., old water) remaining relatively constant throughout the year; and 4) estimates of photosynthesis and respiration rates calculated based on diurnal dissolved oxygen curves that are very different from site to site and provide metrics for comparing instream metabolism.

These examples demonstrate how the scope and resolution of data generated by sensor networks enable identification of trends and analysis of behavior that could not be observed by traditional water quality monitoring or short-term field campaigns. We also discuss how our methods, data collection, and analyses can support the design and implementation of large-scale environmental observatories. It is expected that these analyses will be expanded as the LBRTB datasets mature.

In Section 2.2, we describe the physical setting of the Little Bear River watershed. In Section 2.3 we describe the experimental and sensor network design, data collection procedures, and methods that that have been implemented to support our analyses. We also provide a brief description of the data management and publication procedures and cyberinfrastructure that have been implemented to support the LBRTB. Following these descriptions, in Sections 2.4 and 2.5 we present our results and discuss how the cyberinfrastructure that we have implemented enabled our analyses. Finally, in Section 2.6 we summarize our results.

2.2. Site Description

The Little Bear River in northern Utah, United States (Figure 2.1) drains an area of approximately 740 km^2 and is typical of many semiarid watersheds in the western United States where streamflow is dominated by spring snowmelt and where extensive hydrologic modification for agricultural diversion has taken place. The Little Bear River drains into Cutler Reservoir, a shallow, eutrophic reservoir on the mainstem of the Bear River, which ultimately drains to the Great Salt Lake. The Little Bear River watershed encompasses primarily lower elevation agricultural, mid-elevation range, and higher elevation forested lands. Approximately 70% of the watershed area is grazing land and forest, 19% is irrigated cropland and, 7% is dry cropland. The area is experiencing rapid population growth, with a 32% increase in population between 1990 and 2000 [*U.S. Census Bureau*, 2000].

The headwaters of the Little Bear River are located in the Bear River Mountain Range, which consists, in large part, of a thick sequence of carbonate (limestone and dolomite) rocks that range in age from Cambrian to Mississippian [*Dover*, 1987; *Schaefer et al.*, 2006]. In general, this leads to waters with relatively high and well buffered pH, as well as relatively high specific conductance and dissolved solids concentrations. Elevations in the watershed range from 1,340 m to over 2,700 m. Most of the annual precipitation falls as snow at higher elevations and can exceed 900 mm yr^{-1}, as recorded at the Little Bear River Snowpack Telemetry (SNOTEL) site, with occasional summer storms. Precipitation near the outlet is on the order of 450 mm yr^{-1}, demonstrating the variability in annual precipitation with elevation.

The Little Bear has two principal subdrainages, the East Fork and the South Fork. The South Fork and its major tributary, Davenport Creek, flow northward through forest and range land before the confluence with the East Fork. The East Fork originates in higher elevation, forested land, and flows northwest until it is contained by Porcupine Reservoir, which is used to store water for summer agricultural irrigation. A few miles downstream of Porcupine dam, the East Fork is diverted for irrigation purposes, and for several months of the year, portions of the natural channel are dry. The confluence of the two forks is near the town of Avon, after which the river flows northward through the towns of Paradise and Hyrum. Most of the land adjacent to this stretch of the river is agricultural, including crops and livestock grazing. At Hyrum, the river is contained in Hyrum Reservoir, which is also operated to supply water for irrigation of agricultural areas below the reservoir. Below Hyrum dam, the river flows northwest through lower gradient agricultural land, passing through the towns of Wellsville and Mendon before draining into an arm of Cutler Reservoir.

2.3. Methods

2.3.1. Monitoring Sites

Seven stream monitoring sites have been established along the Little Bear River, two during the summer of 2005 and five more during the summer of 2007. Sites were selected to characterize the major hydrologic conditions in the watershed and to represent the range of land use conditions, with preference given to locations that would provide the most information given our limited resources. In addition to considering hydrology and land use, site selection was dependent on the presence of a bridge or other permanent structure to which the sensors could be mounted, our ability to obtain permission to access the site, our ability to establish a stream cross section suitable for development of a stage-discharge relationship, and our ability to establish communications with the site to retrieve the data. Two sites were located in the unregulated South Fork (Upper South Fork and Lower South Fork), two sites were located where they would be highly influenced by releases from the two reservoirs in the system (East Fork and Wellsville), two sites were located in intermediate locations that would represent the combination of unregulated flows plus reservoir releases (Confluence and Paradise), and the last site was located near the terminus of the river just upstream of Cutler Reservoir (Mendon).

Two continuous weather stations were also installed during the summer of 2007, one near the boundary of the lower watershed and one near the confluence of the East and South Forks. Weather station locations were selected to characterize the upper and lower watershed and were constrained by similar site access and communication limitations. Two USDA NRCS SNOTEL sites provide additional continuous weather and snowpack data for the Little Bear. The Little Bear SNOTEL site is located near the headwaters of the South Fork of the Little Bear at an elevation of approximately 1,994 m, and the Dry Bread Pond SNOTEL site is located in the headwaters of the East Fork at an elevation of approximately 2,545 m. Figure 2.1 shows the location of each of the monitoring sites, which are described in Table 2.1.

2.3.2. Continuous Measurements

At each stream monitoring site, a suite of sensors was permanently installed to provide in-situ discharge and water quality records. Data from each of the stream sensors is recorded electronically at 30-minute resolution, with recorded values representing the average over the 30-minute period. At the two weather station sites, data are collected and recorded electronically at hourly resolution (i.e., hourly average/total values) using tripod mounted sensors. Table 2.2 lists the variables measured at each site, the sensors that are being used, and the manufacturers' reported accuracy and resolution where available.

Continuous discharge is calculated from the stage records according to stage-discharge rating curves that have been developed for each monitoring site. Periodic discharge measurements and water surface elevations are collected at each site for the purpose of establishing and maintaining stage-discharge relationships. Discharge measurements have been made using the area-velocity method [*Buchanan and Somers*, 1969] over a range of different discharges to ensure that the derived relationships are representative of the range of hydrologic conditions at each site. Stream velocities are measured using a Marsh McBirney Flo-Mate Model 2000 velocity meter and depths are measured using a top-setting wading rod.

Stream sensors were installed in the main flow of the river and were enclosed inside PVC pipe housings to protect them from debris and vandalism. The PVC sensor housings were fitted with metal pump screens into which the sensors extend to ensure adequate water flow-through and to protect the sample space around each of the sensors. All sensors are removed and cleaned in the field at least once every two weeks. During each site visit, calibration of the Hydrolab sensors is checked, and recalibration is performed onsite as necessary. The pH sensors are calibrated using both pH 7 and pH 10 buffer solutions, and conductivity sensors are calibrated using a 718 μS cm^{-1} potassium chloride standard. Dissolved oxygen is calibrated to water saturated air using barometric pressure measurements made onsite using a Hydrolab Surveyor (Hach Environmental, Inc.) equipped with a barometric pressure sensor. The turbidity sensors and pressure transducers do not require regular calibration (per the manufacturer’s specifications), although the sensors are checked and cleaned every two weeks along with the Hydrolabs.

The continuous measurements are passed through two levels of quality control. First, the data are plotted and examined for out of range and obviously erroneous data values. Where possible, spurious values are replaced using linear interpolation. In the second level of quality control, data are adjusted for sensor drift using linear drift corrections between the calibration dates as recorded in field notes. All corrections and edits are performed on a copy of the raw data to ensure that the original data are preserved.

2.3.3. Chemistry Sampling

From April 2005 to October 2007, storm event samples and sporadic grab samples from prior studies were available at the Mendon and Paradise sites. Beginning in October of 2007 (at which time in-situ instruments had been installed at all but one site), we began regularly collecting water quality grab samples at all seven sites. Sampling occurs once per week during the spring snowmelt season (March through July) and once every two weeks during the rest of the year. The order in which sites are visited and the day of the week on which sampling occurs are varied in an effort to minimize potential bias due to sampling time of day and day of the week.

In addition to the grab samples, storm event and spring snowmelt event samples have been collected using ISCO 3700 Portable Automated Samplers (Teledyne ISCO, Inc.). These samplers operate by pumping water from the river through tubing into sample bottles held within the sampler, allowing for the collection of multiple samples during an event such as a storm or a period of snowmelt. In general, deployment of the automated samplers has occurred either when precipitation is expected or when a significant snowmelt event is expected.

Phosphorus samples are collected in acid washed 250-mL HDPE bottles, and TSS samples are collected in 500-mL HDPE bottles. Each water quality sample is split for total suspended solids (TSS) and total phosphorus (TP) analysis, with a portion of the sample filtered using a 0.45 μm filter for the analysis of dissolved total phosphorus (DTP). Particulate phosphorus (PP) concentrations are determined by subtracting DTP concentrations from TP concentrations. Laboratory analyses have been performed by labs affiliated with Utah State University and with the State of Utah Division of Water

Quality. For TP and DTP analyses, samples are analyzed using USEPA Method 200.8 (Determination of Trace Elements in Water and Waste by Inductively Coupled Mass Spectroscopy) or using USEPA Method 365.2 (Orthophosphate Ascorbic Acid Manual Single Reagent) preceded by an acid digestion of the sample. The analytical method used depends upon the laboratory performing the analysis. For TSS, samples are analyzed using USEPA Method 340.2 (Total Suspended Solids by Mass Balance) or USEPA Method 160.2 (Residue Nonfilterable Total Suspended Solids). Again, the analytical method used depends on the laboratory performing the analysis. In addition to regular laboratory quality assurance and quality control (QA/QC) procedures, a phosphorus field blank, duplicate, and matrix spike sample are collected at one of the seven sites during each sampling trip, and the site at which QA/QC samples are collected is rotated.

2.3.4. Cyberinfrastructure

The in-situ sensors at each monitoring site are connected to a Campbell Scientific, Inc. datalogger (both CR206 and CR800 dataloggers are used), and the logged data are transmitted via a Campbell Scientific 900-MHz spread spectrum radio telemetry network to the Utah Water Research Laboratory. The data are then automatically loaded into an Observations Data Model (ODM) [see Chapter 3] database using the ODM Streaming Data Loader (http://his.cuahsi.org/odmsdl.html). Laboratory results for water quality samples are entered into the database by hand as they are received from the analytical labs. QA/QC editing to remove obvious errors and correct for instrument drift in the sensor data is performed using the ODM Tools application (http://his.cuahsi.org/odmtools.html) on copies of the raw data series to ensure that the raw data streams are preserved. Derived data series, including discharge and synthetic phosphorus and TSS concentration time series are also stored in the central database to ease data querying, manipulation, and analysis.

The LBRTB data are published using components of the Consortium of Universities for the Advancement of Hydrologic Science, Inc.'s (CUAHSI) Hydrologic Information System (HIS) (http://his.cuahsi.org). Chapter 4 describes details of the HIS data publication system. In short, web services have been implemented on top of the central observations database to provide low-level, programmatic access to the data over the Internet, and the LBRTB website (http://littlebearriver.usu.edu) provides near real time access to the latest observations at each monitoring site as well as data visualization and analysis capability through Internet browser-based interfaces.

2.3.5. Generation of Synthetic Time Series from Surrogate Measures

Despite recent developments in sensor technology, there are still water quality constituents such as phosphorus and TSS that cannot be measured continuously using in-situ sensors. However, many studies have demonstrated the potential for using turbidity as a surrogate for predicting TSS and phosphorus concentrations [*Uhrich and Bragg*, 2003; *Christensen et al.*, 2002; *Stubblefield et al.*, 2007]. At the Mendon and Paradise sites, the period of sensor deployment and sample collection is longer than at the other sites, and approximately 150 grab and storm event samples were available at each site to support calculation of synthetic time series of TP and TSS concentrations using turbidity as a surrogate. Linear regression was used to develop relationships between turbidity and TSS and turbidity and TP for both sites. A number of additional explanatory variables were considered in the regression equations, including discharge, day of the year, hour of the day, whether samples occurred during a storm or not, and whether samples occurred during spring snowmelt versus baseflow conditions. For TP, regression with maximum likelihood estimation (MLE) was performed using techniques described by *Helsel* [2005] to account for censored (i.e., below detection limit) observations. *Spackman Jones et al.* [unpublished data, 2008b] describe in more detail the analyses that were used to derive empirical surrogate relationships for the two sites.

For TSS at both sites, the final regression equations used only turbidity as an explanatory variable. Equation (2.1) shows the model for TSS at the Paradise site, and equation (2.2) shows the model for TSS at the Mendon site:

$$TSS = 3.58 + 1.31 * Turb \tag{2.1}$$

$$TSS = 0.341 + 1.41 * Turb \tag{2.2}$$

where *TSS* is the total suspended solids concentration (mg L^{-1}) and *Turb* is the turbidity (NTU).

For TP, the final regression equations at both sites contained turbidity and an additional categorical variable indicating baseflow versus spring snowmelt conditions. Differentiation between baseflow and snowmelt was done visually by noting the onset and conclusion of the spring snowmelt hydrograph. Additionally, at Mendon the final regression equation contained a variable distinguishing between low (less than 10 NTU) and high (greater than 10 NTU) values of turbidity, which indicates that the relationship between turbidity and TP at Mendon is different at low versus high turbidity. Equation (2.3) gives the model for TP at the Paradise site, and equation (2.4) gives the model for TP at the Mendon site:

$$TP = 0.0209 + 0.000798 * Turb + 0.0386 * Z \tag{2.3}$$

$$TP = -0.0341 + 0.0053 * Turb + 0.0949 * Z - 0.00404 * Turb * Z + 0.0832 * Y - 0.00871 * Y * Turb \tag{2.4}$$

where *TP* is the total phosphorus concentration (mg L^{-1}), *Turb* is the turbidity (NTU), *Z* is a categorical variable for snowmelt (*Z* = 1) versus baseflow (*Z* = 0), and *Y* is a categorical variable for turbidity less than 10 NTU (*Y* = 1) versus turbidity greater than 10 NTU (*Y* = 0). P-values indicating the significance of predictive terms in equations (2.1) – (2.4) were all within the 95% significance level, and the final selected model equations were based on the minimum values of the root mean squared error (RMSE). RMSE values ranged from one third to one half of the means of the observed datasets.

Using the derived relationships, synthetic high-frequency (30-minute resolution) time series of TSS and TP concentrations were calculated from turbidity. The synthetic concentration time series were then used along with the high-frequency discharge data to calculate TSS and TP loads for each half-hour time period within the 2006 and 2007 water years so that we could examine the total loading and temporal patterns in loading for each water year.

2.3.6. Examining Effects of Sampling Frequency on Estimates of Constituent Fluxes

Water quality constituent loadings are commonly determined through collection and analysis of concentration grab samples paired with instantaneous estimates of discharge [*Phillips et al.*, 1999; *Johnes*, 2007]. Several studies have examined how the frequency with which grab samples are collected and the equation used in the calculation affects resulting load estimates [e.g., *Coynel et al.*, 2004; *Johnes*, 2007]. Using the synthetic high-frequency time series of TSS and TP generated at the Paradise site, we investigated the effect of sample frequency on estimates of annual TP and TSS loads. We compared annual load estimates for the 2006 water year at the Paradise site calculated using the high-frequency synthetic time series to annual load estimates calculated from subsets of data created by artificially decimating the synthetic time series. Sub sampling of the synthetic time series was done to simulate hourly, daily, weekly, and monthly sampling frequencies. Excepting the hourly results, sub sampling was done randomly. For example, to simulate daily sampling, we randomly selected one discharge and concentration pair per day for each day of the year and used those values to create an estimate of the annual load. A total of 10,000 annual load estimates were generated for each of the simulated sampling frequencies so that we could examine the resulting distribution of the annual load estimates.

2.3.7. Investigating Hydrologic Pathways and Hydrochemical Response

Assessing water balances, flow paths, and rates is another goal of environmental observatories [*Montgomery et al.*, 2007] that can be supported using continuous high-frequency data. Hydrograph separations based on conservative tracers can be powerful tools for determining contributions to stream discharge from different sources [*Jarvie et al.*, 2001; *McGlynn and McDonnell*, 2003; *Covino and McGlynn*, 2007]. If multiple sources contributing to stream discharge are unique and their signatures are known, end-member mixing analysis can be used to separate the contribution from each source [*Burns et al.*, 2001]. Separation techniques generally use isotope or chemical tracers to define the signatures of each of the end-members. However, laboratory analyses of isotope and chemical tracer concentrations can be expensive, and these constituents cannot be measured with high-frequency over long periods of time. Because of this, many separation studies have focused on individual storm events, leaving longer term catchment behavior uncharacterized.

Our current conceptual model of discharge in the South Fork of the Little Bear is that there is little surface runoff, and that stream discharge is primarily made up of two flow components: 1) slow subsurface flow, or baseflow, which is made up of older water that has a longer residence time in the system; and 2) relatively fast surface and subsurface flows, resulting from spring snowmelt and other storm events throughout the year, which in this paper we refer to as quickflow. Using the high-frequency discharge and specific conductance data collected at the two monitoring sites in the South Fork, we developed continuous, two-component streamflow separations for the two major catchments that make up the South Fork of the Little Bear River (i.e., the Upper South Fork and Davenport Creek). Several previous studies have used specific conductance, which is easily measured with high-frequency using existing sensor technology, as a tracer for hydrograph separation [*Covino and McGlynn*, 2007; *Tetzlaff et al.*, 2007; *Stewart et al.*, 2007]. A two-component separation of the form given in equations (2.5) – (2.7) [e.g., *Pinder and Jones*, 1969; *Jarvie et al.*, 2001; *Stewart et al.*, 2007; *Covino and McGlynn*, 2007] was used to quantify the contribution to stream discharge from two end members:

$$Q_t = Q_1 + Q_2 \tag{2.5}$$

$$\frac{Q_1}{Q_t} = \frac{(C_t - C_2)}{(C_1 - C_2)} \tag{2.6}$$

$$\frac{Q_2}{Q_t} = \frac{(C_t - C_1)}{(C_2 - C_1)} \tag{2.7}$$

where Q_t is the total discharge of the two components, Q_1 and Q_2 represent the discharge of each of the two components, C_t is the tracer concentration within the combined flow (in this case the tracer is specific conductance), and C_1 and C_2 are the tracer concentrations in each of the two flow components. These equations can be solved simultaneously to get the contribution to the total stream discharge from each source.

We were unable to monitor Davenport Creek directly. Instead, continuous time series of discharge and specific conductance were calculated for Davenport Creek (using equations (2.5) – (2.7)) as the difference between the Upper and Lower South Fork monitoring sites since these sites are located just above and below the confluence of Davenport Creek and the South Fork. We then separated stream discharge from the Upper South Fork and Davenport Creek catchments into baseflow and quickflow. Since no direct measurements of baseflow or quickflow conductivities have been made, we adopted the conductivity mass balance method of *Stewart et al.* [2007] and *Jarvie et al.* [2001], which infers the end members from measurements made in the stream. For each catchment, we assigned the baseflow conductivity end member to be equal to the maximum streamflow conductivity, which occurs during the lowest flows (i.e., during the period when stream discharge is made up entirely of baseflow), and the quickflow conductivity end member to be equal to the minimum streamflow conductivity, which occurs during the highest flows (during the period when stream discharge is made up almost entirely of quickflow). End member concentrations were assumed to be constant.

The continuous specific conductance and discharge records for each catchment, along with the derived end members, were then used to calculate the contributions of baseflow and quickflow to stream discharge for the period of record using equations (2.5) – (2.7).

2.3.8. Investigating Ecological Responses

Dissolved oxygen (DO) can be used as an indicator of the general health of a water body and can be used to estimate community metabolism of a stream in terms of gross photosynthesis and respiration rates [*Wang et al.*, 2003]. Generally speaking, DO fluctuations that are near saturation with diurnal variation that is due to temperature and metabolism are characteristic of healthy waters, whereas marked depression of DO below saturation indicates that a stream has been impacted by excess nutrients. Although DO concentrations are controlled by complex physical, chemical, and biological processes, there are three primary processes that contribute to DO dynamics. The first is air-water exchange, or reaeration, which regulates DO to its saturation concentration through exchange with the atmosphere, the second is photosynthesis, which is the process by which plants produce oxygen during the day, and the third is respiration, which is the process by which plants consume oxygen during the night. These three mechanisms can be applied in a mass balance model of the following form:

$$\frac{dC}{dt} = k_a(C_s - C) + P(t) - R \tag{2.8}$$

where C is the DO concentration (mg L^{-1}), t is the time (day), C_s is the saturation DO concentration (mg L^{-1}), k_a is the reaeration rate constant (day^{-1}), $P(t)$ is the photosynthesis rate (mg L^{-1} day^{-1}), and R is the respiration rate (mg L^{-1} day^{-1}). This model assumes that the dissolved oxygen deficit ($C_s - C$) does not vary spatially ($\partial C/\partial x \cong 0$, where x is longitudinal distance). Reaeration is controlled by the physical characteristics of the stream (i.e., surface area, depth, velocity, turbulence, and temperature). Photosynthesis and respiration, however, are biological processes that can be influenced by land use and related pollutant loading and can be important indicators of ecological disturbance [*Mulholland et al.*, 2005].

Using equation (2.8) and the Extreme Value Method (EVM) of *Wang et al.* [2003], we calculated average photosynthesis and respiration rates at four sites (Lower South Fork, Paradise, Wellsville, and Mendon) for a one week period at the beginning of July 2008. The EVM assumes that the change in DO concentration (*dC/dt*) is equal to zero at the minimum and maximum values of the DO diurnal curve and uses these extreme points to estimate the respiration and photosynthesis rates respectively. At the minimum DO concentration, which typically occurs at night or early morning when there is no photosynthesis (*P(t)* = 0), equation (2.8) simplifies to:

$$R = k_a\left(C_{s,min} - C_{min}\right) \tag{2.9}$$

where C_{min} is the minimum DO concentration (mg L^{-1}) and $C_{s,min}$ is the saturation DO concentration corresponding to the temperature at C_{min} in the diurnal curve (mg L^{-1}). At the maximum DO concentration, which generally occurs during the early afternoon, equation (8) simplifies to:

$$P(t_{maxC}) = R - k_a\left(C_{s,max} - C_{max}\right) \tag{2.10}$$

where $P(t_{maxC})$ is the photosynthesis rate (mg L^{-1} day^{-1}) at the time of the maximum DO concentration and $C_{s,max}$ is the saturation DO concentration corresponding to the temperature at C_{max} in the diurnal curve (mg L^{-1}).

Photosynthesis as a function of time was approximated as a half sine wave during daylight hours and zero at night [*Chapra*, 1997]:

$$P(t) = P_{max} \sin\left(\frac{\pi t}{f}\right), \quad 0 \leq tf \tag{2.11}$$

$$P(t) = 0, \qquad f \leq t \leq \tau$$

where P_{max} is the maximum photosynthesis rate (mg L^{-1} day^{-1}), f is the photo-period (hr), τ is the diurnal period (24 hr), and t is measured starting at sunrise. The maximum photosynthesis rate was calculated using equation (2.11) where $P(t) = P(t_{maxC})$ and $t = t_{maxC}$:

$$P_{max} = \frac{P(t_{maxC})}{\sin(\pi t_{maxC}/f)} \tag{2.12}$$

Since solar noon occurs at 0.5f, t_{maxC} was calculated as:

$$t_{maxC} = \Delta t + 0.5f \tag{2.13}$$

where Δt is the time shift of the maximum DO concentration from the solar noon (hr). Finally, the average photosynthesis rate was estimated from the maximum value as:

$$P_{ave} = P_{max}\left(\frac{2f}{\pi\tau}\right) \tag{2.14}$$

where P_{ave} is the average photosynthesis rate (mg L^{-1} day^{-1}).

Using the EVM, average photosynthesis and respiration rates were calculated at each site for each of the days and then all of the days were averaged to estimate the overall average rates at each site for the entire period. Reaeration rate constants (k_a) were estimated for each site using empirical methods presented by *Chapra* [1997] that are based on stream depth and velocity. Saturation DO concentrations were also calculated using equations provided by *Chapra* [1997] based on water temperature and elevation.

2.4. Results

2.4.1. Synthetic Time Series Generated from Surrogate Measures

Figure 2.2 shows discharge and synthetic high-frequency time series of derived TSS and TP at the Paradise site for water years 2006 and 2007. During both years, predicted concentrations of TP and TSS associated with early spring snowmelt events were very high, exceeding 1,500 mg L^{-1} for TSS and 1 mg L^{-1} for TP, and daily fluctuations that were highly dependent on discharge were as high as 1 mg L^{-1} for TP and 2,000 mg L^{-1} for TSS. Predicted concentrations tapered off through the remainder of the snowmelt period and were very low during the summer and winter baseflow periods except for a few spikes related to storm events. Similar timing was observed during both years; however, 2007 was a low water year in the Little Bear and the magnitude and duration of elevated spring snowmelt concentrations was lower during 2007.

The annual TP and TSS load estimates based on the high-frequency synthetic time series were vastly different for the two water years at Paradise. In 2006, the estimated annual TSS load was approximately 1.1 X 10^7 kg and the TP load was approximately 1.2 X 10^4 kg, whereas in 2007 the annual TSS load was approximately 1.8 X 10^6 kg and the TP load was approximately 3 X 10^3 kg. Figure 2.3 shows the estimated cumulative percent of annual discharge and the total annual TSS and TP loads as a function of time for the two water years. For both water years, and for both TSS and TP, the first 3 months of the water year and the last 4 contribute less than 10% of the total annual load each, which means that approximately 80% of the annual loading at this site occurs during only 5 months of the year. A single event that spanned several days during

January of 2006 contributed approximately 5% of the total annual TP and TSS loads, demonstrating the importance of individual events, but the vast majority of the annual loading in all cases was associated with the period of spring snowmelt and, in particular, the beginning of the spring snowmelt period. Figure 2.4 shows discharge and 30-minute TSS loads for the 2006 water year and highlights the early spring loading. In 2006, approximately 60 – 65% of the annual TP and TSS load occurred over a period of approximately 2 – 3 weeks. Figure 2.3 also shows that in general, a greater percentage of the annual loads occurred earlier in 2007 than in 2006, although the last 5 – 6 months of the water years were similar on a percentage loading basis. The divergence between the cumulative TSS and TP loading during the snowmelt period (Figure 2.3) is due to the categorical variable in the TP model, which switches the relationship between turbidity and TP during the snowmelt period and is not present in the TSS model.

2.4.2. Effects of Sampling Frequency on Estimates of Constituent Fluxes

Figure 2.5, which shows synthetic TSS concentrations for the period between February and June of 2006 at the Paradise site, illustrates how much information is lost as sample frequency drops from half hourly (based on the high-frequency data) to weekly and monthly (based on random subsets of the continuous data), which are common sampling frequencies used in traditional monitoring programs. These results illustrate how weekly and monthly samples miss nearly all of the system dynamics and even daily samples fail to characterize the variability in TSS concentrations which, in this example, is primarily driven by the daily snowmelt cycle during spring conditions. Similar results have been generated for TP.

In Figure 2.6, annual loads at the Paradise site calculated using the entire synthetic time series (half-hourly resolution) are compared to annual load estimates created by sub sampling from the half-hourly data at hourly, daily, weekly, and monthly time scales. Across the sites and variables at which this analysis was completed there was relatively little difference between the half-hourly and hourly results, indicating that little resolution would be lost by sampling hourly. However, resolution was lost at the daily, weekly, and monthly time scales, and annual load estimates generated by random sub sampling at these time scales were often several times greater or less than the half-hourly estimates. *Spackman Jones et al.* [unpublished data, 2008a] provide a more in depth analysis of the effects of sampling frequency on TP and TSS load estimates for the Little Bear that considers additional factors such as the hour of the day on which sampling occurs and the day of the week.

2.4.3. Source Water Contributions

The hydrochemical data collected at the two monitoring sites in the South Fork of the Little Bear (and those calculated for Davenport Creek) show a distinct difference in the specific conductance of baseflow conditions versus spring snowmelt conditions (Figure 2.7). In general, specific conductance is inversely related to discharge, and the patterns in specific conductance are similar at both monitoring sites and for Davenport Creek. Conductivity is high during baseflow conditions and is on the order of approximately 400 μS cm^{-1}. As discharge increases with spring snowmelt, conductivity decreases to less than half of baseflow conductivity as the stream water becomes diluted with snowmelt. This pattern is most pronounced at the Upper South Fork site, where conductivity decreases from greater than 400 µS cm^{-1} under baseflow conditions to a minimum of 114 µS cm^{-1} during one of the spring discharge peaks. Figure 2.8 shows conductivity plotted versus discharge for the Upper South Fork and Davenport Creek. The relatively consistent 1:1 relationship between discharge and conductivity in these figures indicates that this relationship has little hysteresis or seasonal dependence. Low flow conductivities are similar in both catchments, while high flow conductivities approach a minimum value that is a little different in each catchment (~100 µS cm^{-1} in the Upper South Fork and ~150 µS cm^{-1} in Davenport Creek).

Figure 2.9 shows the contributions of baseflow and quickflow in the Upper South Fork and Davenport Creek catchments resulting from the separation analysis. In this figure, precipitation and snow water equivalent data are from the Little Bear SNOTEL site. Over the period between November 1, 2007 and July 31, 2008, baseflow accounted for approximately 43% of the total discharge in the Upper South Fork catchment, and quickflow contributed approximately 57%. Within the Davenport Creek catchment, the total discharge for the same period was made up of approximately 37% baseflow and 63% quickflow. The greater contribution of quickflow in the Davenport Creek catchment is due to two later peaks in the quickflow hydrograph that occurred in mid May to early June in Davenport Creek but not in the Upper South Fork. Based on the precipitation data from the Little Bear River SNOTEL site, it appears that these two peaks are related to precipitation events. The snow water equivalent data indicate that the snow was gone in the Upper South Fork catchment at the time of these precipitation events, which explains the lack of observed response in the quickflow hydrograph for the Upper South Fork. However, the Davenport Creek catchment incorporates some higher elevation areas, and it appears that there may have been a rapid melt of remaining high elevation snow caused by these two precipitation events. Observations from nearby SNOTEL sites support this. The Ben Lomond Peak SNOTEL site at 2,438 m elevation and located southwest of the Little Bear SNOTEL site maintained snow well into June, and the Dry Bread Pond SNOTEL site at 2,545 m elevation did not melt out until the beginning of June indicating that there was likely still snow in the upper portions of the Davenport Creek catchment when these precipitation events occurred.

2.4.4. Diurnal Patterns in Hydrochemical Response

Diurnal variability in discharge and specific conductance at the Upper South Fork monitoring site is shown in Figure 2.10. Panel (a) shows the month of April 2008 and demonstrates diurnal patterns in specific conductance that occur during snowmelt. Discharge peaks occur during the late afternoon and early evening near the end of the snowmelt period each day, and the troughs in the daily discharge cycle occur in the early morning around sunrise when air temperatures are coldest. Observed daily fluctuations in discharge during the snowmelt period were as large as 7 m^3 s^{-1}, but were generally on the order of less than 4.2 m^3 s^{-1} depending on the weather conditions. During the snowmelt period, conductivity behaved exactly opposite to discharge. Conductivity peaks occur during the early morning when snowmelt is minimum, and daily troughs in conductivity occur simultaneously with the discharge peaks, with daily fluctuations in conductivity of 30 – 60 μS cm^{-1}.

Panel (b) of Figure 2.10 shows conductivity and discharge at the Upper South Fork site during the month of July 2008, which is within the period of baseflow recession. Air temperatures were hot during this period, there was no snowmelt, and very little precipitation occurred, indicating that all of the flow in the stream is from subsurface sources. Much smaller and more uniform diurnal fluctuations in discharge (on the order of approximately 0.03 $m^3 s^{-1}$ per day) and conductivity (approximately 15 – 20 $\mu S cm^{-1}$) were observed during this period. Maximum conductivity values occur near or after midnight (approximately 11:00 PM – 3:00 AM), and minimum values occur during the afternoon (approximately 1:00 PM – 5:00 PM). Daily discharge peaks in the morning (8:30 AM – 12:30 PM), and daily minimum discharge values occur at night, just before maximum conductivity values (9:00 PM – 12:30 AM). The timing of these diurnal fluctuations indicates a time lag between discharge and conductivity.

2.4.5. Ecological Responses

Figure 2.11 shows DO concentrations and dissolved oxygen deficits at four of the seven stream monitoring sites during the first week of July 2008. The Lower South Fork and Paradise sites, which are located in the upper portion of the watershed, exhibit DO concentrations that are almost always near or above saturation concentrations, whereas the Wellsville and Mendon sites, which are located in the lower watershed and are influenced by higher density agricultural areas, exhibit DO concentrations that are primarily below saturation.

Table 2.3 shows that there are large differences between the respiration and photosynthesis rates among the four sites. Photosynthesis and respiration rates are low at the Lower South Fork site, where we have observed relatively little periphyton growth and where there is little influence from agricultural lands. At the Paradise and Wellsville sites, our observations from the field are consistent with the much higher photosynthesis and respiration rates shown in Table 2.3. During July, the water is clear and periphyton are dense, especially at Wellsville where they sometimes fill the channel. At Mendon, the rates are much lower and may be limited by water clarity (average turbidity during these days at Mendon was 46 NTU, which is high compared to 6.4 NTU at Paradise and 1.2 NTU at Wellsville).

A closer inspection of the diurnal curves revealed that three out of the four sites have similar timing and follow the assumptions of the conceptual model described above. At Mendon, Wellsville, and Paradise, DO concentrations are lowest during the night or early morning when there is no photosynthesis and are highest during the early afternoon when solar radiation and photosynthesis are greatest. However, the Lower South Fork site does not follow this pattern. Figure 2.12 shows a close-up view of the diurnal curves for all four sites on July 5, 2008. DO at the Upper South Fork site peaks at 9:30 AM MST and is lowest at 7:30 PM MST. It appears that since the photosynthesis and respiration rates are relatively low at this site, DO concentrations are driven much more by diurnal temperature fluctuations than instream metabolism. The EVM estimate of the respiration rate (and the photosynthesis rate, which is calculated from the respiration rate) may be subject to error because the minimum DO occurs during the photo-period, when photosynthesis is likely not equal to zero.

2.5. Discussion

The need for high-frequency data is already well established [*Jarvie et al.*, 2001; *Kirchner et al.*, 2004; *Tetzlaff et al.*, 2007]. *Kirchner et al.* [2004] liken trying to infer hydrochemical functioning of a catchment using weekly or monthly grab samples to trying to understand a Beethoven symphony by hearing one note every minute or two. In the following sections, we discuss the value of high-frequency data and provide specific examples of how it has assisted us in evaluating dynamic catchment behavior.

2.5.1. Estimating Constituent Fluxes

Our loading analyses show that TP and TSS loads estimated using weekly or monthly sampling, which are frequencies widely used for assessing mass balances of water quality constituents, for calibrating dynamic water quality models, for assessing compliance with water quality standards, and for measuring trends are, at best, order of magnitude estimates of the true annual loading and tend to, in the majority of simulations, underpredict the true annual load when compared to loads calculated from the half-hourly synthetic data. There was even significant spread in annual load estimates from daily sampling. Because the distributions of discharge, TSS, and TP concentrations are skewed low (i.e., high discharge and concentrations only happen a small portion of the time), any one random set of weekly or monthly samples has a high probability of sampling only lower flows and concentrations, and thus the probability is high that the annual load estimated from the sample set will underestimate the true load. The means of the collections of 10,000 annual load estimates from daily, weekly, and monthly sub sampling were actually very similar to the annual load calculated using the half-hourly data; however, for both TP and TSS at Paradise approximately 53% of the annual load estimates calculated from random daily subsets were less than the mean of all of the annual load estimates from random daily subsets. This number was approximately 68% for random weekly subsets, and approximately 77% for random monthly subsets.

TSS loads estimated from the high-frequency synthetic time series were an order of magnitude greater in 2006 than they were in 2007, and TP loads in 2006 were nearly 4 times greater than those in 2007. These differences demonstrate that year to year load variability is significant, that it is highly influenced by differences in discharge, and that characterizing multiple water years is important in understanding how watersheds behave. We also found that more than half of the annual loading of TP and TSS for both years occurred during a 2-week to 1-month long time window. Cumulative plots of loading and discharge over the two water years illustrate the timing of the TSS and TP loads and show that they do not simply follow the same timing as the discharge. The period of early spring snowmelt is critically important to TP and TSS loading in the Little Bear River, which is likely representative of many snowmelt driven watersheds in the western United States. Traditional grab sampling programs using a weekly or bi-weekly sample frequency would get one to two samples during this period, and monthly sampling might miss it entirely.

The observations made above demonstrate the type of information that can be extracted from high-frequency data. The implications of this type of information are far reaching in the water quality community where low frequency data are routinely used to estimate mass balances for water quality constituents under USEPA’s Total Maximum Daily Load (TMDL) program. Significant overestimation of loads would result in required load reductions that are too strict, an error that could have multi-million dollar consequences for point sources of pollution whose discharge permits are tied to TMDL load reductions. Conversely, underestimation of loads may result in required load reductions that do not fully restore water quality and are not protective of the environment.

In the absence of in-situ sensors for phosphorus and suspended solids, the methods that we have employed in the LBRTB hold much promise for application in environmental observatories for providing relatively inexpensive, high-frequency estimates of TP and TSS concentrations, especially since large-scale environmental observatories will require estimates such as these at many locations and over long time periods to characterize the spatial and temporal variability in water quality constituent fluxes. To recreate the 2-year long time series shown in Figure 2.2 for the Paradise site using grab samples, the cost of sample analytical costs alone would exceed \$500,000 (estimated using our current analytical costs for TP and TSS analysis), and the logistics of collecting, processing, and analyzing samples of this frequency over an extended time period would be impossible. We estimate that the total cost of developing the time series shown in Figure 2.2 using surrogate sampling was on the order of approximately \$50,000, which includes the monitoring equipment, field work, sample analytical costs, and analysis time to develop the surrogate relationships.

2.5.2. Investigating Hydrologic Pathways and Hydrochemical Response

The conceptual model of discharge in the South Fork of the Little Bear River that we tested using the two-component separation is that stream discharge is made up predominantly of subsurface baseflow and quickflow from snowmelt that includes some surface runoff. The observed difference in conductivity between the portion of the hydrograph dominated by baseflow and the portion dominated by spring snowmelt (i.e., quickflow) is consistent with this model. Diurnal discharge and conductivity data during the spring snowmelt period also seem to be consistent with this two-component model. As low conductivity quickflow associated with snowmelt increases during the day, conductivity in the stream decreases.

An additional line of evidence is that TSS and TP concentrations and loads at Paradise are highest during the beginning of the spring hydrograph. In general, these constituents do not move via subsurface pathways, so the fact that spikes in TSS and TP concentrations occur suggests that some surface runoff occurs early in the spring when snow close to active streams is melting, carrying high surface runoff loads of TSS and TP to the stream. This is likely augmented by mobilization of sediment from the stream banks and bed, which happens more during the rising limb of the hydrograph. As snowmelt progresses, it is likely that three things happen: 1) sediment stored within the channel is washed through the system by higher flows; 2) the flow pathway delivering water to the stream increasingly switches from surface to subsurface as snowmelt moves further from active streams, effectively eliminating the pathway carrying TSS and TP to the stream; and 3) snowmelt moves from the predominantly agricultural lowland areas that are close to active streams to upland areas where available sources of TSS and TP are reduced.

The hydrograph separation results show that the baseflow component is relatively constant throughout the year and that the baseflow does not extend into the peaks of the spring snowmelt hydrograph. This is somewhat at odds with some previous isotopic studies elsewhere that have shown a preponderance of “old” water in hydrograph peaks

[*McDonnell*, 1990; *Shanley et al.*, 2002; *Kirchner*, 2003], although these studies are generally done on an individual event basis and not over long periods of time. The observed decrease in specific conductance with increased discharge during the spring snowmelt hydrograph means that newer water from lower conductivity snowmelt is predominating in the stream, essentially diluting the baseflow, and that quickflow exhibits a chemical signature that is different from baseflow and likely results from a relatively short contact time with the soil when compared to baseflow, which is likely from a deeper flow pathway.

The period of baseflow recession presents a challenge for the two component model. During a period where there is no snowmelt and very little precipitation, conductivity is slowly increasing as discharge is slowly decreasing, with superimposed diurnal fluctuations in both. The overall trend suggests that the watershed is drying as the remainder of the quickflow component leaves the system. However, the diurnal fluctuations in discharge and specific conductance that are superimposed on the overall trend are not explained by the model. Although these diurnal fluctuations appear to be inversely related (i.e., peaks in discharge generally line up with troughs in specific conductance), there is a time lag that offsets the curves, with conductance peaks lagging discharge troughs by a few hours, perhaps reflecting the difference in velocity of flow fluctuations that travel with a wave celerity compared to conductance that travels with water velocity.

Several other studies have attributed diurnal patterns in discharge and specific conductance during summer low flow periods to the effects of water use by vegetation and instream photosynthesis and respiration [*Bond et al.*, 2002; *Wondzell et al.*, 2007;

Tetzlaff et al., 2007]. *Tetzlaff et al.* [2007] suggest that diurnal fluctuations involve increased capillary tensions in riparian groundwater arising from high rates of potential evapotranspiration restricting seepage during the day when transpiration rates are highest. *Wondzell et al.* [2007] examined the time lag between maximum estimated evapotranspiration and minimum discharge and attributed changes in the amplitude and time lag of the peaks over time to changes in flow velocity in the stream that affect the rate at which the effects of evapotranspiration are propagated through a catchment. *Bond et al.* [2002] conceptualize that changes in the timing and amplitude of the peaks that occur as summer progresses are related to a transition of streamflow to deeper flow paths with less vegetative water use from shallow flow paths. If we assume that the fluctuations we have observed are driven by evapotranspiration that peaks around midday, then the wave travel time from the effective location where evapotranspiration is impacting discharge to the monitoring site would need to be about 10 hours, as we observe troughs in discharge around 10:00 PM. Evapotranspiration that removes water from the soil layers may increase specific conductance either by reducing dilution of the higher conductance baseflow or by not appreciably taking up constituents that contribute to conductivity. This effect should cause a peak in the specific conductance from evapotranspiration. The observed lag of about 14 hours from midday to the conductance peak (which usually occurs around 12:00 AM to 2:00 AM) would be consistent with a water velocity that is smaller than flow wave celerity.

The differences in diurnal behavior of discharge and specific conductance during the snowmelt period versus the baseflow recession period are somewhat of a serendipitous discovery. However, they also demonstrate an important limitation of hydrograph separation studies based on relatively infrequent isotope or chemical tracer samples that do not consider diurnal variability. Specific conductance is arguably not the best conservative tracer, but it can be measured in-situ with high-frequency and can provide an important line of evidence in investigating hydrologic pathways and hydrochemical response. Additionally, even though the diurnal variations in discharge and specific conductance observed during the baseflow recession period are relatively small when compared to the snowmelt period, they are still interesting and illustrative of how high frequency measurements provide opportunities for studying hydrologic processes and for connecting with other disciplines in studying potential linkages between hydrology and riparian and instream biological processes.

2.5.3. Investigating Ecological Response

The processes controlling dissolved oxygen concentrations are inherently diurnal in nature. The analysis that we performed to estimate photosynthesis and respiration rates would not have been possible without observations of DO concentrations that characterize the entire diurnal DO curve. The DO deficits and rates derived from the high-frequency data are useful indicators of stream metabolism. Our results show that there are large differences in these rates at each site, and we are now investigating the degree to which they are useful in evaluating the effects of human disturbances at the catchment scale (i.e., why are metabolism rates higher at Paradise and Wellsville than at Mendon and the Lower South Fork site?). Although our analysis was limited to a brief period during critical summer low flow and high water temperatures, high-frequency data collected over long time periods also enable estimation of how photosynthesis and respiration rates change seasonally and in response to human disturbances such as agricultural diversions, reservoir releases, and agricultural return flows. Additionally, we have identified one out of four monitoring sites where the most basic assumptions of the EVM conceptual model are not met. It is anticipated that this will happen often within environmental observatories and that insights from high-frequency data will drive development of the next generation of hydrologic and water quality models.

2.5.4. The Supporting Role of Cyberinfrastructure

The cyberinfrastructure that we have implemented within the LBRTB provides an end-to-end system for collecting, managing, analyzing, and publishing observational data. The analyses presented in this paper made extensive use of this system. First, without the sensor network and the high-frequency data that it has produced, none of these analyses would have been possible. The communication system enables us to retrieve data in a timely manner, and it also enables us to monitor the status of the system in real time, which is important in identifying and responding to malfunctions within the sensor network to avoid data gaps.

Organization of the data within a central ODM database was perhaps the most critical step, with several important implications. First, the seamless, automated linkage between sensors and database reduces errors in transcription of the datalogger files, ensures the integrity of the raw data streams, and ensures that data are organized and tagged with appropriate metadata. Second, ODM and the ODM Tools application enable us to manage data versioning, which is important in preserving raw sensor data streams and creating quality controlled versions of the data for use in our analyses. Third, implementation of ODM within a Relational Database Management System (RDBMS) enabled us to use Structured Query Language (SQL) to manipulate and subset data through coded queries. This was important in correctly matching and retrieving subsets of data. For some of our analyses, we were able to write code that directly interfaced with the database to retrieve data in a structured way that eliminated the need for intermediate data processing steps, saving time and eliminating potential data manipulation errors. Finally, publication of the data using the CUAHSI HIS data publication system ensures that the LBRTB data are publicly available and can be used by other investigators to support additional analyses.

2.5.5. Where to Go From Here?

Our study of high-frequency water quality data collected in the Little Bear has informed our conceptual model of the behavior of the Little Bear River watershed, but it has also raised questions that we did not anticipate at the outset and that warrant further investigation. What is the role of vegetation in the timing and magnitude of diurnal fluctuations in specific conductance and discharge during the period of baseflow recession? Why do high flow specific conductance values differ between the Upper South Fork and Davenport Creek catchments? Why do the dissolved oxygen data at the Lower South Fork Site not follow the conceptual model when the other sites we examined do? These questions may be important, especially in linking understanding of hydrologic processes with ecological responses.

Other, more practical questions related to the use of surrogate relationships for environmental observatory design and implementation have also emerged. How many grab samples are really needed to establish surrogate relationships between turbidity and TSS and TP, do the relationships change over time, how often do we need to sample to maintain the relationships, and when should the samples be collected to gain the most information? These questions aim at how to best quantify fluxes given the technology that we currently have while minimizing costs and achieving acceptable accuracy. While we estimated above the large (and unrealistic) cost of quantifying high-frequency TP and TSS using grab samples, the design of efficient sampling protocols that take advantage of the availability of high-frequency surrogate data generated by in-situ sensors needs to be informed by answers to these more nuanced questions.

2.6. Conclusions

This research has demonstrated how high-frequency sensor data collected at multiple sites can provide multiple lines of evidence to improve hydrologic and hydrochemical process understanding. Coupled with generation of surrogate relationships, the high-frequency data collected in the LBRTB suggest first that the spring snowmelt period is the dominant TSS and TP load generation period, and the period of early snowmelt generates the vast majority of the annual TSS and TP load via surface pathways from snowmelt close to the streams that carry TP and TSS loads. Second, water quality constituent loads estimated using weekly or monthly data are not representative of the high variability in discharge and constituent concentrations, and tend to, in the majority of cases, under predict the true loading because of the high probability that peaks in discharge and concentration are missed, and should be considered as order of magnitude estimates of the true loading.

The two component hydrograph separation supported our conceptual model of discharge in the unregulated portions of the Little Bear River, which may be applicable to many snowmelt driven watersheds that are similar to the Little Bear River. Discharge from slow subsurface pathways (i.e., baseflow) is relatively constant throughout the year and does not extend to a great degree into the peaks of the spring snowmelt hydrograph. According to the simple mixing model, more than half of the annual discharge is from fast pathways (i.e., quickflow) that dominate the spring snowmelt hydrograph and dilute the relatively constant baseflow. The chemical signatures of baseflow and quickflow appear to be distinct, suggesting that the two flow paths have very different residence times within the system.

Metrics based on high-frequency profiles of DO concentrations and saturation deficits, such as estimates of photosynthesis and respiration rates, are useful indicators of instream metabolism and can easily be calculated from high-frequency data. In the Little Bear River, we found that these rates were very different from site to site, and because they are related to physical, chemical, and biological processes, they represent an opportunity for better understanding the interactions among hydrologic, hydrochemical, and biological processes. They may also provide useful indicators for quantifying the degree to which sites and their contributing catchments have been affected by human disturbance.

The results of our analyses demonstrate the need for and value of high-frequency, continuous time series of discharge and hydrochemical variables. Indeed, the observing system, surrogate methods, and cyberinfrastructure that we have demonstrated are advances to the infrastructure available for the design and implementation of environmental observatories and together have enabled us to gain insights into the importance and relative magnitude of hydrologic pathways and responses that are only possible through high-frequency data. Data and analyses such as these, as well as the cyberinfrastructure that enabled them, make it possible for us to better understand the processes that control the fluxes, flow paths, and stores of both water and water-borne constituents. They also present challenges for current hydrologic and water quality models, which typically lack appropriate mechanisms for representing these types of responses on the time scales at which they were observed. Without this type of information, we have no way of testing many of the concepts and assumptions that are the basis of our current understanding of hydrological processes, and our ability to predict hydrologic and water quality response will remain constrained.

2.7. Data Availability

The data referenced in this paper are available via the LBRTB website http://littlebearriver.usu.edu, which is maintained by the Utah Water Research Laboratory at Utah State University. Raw data streaming from the sensors in the LBRTB are available on the website within hours of being collected. Quality controlled data are also available, and are periodically added to the database as quality control procedures are completed.

References

Bond, B. J., J. A Jones, G. Moore, N. Phillips, D. Post, and J. J. McDonnell (2002), The zone of vegetation influence on baseflow revealed by diel patterns of streamflow and vegetation water use in a headwater basin, *Hydrological Processes*, *16*, 1671-1677, doi:10.1002/hyp.5022.

Buchanan, T. J. and W. P. Somers (1969), Discharge measurements at gaging stations, Book 3, Chapter A8, Applications of Hydraulics, in *Techniques of water-resources investigations of the United States Geological Survey*, United States Department of the Interior, Washington, D. C. (Available at http://pubs.usgs.gov/twri/twri3a8/html/pdf.html)

Burns, D. A., J. J. McDonnell, R. P. Hooper, N. E. Peters, J. E. Freer, C. Kendall, and K. Beven (2001), Quantifying contributions to storm runoff through end-member mixing analysis and hydrologic measurements at the Panola Mountain Research Watershed (Georgia, USA), *Hydrological Processes*, *15*, 1903-1924, doi:10.1002/hyp.246.

Burns, R., A. Terzis, and M. Franklin (2006), Design tools for sensor-based science, IEEE Workshop on Embedded Networked Sensors, Harvard University, Cambridge, Mass., May 2006. (Available at http://www.eecs.harvard.edu/emnets/papers/terzisEmnets06.pdf)

Chapra, S.C. (1997), *Surface Water-Quality Modeling*, 844 pp., McGraw-Hill, New York.

Christensen, V. G., P. P. Rasmussen, and A. C. Ziegler (2002), Real-time water quality monitoring and regression analysis to estimate nutrient and bacteria concentrations in Kansas streams, *Water Science and Technology*, *45*(9), 205-211.

Covino, T. P., and B. L. McGlynn (2007), Stream gains and losses across a mountain-to-valley transition: Impacts on watershed hydrology and stream water chemistry, *Water Resour. Res.*, *43*, W10431, doi:10.1029/2006WR005544.

Coynel, A., J. Schafer, J. E. Hurtrez, J. Dumas, H. Etcheber, and G. Blanc (2004), Sampling frequency and accuracy of SPM flux estimates in two contrasted drainage basins, *Science of the Total Environment*, *330*, 233-247, doi:10.1016/j.scitotenv.2004.04.003.

Dover, J. H. (1987), Geologic map of the Mount Naomi roadless area, Cache County, Utah, and Franklin County, Idaho, Department of the Interior, U. S. Geological Survey Miscellaneous Field Studies Map MF-1566-B.

Fisher, J., X. Meng, R. Rice, C. Butler, N. Molotch, T. C. Harmon, and R. Bales (2007), The Sierra Nevada-San Joaquin Hydrologic Observatory (SNSJHO): A WATERS Network Test Bed, *EOS Trans. AGU*, *88*(52), Fall Meet. Suppl., Abstract H13A-0963.

Glasgow, H. B., J. M. Burkholder, R. E. Reed, A. J. Lewitus, and J. E. Kleinman (2004), Real-time remote monitoring of water quality: A review of current applications, and advancements in sensor, telemetry, and computing technologies, *J. Experimental Marine Biology and Ecology*, *300*, 409-448, doi:10.1016/j.jembe.2004.02.022.

Hart, J. K. and K. Martinez (2006), Environmental sensor networks: A revolution in earth system science?, *Earth-Science Reviews*, *78*, 177-191, doi:10.1016/j.earscirev.2006.05.001.

Helsel, D. R. (2005), *Nondetects and Data Analysis: Statistics for Censored Environmental Data*, 250 pp., John Wiley and Sons, New York.

Jarvie, H. P., C. Neal, R. Smart, R. Owen, D. Fraser, I. Forbes, and A. Wade (2001), Use of continuous water quality records for hydrograph separation and to assess short-term variability and extremes in acidity and dissolved carbon dioxide for the River Dee, Scotland, *The Science of the Total Environment*, *265*, 85-98, doi:10.1016/S0048-9697(00)00651-3.

Johnes, P. J. (2007), Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density, *J. Hydrology*, *332*, 241-258, doi:10.1016/j.jhydrol.2006.07.006.

Kirchner, J. W. (2003), A double paradox in catchment hydrology and geochemistry, *Hydrological Processes*, *17*, 871-874, doi:10.1002/hyp.5108.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, 42, W03S04, doi:10.1029/2005WR004362.

Kirchner, J. W., X. Feng, C. Neal, and A. J. Robson (2004), The fine structure of water-quality dynamics: the (high-frequency) wave of the future, *Hydrological Processes*, *18*, 1353-1359, doi:10.1002/hyp.5537.

Lundquist, J. D., D. R. Cayan, and M. D. Dettinger (2003), Meteorology and hydrology in Yosemite National Park: A sensor network application, in *Information Processing in Sensor Networks*, pp. 518-528, Proceedings Second International Workshop, IPSN 2003, Palo Alto, Calif., April 22-23, 2003, Edited by F. Zhao and L. Guibas, LNCS 2634, Springer-Verlag, doi:10.1007/3-540-36978-3.

McDonnell, J. J. (1990), A rationale for old water discharge through macropores in a steep, humid catchment, *Water Resour. Res.*, *26*(11), 2821-2832.

McGlynn, B. L., and J. J. McDonnell (2003), Quantifying the relative contributions of riparian and hillslope zones to catchment runoff, *Water Resour. Res.*, *39*(11), 1310, doi:10.1029/2003WR002091.

Minsker, B., D. Maidment, D. Hodges, P. Montagna, and J. Bonner (2006), An environmental information system for hypoxia in Corpus Christi Bay: A WATERS Network Testbed, *EOS Trans. AGU*, *87*(52), Fall Meet. Suppl., Abstract H21F-1432.

Montgomery, J. L., T. Harmon, W. Kaiser, A. Sanderson, C. N. Haas, R. Hooper, B. Minsker, J. Schnoor, N. L. Clesceri, W. Graham, and P. Brezonik (2007), The WATERS Network: An integrated environmental observatory network for water research, *Environ. Sci. Technology*, *41*(19), 6642-6647. (Available at http://pubs.acs.org/subscribe/journals/esthag/41/i19/pdf/100107feature_waters.pdf)

Moore, J. N., J. T. Harper, W. W. Woessner, and S. Running (2007), Headwaters of the Missouri and Columbia Rivers WATERS Test Bed site: Linking time and space of snow melt runoff in the crown of the continent, *EOS Trans. AGU*, *88*(52), Fall Meet. Suppl., Abstract H13A-0964.

Mulholland, P. J., J. N Houser, and K. O. Maloney (2005), Stream diurnal dissolved oxygen profiles as indicators of in-stream metabolism and disturbance effects: Fort Benning as a case study, *Ecological Indicators*, *5*, 243-252, doi:10.1016/j.ecolind.2005.03.004.

Phillips, J. M., B. W. Webb, D. E. Walling, and G. J. L. Leeks (1999), Estimating the suspended sediment loads of rivers in the LOIS study area using infrequent samples, *Hydrological Processes*, *13*(7), 1035-1050.

Pinder, G. F., and J. F. Jones (1969), Determination of the ground-water component of peak discharge from the chemistry of total runoff, *Water Resour. Res.*, *5*(2), 438-445.

Schaefer, D. H., S. A. Thiros, and M. R. Rosen (2006), Ground-water quality in the carbonate-rock aquifer of the Great Basin, Nevada and Utah, 2003, U.S. Department of the Interior, U.S. Geological Survey, Scientific Investigations Report 2005-5232. (Available at http://pubs.usgs.gov/sir/2005/5232/PDF/SIR2005_5232.pdf)

Shanley, J. B., C. Kendall, T. E. Smith, D. M. Wolock, and J. J. McDonnell (2002), Controls on old and new water contributions to stream flow at some nested catchments in Vermont, USA, *Hydrological Processes*, *16*, 589-609, doi:10.1002/hyp.312.

Stevens, D. K., D. G. Tarboton, J. S Horsburgh, A. Spackman, and N. O. Mesner (2007), Little Bear River Test-Bed: Tools for environmental observatory design and implementation, *EOS Trans. AGU*, *88*(52), Fall Meet. Suppl., Abstract H13A-0980.

Stewart, M., J. Cimino, and M. Ross (2007), Calibration of base flow separation methods with streamflow conductivity, *Ground Water*, *45*(1), 17-27.

Stubblefield, A. P., J. E. Reuter, R. A. Dahlgren, and C. R. Goldman (2007), Use of turbidometry to characterize suspended sediment and phosphorus fluxes in the Lake Tahoe basin, California, USA, *Hydrological Processes*, *21*(3), 281-291, doi:10.1002/hyp.6234.

Szlavecz, K., A. Terzis, S. Ozer, R. Musăloiu-E., J. Cogan, S. Small, R. Burns, J. Gray, and A. Szalay (2006), Life under your feet: An end-to-end soil ecology sensor network, database, web server, and analysis service, Microsoft Technical Report MSR-TR-2006-90. (Available at ftp://ftp.research.microsoft.com/pub/tr/TR-2006-90.pdf)

Tetzlaff, D., S. Waldron, M. J. Brewer, and C. Soulsby (2007), Assessing nested hydrological and hydrochemical behavior of a mesoscale catchment using continuous tracer data, *J. Hydrology*, *336*, 430-443, doi:10.1016/j.jhydrol.2007.01.020.

Tomlinson, M. S., and E. H. De Carlo (2003), The need for high resolution time series data to characterize Hawaiian streams, *J. Am. Water Resour. Assoc. (JAWRA)*, *39*(1), 113-123, doi:10.1111/j.1752-1688.2003.tb01565.x.

Uhrich, M. A. and H. M. Bragg (2003), Monitoring instream turbidity to estimate continuous suspended sediment loads and yields and clay-water volumes in the Upper North Santiam River Basin, Oregon, 1998-2000, U. S. Department of the Interior, U.S. Geological Survey, Water-Resources Investigations Report 03-4098. (Available at http://pubs.usgs.gov/wri/WRI03-4098/pdf/wri034098.pdf)

U.S. Census Bureau (2000), Department of Commerce, Bureau of the Census, Washington, DC: Government Printing Office, Downloaded place- and tract-level data from 2000 Census of Population from http://www.census.gov on June 1, 2004.

Wang, H., M. Hondzo, C. Xu, V. Poole, and A. Spacie (2003), Dissolved oxygen dynamics of streams draining an urbanized and an agricultural catchment, *Ecological Modeling*, *160*, 145-161, doi:10.1016/S0304-3800(02)00324-1.

Welty, C., A. J. Miller, R. J. Ryan, N. Crook, T. Kerchkof, P. Larson, J. Smith, M. L. Baeck, S. Kaushal, K. Belt, M. McGuire, T. Scanlon, J. Warner, R. Shedlock, L. Band, and P. Groffman (2007), Baltimore WATERS Test Bed – Quantifying groundwater in urban areas, *EOS Trans. AGU*, *88*(52), Fall Meet. Suppl., Abstract H13A-0966.

Wondzell, S. M., M. N. Gooseff, and B. L. McGlynn (2007), Flow velocity and the hydrologic behavior of streams during baseflow, *Geophysical Res. Letters*, *34*, L24404, doi:10.1029/2007GL031256.

Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental design and initial results from the Mahurangi River Variability Experiment: MARVEX, in *Observations and Modelling of Land Surface Hydrological Processes*, edited by V. Lakshmi, J. D. Albertson and J. Schaake, pp. 201-213, Water Resources Monographs, American Geophysical Union, Washington, D. C.

Table 2.1. Little Bear River Monitoring Sites

Site Number	Site Name	Latitude	Longitude	Site Description
1	Upper South Fork	41.4954	-111.818	Unregulated watershed relatively unimpacted by agricultural or urban pollutant sources.
2	Lower South Fork	41.5065	-111.8151	Unregulated. Located on the South Fork below the confluence with its major tributary, Davenport Creek.
3	East Fork	41.5292	-111.7993	Located below Porcupine Reservoir on the East Fork. During the summer irrigation season, the entire East Fork is diverted at this location, leaving the downstream river channel dry during most years.
4	Confluence	41.5361	-111.8305	Located below the confluence of the East and South Forks. During summer, this site is primarily South Fork water as the East Fork is entirely diverted for irrigation.
5	Paradise	41.5756	-111.8552	Located a short distance upstream of Hyrum Reservoir and representative of the cumulative effects of the watershed above Hyrum Reservoir.
6	Wellsville	41.6435	-111.9176	Located a short distance downstream of Hyrum Reservoir. Winter flow is primarily groundwater because there are no releases from Hyrum Dam. When Hyrum Reservoir fills in the spring, high flows associated with spills from the reservoir pass this site. Summer flow is essentially groundwater as releases from Hyrum Dam are diverted for irrigation immediately below the dam and do not contribute to river flow.
7	Mendon	41.7185	-111.9464	Near the terminus of the river, just upstream of the confluence with Cutler Reservoir. Influenced primarily by releases from Hyrum Reservoir and agriculture return flows.
8	Lower Watershed Weather Station	41.667	-111.8906	Located near the border of the watershed and characteristic of the lower watershed below Hyrum Reservoir.
9	Upper Watershed Weather Station	41.5355	-111.8059	Located near the confluence of the South and East Forks and characteristic of the mid to upper watershed.
10	Little Bear SNOTEL	41.40	-111.53	Located in the headwaters of the South Fork.
11	Dry Bread Pond SNOTEL	41.40	-111.82	Located in the headwaters of the East Fork.

Table 2.2. Variables Measured at Each Monitoring Site and Sensor Specifications

Variable	Sensor	Specifications
Stream Monitoring Sites		
Stage	SPXD-600 Pressure Transducer KWK Technologies, Inc.	Accuracy: ±1% of the full measurement span
Turbidity	DTS-12 turbidity sensor Forest Technology Systems, Inc.	Accuracy: ±2% 0 to 500 NTU and ±4% 501 to 1600 NTU
Water Temperature	Hydrolab MiniSonde5 thermistor Hach Environmental, Inc.	Accuracy: ±0.1 °C Resolution: 0.01 °C
Dissolved Oxygen Concentration	Hydrolab MiniSonde5 optical LDO sensor Hach Environmental, Inc.	Accuracy: ±0.1 mg L^{-1} at < 8 mg L^{-1} and ±0.2 mg L^{-1} at > 8 mg L^{-1} Resolution: 0.01 mg L^{-1}
pH	Hydrolab MiniSonde5 reference electrode Hach Environmental, Inc.	Accuracy: ±0.2 pH units Resolution: 0.01 pH units
Specific Conductance	Hydrolab MiniSonde5 4-electrode, temperature compensated conductivity sensor Hach Environmental, Inc.	Accuracy: ±0.5% Resolution: 0.001 mS cm^{-1}
Weather Monitoring Sites		
Precipitation	TE25 tipping bucket rain gage with a 20.32 cm orifice Texas Electronics	Accuracy: ±1% up to 2.54 cm hr^{-1} Resolution: 0.254 mm
Air Temperature	CS215 temperature and relative humidity sensor Campbell Scientific, Inc.	Accuracy: ±0.4 °C from +5 °C to +40 °C, and ±0.9 °C from -40 °C to +70 °C
Relative Humidity	CS215 temperature and relative humidity sensor Campbell Scientific, Inc.	Accuracy: ±2% at 25 °C in the 10-90% range and ±4% in the 0-100% range
Wind Speed	R. M. Young Wind Sentry Set	Accuracy: ±0.5 m s^{-1}
Wind Direction	R. M. Young Wind Sentry Set	Accuracy: ±0.5 degrees
Solar Radiation	PYR-P Silicon Pyranometer Apogee Instruments, Inc.	Accuracy: 5% for daily total radiation
Barometric Pressure	Setra 278 Barometric Pressure Sensor	Accuracy: ±0.5 mb at +20 °C

Table 2.3. Average DO Deficit (D), Rate Constant (k_a), Respiration Rates (R), and Photosynthesis Rates (P) Calculated Using the Extreme Value Method for the Period Between July 1, 2008 and July 7, 2008

Site	D_{avg} (mg L^{-1})	k_a (day^{-1})	R (mg L^{-1} day^{-1})	P_{avg} (mg L^{-1} day^{-1})
Mendon	-1.62	2.1	6.2	3.7
Wellsville	-0.97	44.1	100.8	58.1
Paradise	0.61	42.0	29.6	56.3
Lower South Fork	-0.06	12.3	4.7	6.2

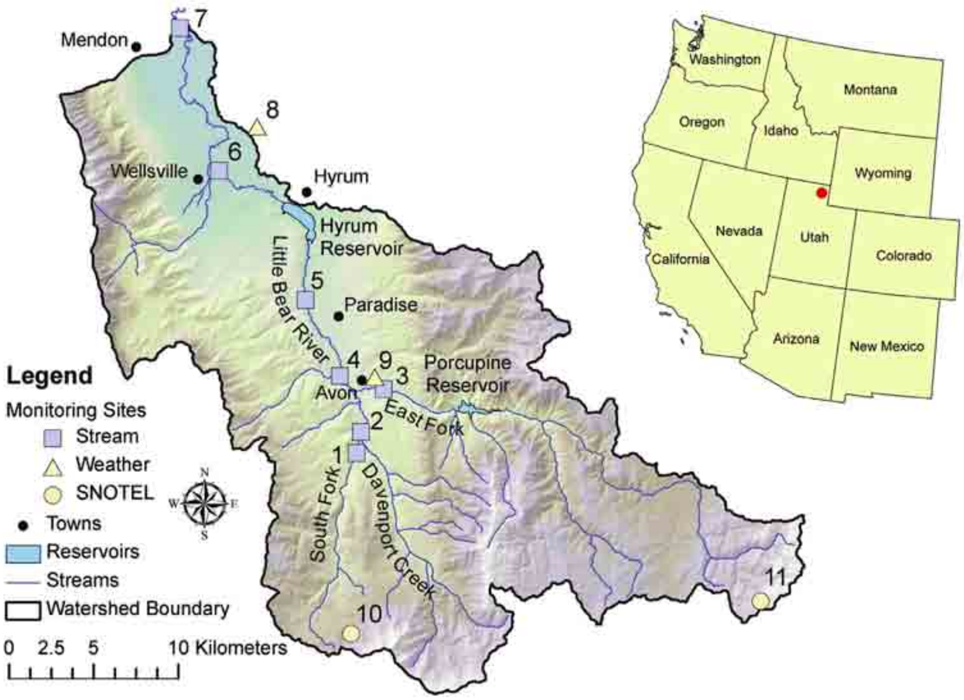


Figure 2.1. Little Bear River watershed. Descriptions of sampling sites are contained in Table 2.1.

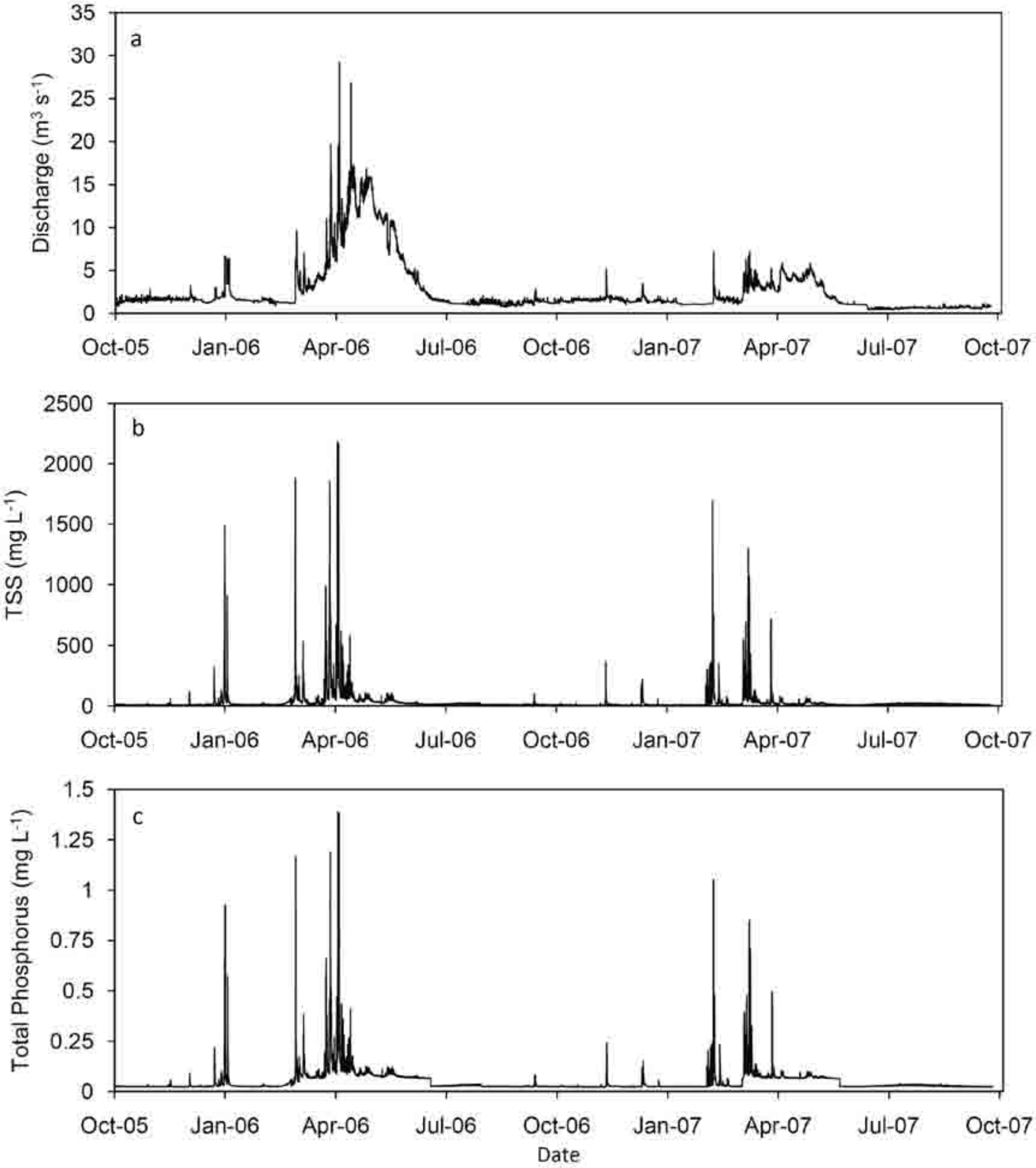


Figure 2.2. Continuous (half hourly) estimates of discharge (a), total suspended solids concentration (b), and total phosphorus concentration (c) at the Paradise site.

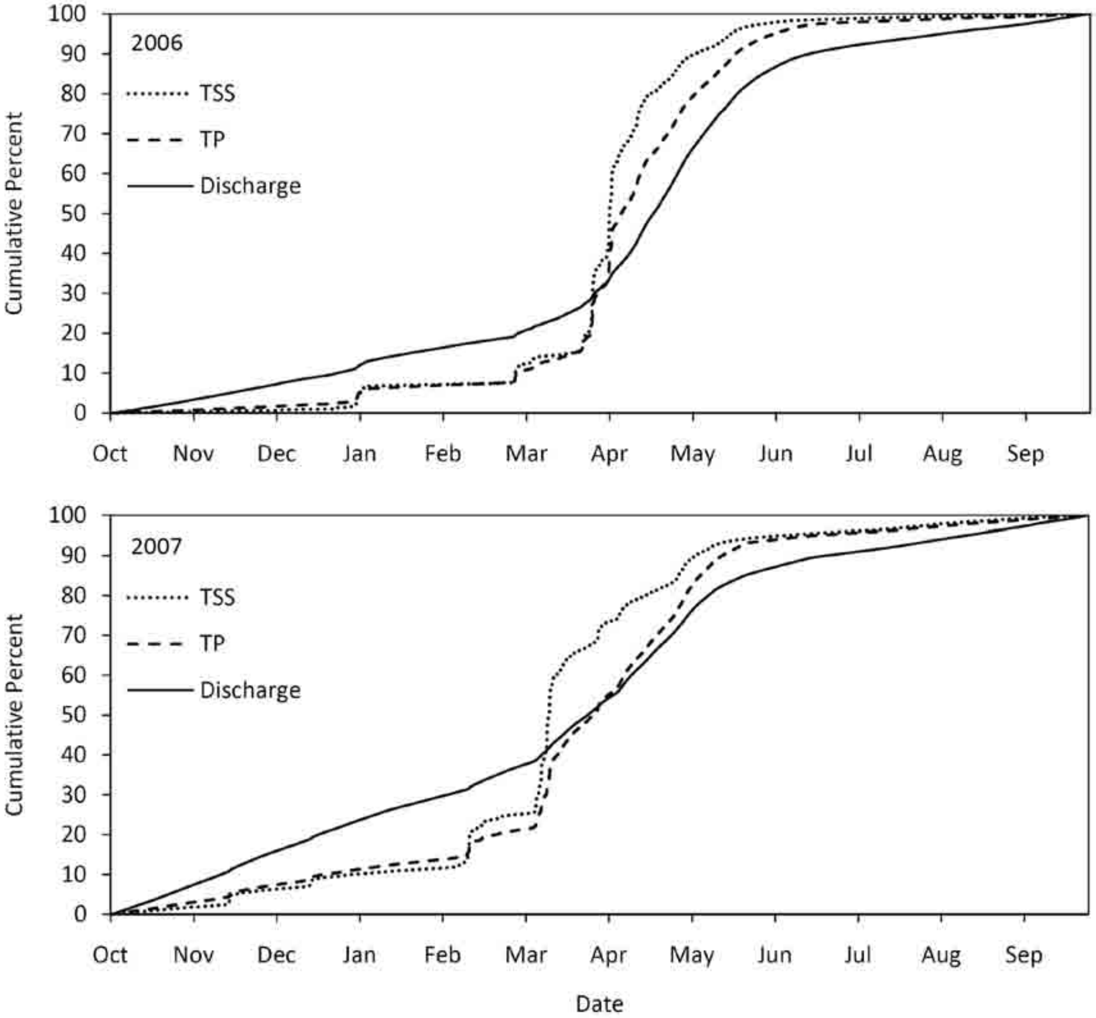


Figure 2.3. Cumulative percent of annual discharge, TSS, and TP loads contributed by date for water years 2006 and 2007 at the Paradise site.

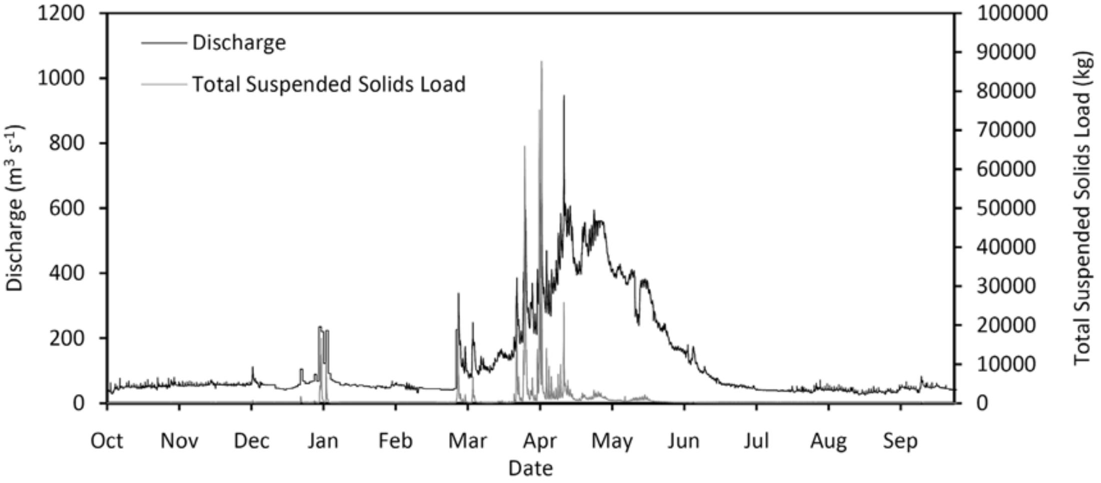


Figure 2.4. Discharge and 30-minute total suspended solids loads estimated using the synthetic concentration time series for the Paradise site during water year 2006.

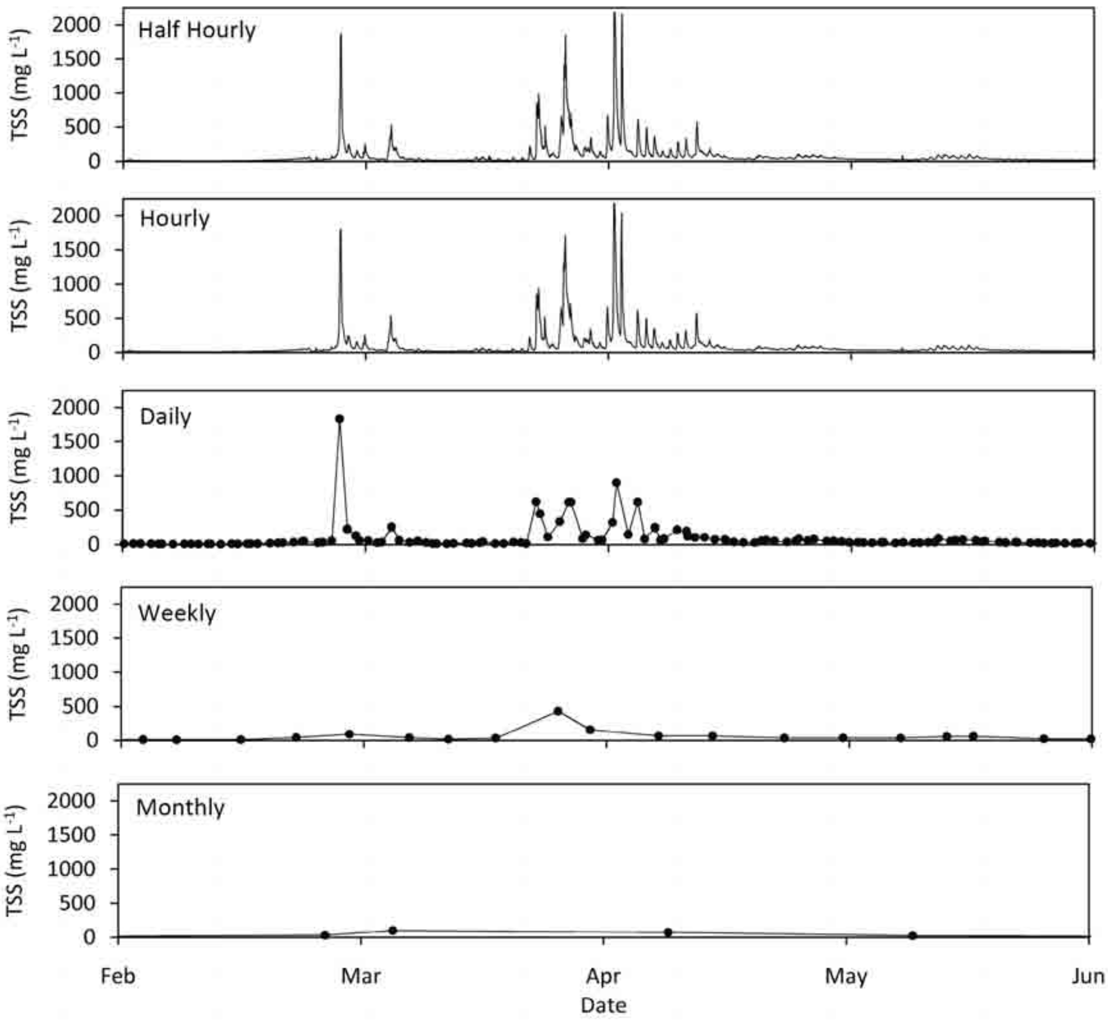


Figure 2.5. Total suspended solids concentrations at the Paradise site during spring of 2006 at varying sampling frequencies as sub sampled from the synthetic concentration estimates. The daily, weekly, and monthly time series are randomly selected points.

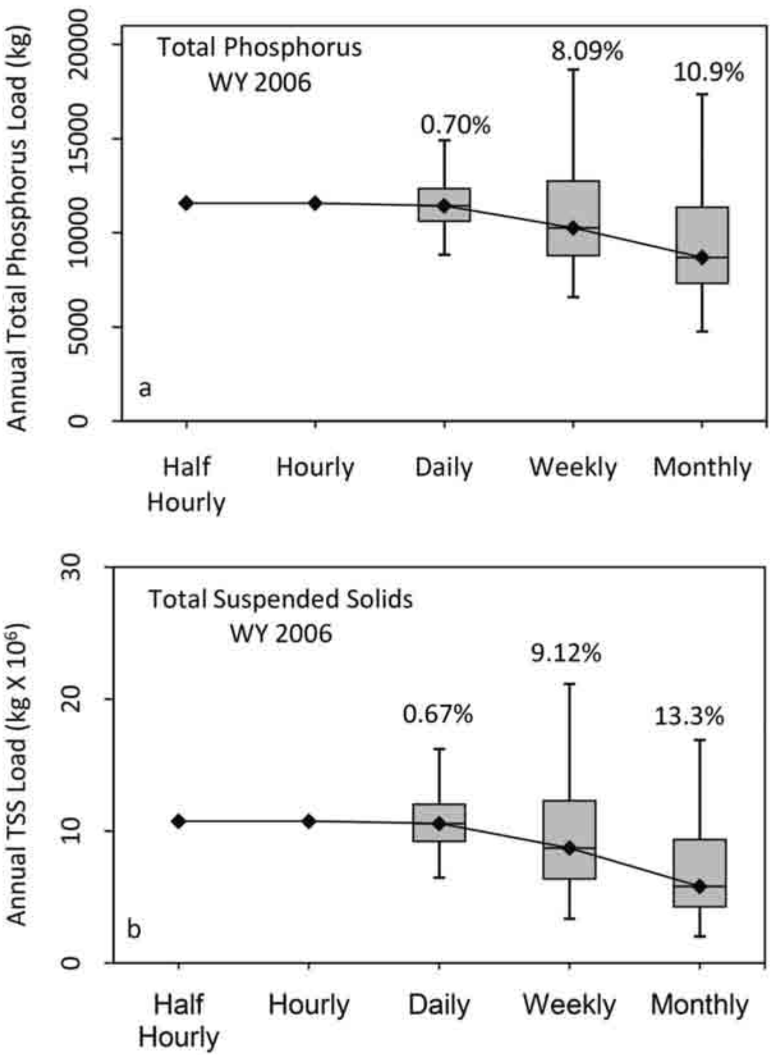


Figure 2.6. Box and whisker plots showing the results of varying sampling frequencies on estimated TP (a) and TSS (b) loads at the Paradise site for water year 2006. The half hourly result uses all of the continuous data, hourly represents the load estimate from sub sampling on the hour, and daily, hourly, and monthly box plots represent 10,000 estimates of the annual load given randomly selected sample times within each day, week, or month. The boxes represent the first and third quartiles and the whiskers represent the lower and upper adjacent values. The medians of each of the sets of realizations are also indicated. The percentages above the upper whisker represent the portion of load estimates that fell above the upper adjacent level.

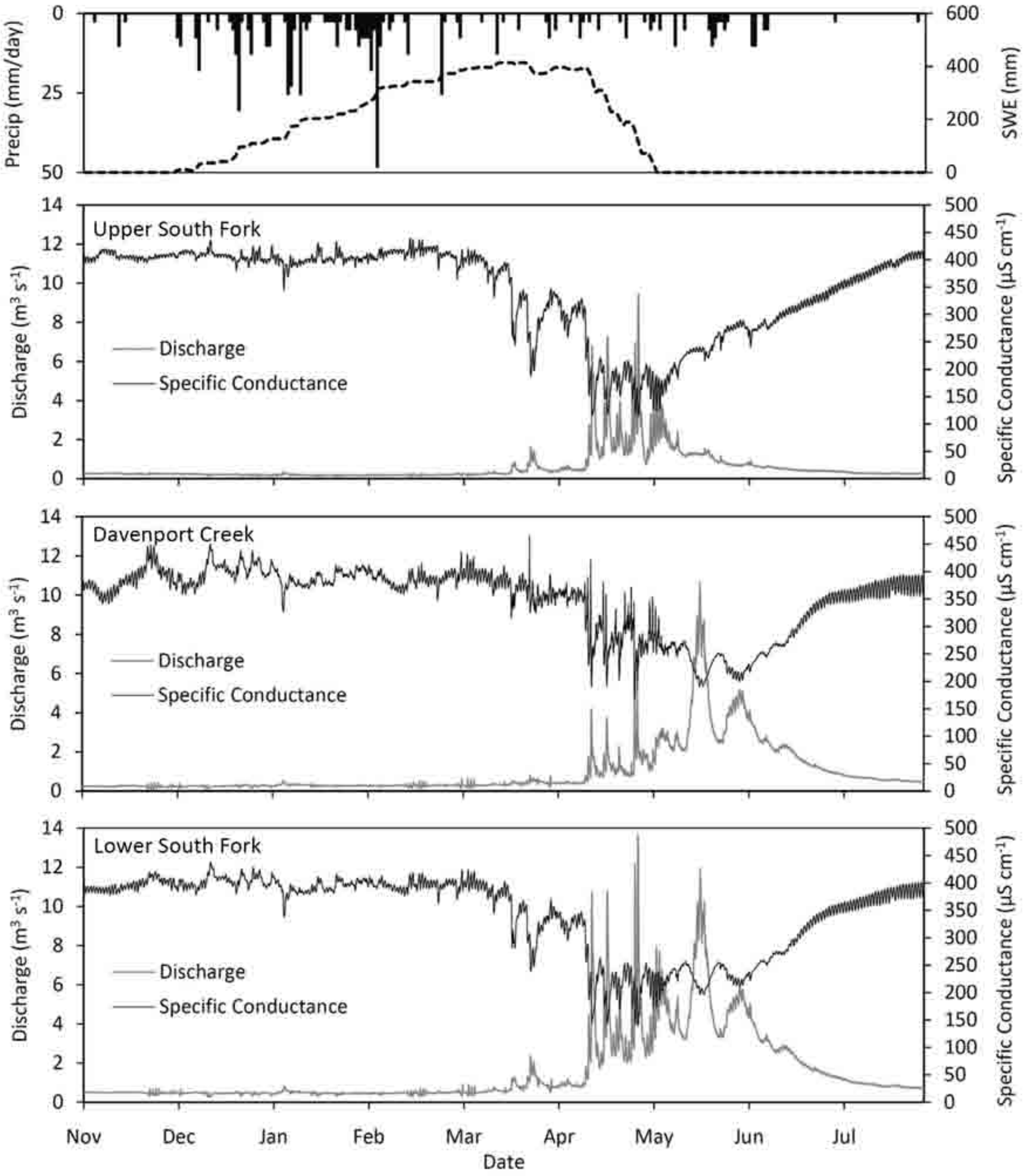


Figure 2.7. Discharge and specific conductance for the period between November 1, 2007 and July 31, 2008 in the South Fork and Davenport Creek. Precipitation and snow water equivalent are from the Little Bear SNOTEL site.

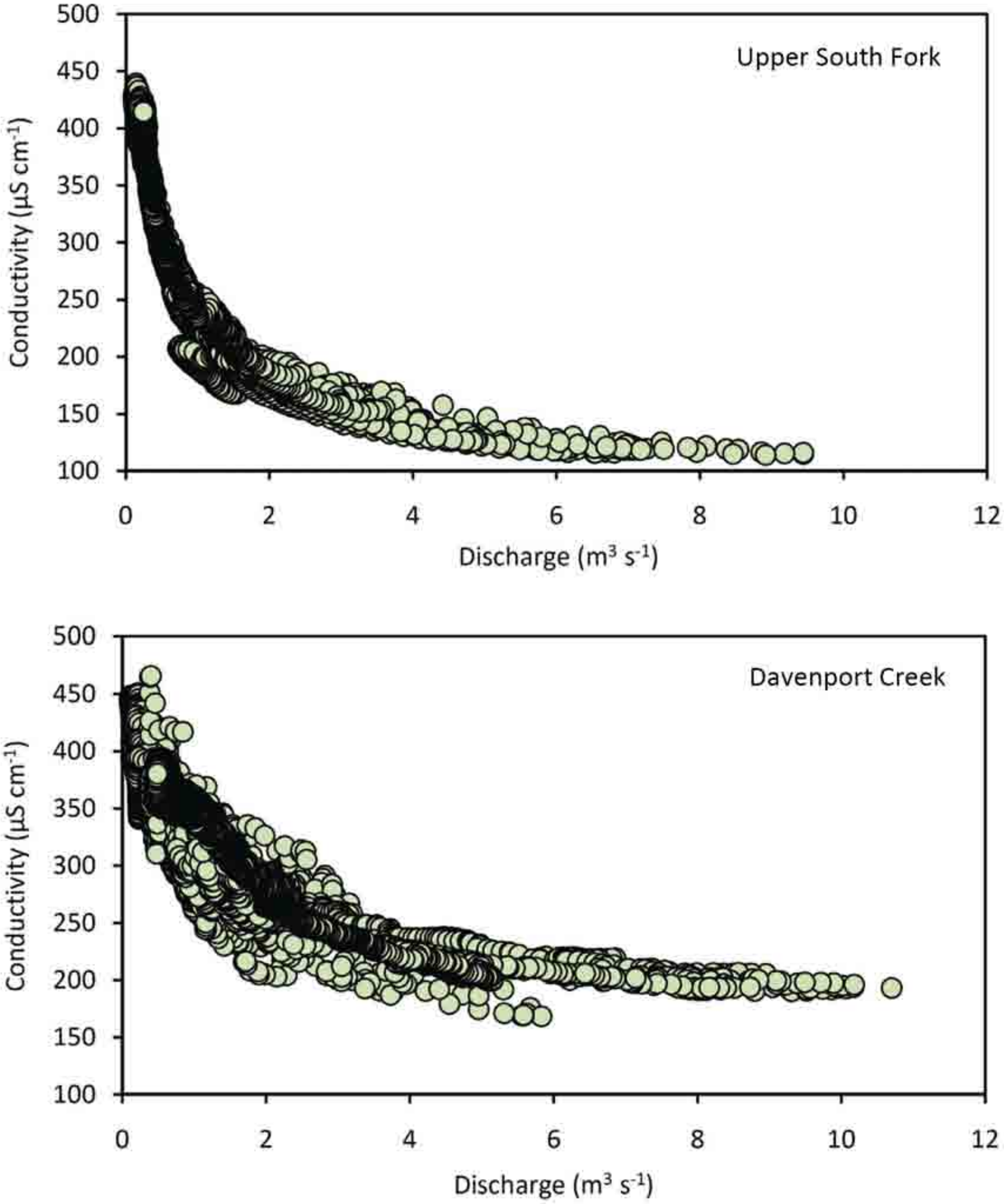


Figure 2.8. Specific conductance plotted versus discharge for the Upper South Fork and Davenport Creek catchments for the period between November 1, 2007 and July 31, 2008.

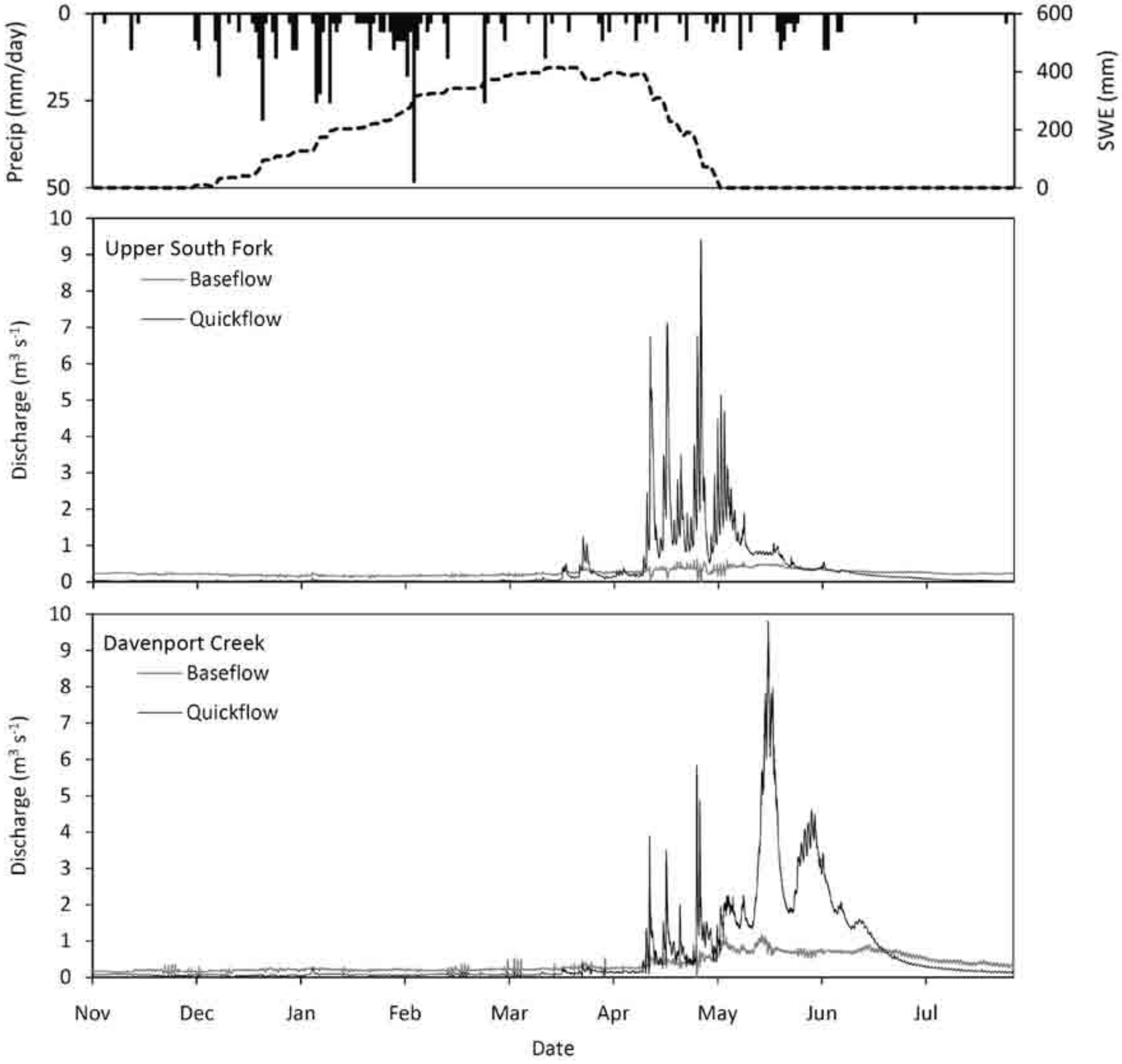
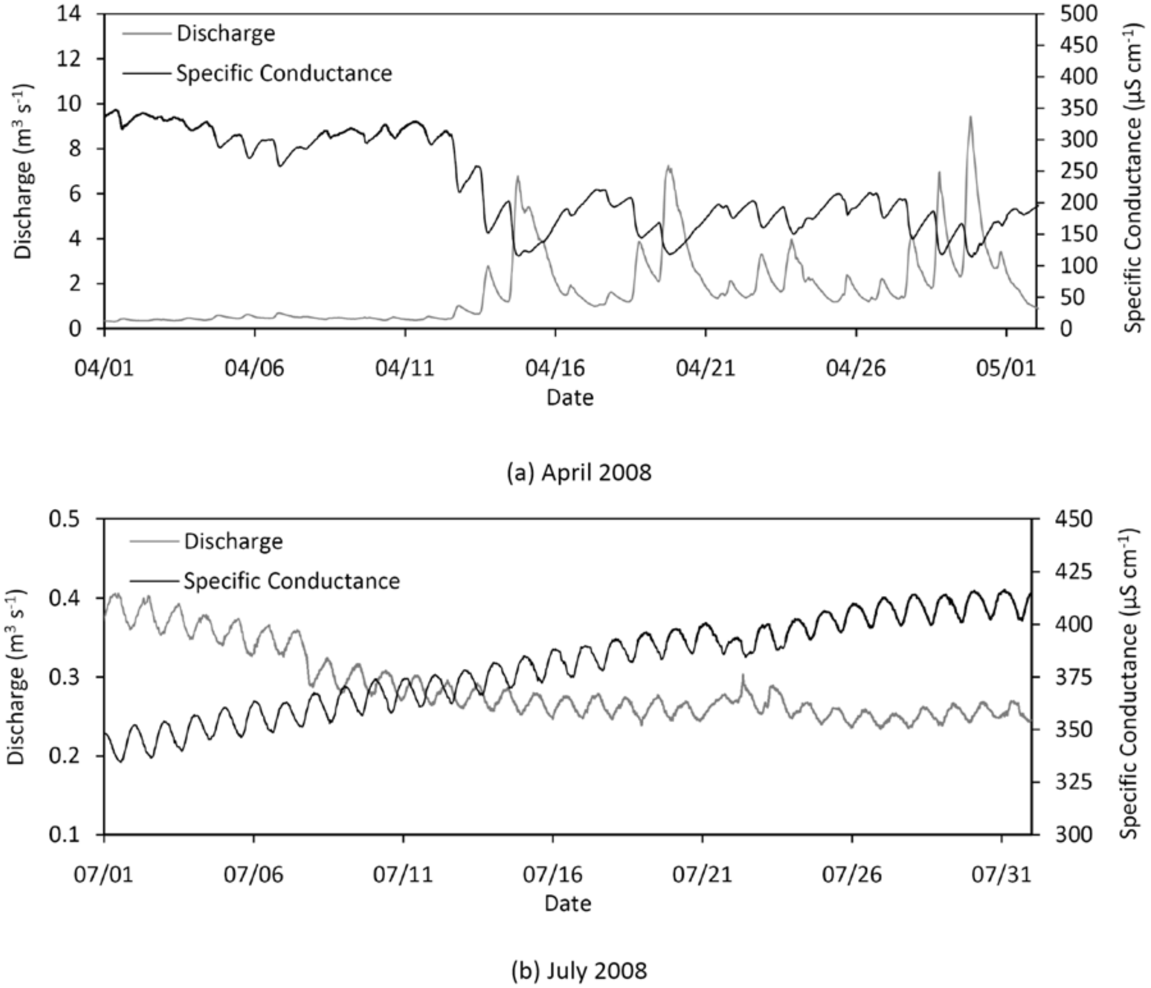


Figure 2.9. Hydrograph separation results for the Upper South Fork and Davenport Creek catchments based on 30-minute discharge and specific conductance data for the period between November 1, 2007 and July 31, 2008. Precipitation and snow water equivalent are from the Little Bear SNOTEL site.



(a) April 2008

(b) July 2008

Figure 2.10. Diurnal patterns in specific conductance at the Upper South Fork monitoring site during April of 2008 (a) and July of 2008 (b).

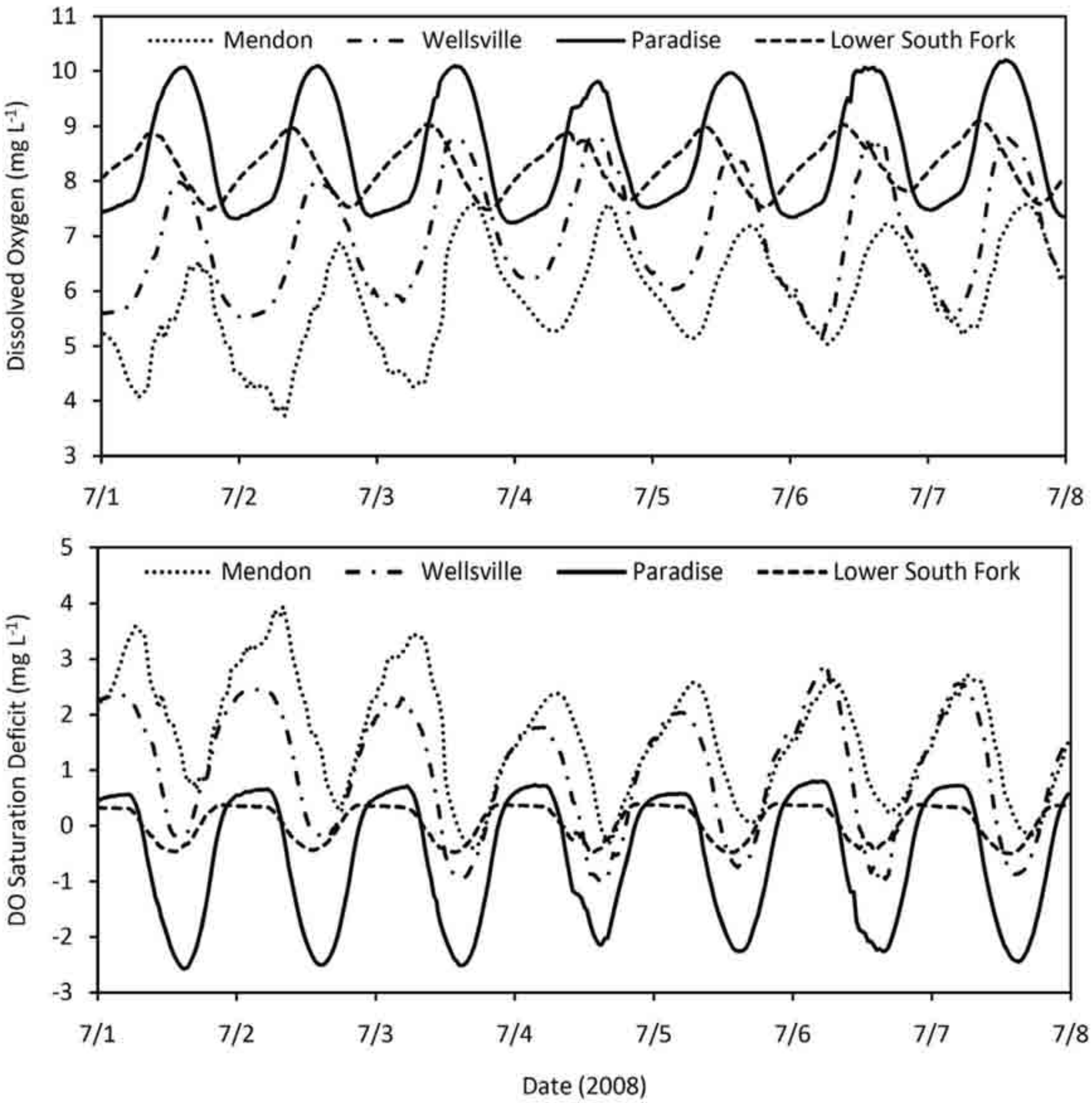


Figure 2.11. Dissolved oxygen concentrations and dissolved oxygen deficits at the Mendon, Wellsville, Paradise, and Lower South Fork sites during the first week of July 2008.

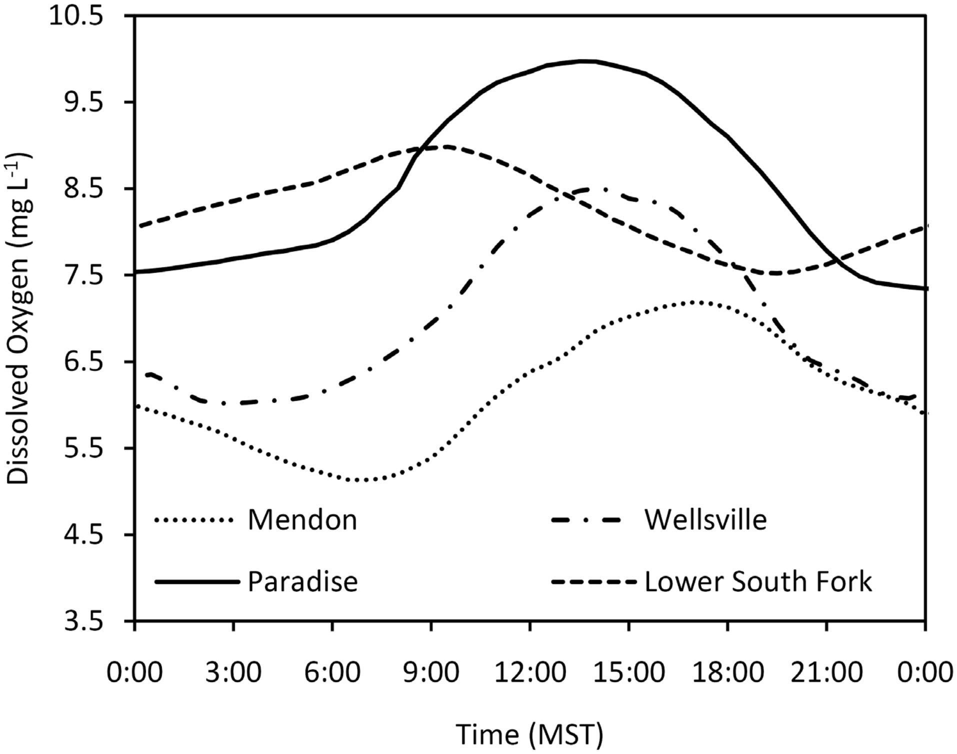


Figure 2.12. Dissolved oxygen concentrations on July 5, 2008 at the Mendon, Wellsville, Paradise, and Lower South Fork sites.

CHAPTER 3

A RELATIONAL MODEL FOR ENVIRONMENTAL AND WATER RESOURCES DATA[1]

Abstract

Environmental observations are fundamental to hydrology and water resources, and the way these data are organized and manipulated either enables or inhibits the analyses that can be performed. The Observations Data Model presented here provides a new and consistent format for the storage and retrieval of point environmental observations in a relational database designed to facilitate integrated analysis of large datasets collected by multiple investigators. Within this data model, observations are stored with sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and to provide traceable heritage from raw measurements to useable information. The design is based upon a relational database model that exposes each single observation as a record, taking advantage of the capability in relational database systems for querying based upon data values and enabling cross dimension data retrieval and analysis. This paper presents the design principles and features of the Observations Data Model and illustrates how it can be used to enhance the organization, publication, and analysis of point observations data while retaining a simple relational format. The contribution of the data model to water resources is that it

[1] Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resour. Res., 44*, W05406, doi:10.1029/2007WR006392. Copyright 2008 by the American Geophysical Union.

represents a new, systematic way to organize and share data that overcomes many of the syntactic and semantic differences between heterogeneous datasets, thereby facilitating an integrated understanding of water resources based on more extensive and fully specified information.

3.1. Introduction

Environmental observations are fundamental to hydrology and water resources, and the manner in which the data are collected, organized, and manipulated either enables or inhibits their scientific analysis [*Tomasic and Simon*, 1997; *Pokorný*, 2006]. When scientists and engineers want to search for and use environmental observations data, they are generally faced with the following problems [*Tomasic and Simon*, 1997]: (1) data are not sufficient or do not exist; (2) data are not published and are hard to locate; (3) data are not easy to access, they are either private or expensive, or require costly pre-processing before they can be used; (4) data are not easy to use because they are inconsistent or non-compatible; and (5) data are not adequately documented. Addressing these issues is one of the main challenges influencing recent developments in environmental information systems, which include water resources and hydrologic information systems [*Bouganim et al.*, 2001; *Pokorný*, 2006].

Even for datasets that have been published for widespread use, points three through five above still apply. Generally, datasets published on public web sites are in file-based systems that are different syntactically (e.g., file types, file formats, and data structure) and semantically (e.g., variable names, units, and descriptive metadata) from one data source to the next. In accessing these data archives, users are faced with the daunting task of navigating through directories and supporting files to find all of the metadata necessary for interpreting and using the data. There is a fundamental need within the hydrologic and environmental engineering communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. This need is being driven by the ever increasing number of environmental observations being produced as sensor technology improves, as the number, size, and complexity of environmental monitoring programs grow (including efforts to establish a national network of large scale environmental observatories), and as engineers and scientists realize that it is as important to characterize the environment with observations as it is to describe it with models and simulations. It is critical that the data, when published, be carefully annotated with metadata so that they can be unambiguously interpreted and used.

In this paper we present a logical database design for an Observations Data Model (ODM) that advances the information science knowledge base of water resources research. We describe a relational model that eases access to and manipulation of time series of observations from experimental sites and watersheds and facilitates data publishing, querying, retrieval, and analysis among domains and investigators. This design identifies the entities, attributes, and relationships required to represent observations, but it is independent of its physical implementation (i.e., it can be implemented within any relational database management system). This system has been implemented and used to publish a wide range of environmental data at 11 Test Bed sites that are part of an effort to advance environmental observatory design (http://www.watersnet.org/wtbs/index.html). The experience in implementing this model at these 11 sites has demonstrated the generality and effectiveness of ODM.

ODM is focused on observations made at a point, such as those made at a streamflow gage or a stationary weather station, although observations recorded from moving platforms or along routes can also be represented by treating location as an observation. The representation of spatially distributed data in ODM is limited to the presentation of time series of point observations that are at different spatial locations. ODM does not include raster datasets, for which we envision a different data model being developed. However, distributed time series data (e.g., time series of raster datasets such as weather radar observational grids) can be represented within ODM by using grid cell centers as observation sites.

ODM is the result of an effort to create a generic model of observational data from a range of water resources disciplines (hydrology, environmental engineering, meteorology, etc.) and to accommodate a range of different variables (precipitation, streamflow, water quality). The model has drawn upon input from community surveys and reviews [*Bandaragoda et al.*, 2005, 2006; *Tarboton*, 2005]. ODM has been applied to physical and chemical data from water systems, climate and weather observations, and aquatic biology measurements such as species distributions, and it is this flexibility that is largely responsible for its utility. ODM's ability to store and enable access to similarly formatted data and metadata from multiple domains, for example streamflow data and climate data for inputs to a hydrologic model, can greatly enhance the use of these data and can result in significant time savings and value added to the data. Additionally, the consistent format for data and metadata that ODM provides enables the development of standardized software applications on top of ODM. ODM enables easy and automated access to the data through a relational database management system, which enables multiple software developers to create compatible applications as well as the reuse of code for standard tasks such as data discovery and retrieval.

Additionally, ODM represents a new opportunity for many within the water resources community to approach the management, publication, and analysis of their data systematically – i.e., moving from collections of ASCII text or spreadsheet files to a relational data model that removes the burden of learning and interpreting diverse file formats from the data end user. Systematic data management using relational database systems has advanced data mining, predictive modeling, and deviation detection within the business community, where most operational data is stored in relational databases due to their reliability, scalability, available tools, and performance [*Connolly and Begg*, 2005]. The systematic data analysis capabilities that a relational data model enables have the potential to stimulate similar advances in the water resources area.

In this paper we describe the structure and features of ODM and discuss its implementation for data management in prototype environmental observatories. Section 3.2 discusses existing standards for environmental observations data. Section 3.3 describes the requirements considered in designing ODM. Section 3.4 gives the structure of ODM and describes some of its features. Section 3.5 provides examples of water resources data that have been incorporated into ODM, and Section 3.6 discusses the implementation of ODM within a national network of environmental observatory Test Beds.

3.2. Existing Standards for Environmental Observations

Much work has already been done to develop standards for exchanging information describing the collection, analysis, and reporting of environmental data. The Environmental Data Standards Council (EDSC) has developed a set of Environmental Sampling, Analysis, and Results Data Standards specifically for this purpose [*Environmental Data Standards Council*, 2006]. A similar standard has been developed by the National Water Quality Monitoring Council (NWQMC) specifically for water quality data elements [*National Water Quality Monitoring Council*, 2006], and the Open Geospatial Consortium (OGC) has developed a best practices document called "Observations and Measurements" that describes terminology and presents a framework and encoding for measurements and relationships between them [*Open Geospatial Consortium*, 2006]. These standards are focused primarily on the data elements required to facilitate the exchange of environmental observations without considering the format for persistent data storage such as in a relational database. In designing ODM, we strove to include the most important attributes of observations from these standards in a logical data model design that can be physically implemented in relational database management systems.

ODM's purpose is to manage the *storage* and *retrieval* of observations data as part of a broader hydrologic information system (HIS) that also provides data *discovery*, *analysis*, and *exchange* capability through software applications built on top of ODM. For example, within the HIS being developed by the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI), the main mechanism for the exchange of environmental observations is the WaterOneFlow web services (http://his.cuahsi.org/wofws.html). Web services are applications that provide the ability to pass information between computers over the Internet [*Goodall et al.*, 2008]. The WaterOneFlow web services transmit data extracted from an ODM database encoded as eXtensible Markup Language (XML) and formatted using an XML schema called WaterML [*Open Geospatial Consortium*, 2007]. This separation between content (i.e., the data stored in an ODM database) and presentation (i.e., the format of the data when it is transmitted) is an important aspect of the overall HIS design.

3.3. ODM Design Requirements

An observation is an event that results in a value describing some phenomenon [*Open Geospatial Consortium*, 2006]. Observation values are not self describing, and, because of this, interpretation of a particular set of observations requires contextual information, or metadata. Metadata is the descriptive information about data that explains the measurement attributes, their names, units, precision, accuracy, and data layout, as well as the data lineage describing how the data was measured, acquired, or computed [*Gray et al.*, 2005]. The importance of recording fundamental metadata to help others discover and access data products is well recognized [*Michener et al.*, 1997; *Bose*, 2002; *Gray et al.*, 2005]. ODM was designed to store environmental observations along with sufficient metadata to provide traceable heritage from raw measurements to usable information, allowing observations stored in ODM to be unambiguously interpreted and used.

Environmental observations are identified by the following fundamental characteristics: (1) the location at which the observations were made (space); (2) the date and time at which the observations were made (time); and (3) the type of variable that was observed, such as streamflow, water quality concentration, etc. (variable). In addition to these fundamental characteristics, there are many other attributes that provide additional information necessary for interpretation of observational data. These include the methods used to make observations, qualifying comments about the observation, and information about the organization that made the observation.

Table 3.1 presents general attributes that are important in interpreting and establishing the provenance of an observation. This list of attributes was compiled from comments received from a community review of a preliminary version of ODM [*Tarboton*, 2005]. All of the information contained in Table 3.1, except for the value of the observation itself, can be considered metadata. The ODM logical data model given in the following section has been designed to store observation values and their supporting metadata in a structured way.

3.4. ODM Logical Data Model

The logical data model for ODM is shown in Figure 3.1. The DataValues table at the center stores the numeric values for observations and links (foreign keys) to all of the data value level attributes. Most of the attribute details are stored in the tables surrounding the DataValues table to avoid redundancy. The relationships between tables are shown, along with all of the required primary and foreign keys. Each of these relationships has a name, which is indicated by a text label, and a directionality that is indicated by an arrow. For example, the relationship between the Sources table and the DataValues table is named "Generate" and has directionality that points from the Sources table to the DataValues table. This indicates that data sources *generate* data values. Additionally, the cardinality, or numeric relationship between entities in each of the tables, is shown at either end of each of the relationship lines. For example, the relationship line between the Variables and DataValues tables has "1..1" at the Variables end, and "0..*" at the DataValues end, indicating that there is one and only one variable associated with 0 or many DataValues (i.e., there is a one-to-many relationship between variables and data values) and that variables *characterize* data values. The subsections that follow describe how ODM encodes observations and their supporting metadata. Readers are referred to *Tarboton et al.* [2007] for the complete ODM design specifications and data dictionary.

3.4.1. Monitoring Site Geography, Location, and Offset

Within ODM, the geographic location of monitoring sites is specified through latitude and longitude coordinates as well as elevation information recorded in the Sites table. Additionally, ODM provides the option to specify local coordinates, which may be in a standard geographic projection (e.g., Universal Transverse Mercator) or a locally defined coordinate system specific to a study area. Both the spatial reference system associated with the horizontal and vertical coordinates and the accuracy with which the location of a monitoring site is known can be quantified within ODM. The field *PosAccuracy_m* is a numeric value intended to specify the uncertainty in the spatial location information.

Each monitoring site has a unique identifier that can be logically linked to one or more objects in a Geographic Information System (GIS) data model. Figure 3.2 depicts relationships between monitoring sites within an ODM database and points in a GIS data model. The GIS data model depicted in Figure 3.2 is Arc Hydro, which is a data structure for linking stream networks, monitoring points and watersheds within a GIS [*Maidment*, 2002]. This linkage between unique monitoring site identifiers and GIS object identifiers is generic and suitable for use with any geographic data model that includes the location of monitoring sites. For example, a linear referencing system on a river network, such as the National Hydrography Dataset [*Dewald*, 2006], might be used to specify the location of a site on a river network. Information from direct addressing relative to hydrologic objects, such as position of a stream gage along a stream reach, is often of greater value to a user than latitude and longitude information [*Maidment*, 2002].

The location at which observations were made may also be qualified by an offset, which is used to record the location of an observation relative to an appropriate local reference point, such as depth below the water surface. In some cases, such local reference is required for proper interpretation of the data. For example, observations of water temperature or dissolved oxygen may be made at a number of different depths at a location within a water body. The offset would be used to quantify the depth of each measurement below the surface. Within ODM, an offset is specified by a numeric value that is the offset distance, the units of the offset, and an offset description that defines the type of offset (e.g., below the water surface or above ground level).

3.4.2. Variable Information

The variables that can be represented in ODM range from hydrologic variables such as discharge and gage height to water quality variables such as nutrient and sediment concentrations to meteorological variables such as air temperature and precipitation as well as many others. The most fundamental attribute of an environmental variable is its name (e.g., discharge or temperature), but there are several other variable attributes recorded in ODM that are important, including: (1) the units of the observations for a variable (e.g., m^3 s^{-1}); (2) the medium in which the observations are made (e.g., surface water or sediment); (3) the regularity with which observations are made; (4) the support, spacing, and extent of observations; and (5) the nature of the observation as an actual measurement (e.g., stage) or a derived value (e.g., discharge derived from stage). All of this information is represented at the variable level within ODM.

3.4.2.1. Time Support, Spacing, and Extent

To interpret values that comprise a time series or set of observations, it is important to know the time scale information associated with the values. *Blöschl and Sivapalan* [1995] review the important issues. Any set of observations is quantified by a scale triplet comprising support, spacing, and extent. Extent is the full range of time over which the observations occur, spacing is the time between observations, and support is the averaging interval implicit in any observation. In ODM, the time support associated with observations is specified by a numeric value that quantifies the support and an indication of the units associated with the support value. Extent and spacing are properties of multiple observations and are defined by the set of dates and times associated with the observations. Dates and times associated with observations are stored in local time (in the time zone in which the observation was made), UTC time, and ODM also stores the UTC offset to ensure that dates and times are unambiguous.

3.4.2.2. Data Types

The environmental processes that we wish to characterize through observation may be dynamic and continuous in nature, but our ability to measure them is constrained to particular instants or intervals of time. To interpret environmental observations, it is important to know whether an observation is an instantaneous result, such as in the case of water quality variables where a sample is collected at an instant in time, or whether the observation is a cumulative or incremental value resulting from a measurement device such as a rain gage that accumulates a quantity over time. In ODM this information is referred to as the data type and is recorded in the DataType attribute in the Variables table. Table 3.2 lists the major data types that can be represented within ODM. This list expands upon the data types listed by *Maidment* [2002], and it is anticipated that as more data types are incorporated into specific ODM instances that this list will grow.

3.4.2.3. Samples and Methods

The method used to make a measurement is important for its interpretation. Within ODM, individual observation values can be associated with a record in the Methods table that describes how a physical observation was made or collected. Descriptive information about each measurement method can be stored and can include specific and detailed information about the technique or equipment used. In the case of observations derived from laboratory samples, ODM provides the additional feature of storing information in the Samples table to link individual observations to the specific physical samples analyzed in a laboratory. Details about the laboratory methods and protocols used in analyzing the samples can be stored in the LabMethods table.

3.4.3. Quality Control

Data versioning and quality control are key concepts in environmental data management where raw data streams in from in-situ sensors through telemetry networks. Raw sensor data can contain a variety of errors caused by equipment malfunction, instrument drift, improper calibration, vandalism, or other causes. In most cases, raw sensor data are not useful for defensible scientific analyses until they have been filtered through a quality control process. To accommodate quality control measures and data versioning, each observation stored in ODM is assigned a quality control level that indicates the level of quality control to which a value has been subjected. The quality control levels used within ODM are stored in the QualityControlLevels table and have been adapted from those used by other earth observatory projects and communities [*Ahern*, 2004; *NASA*, 2005] so that ODM is consistent with these other efforts. The definitions for the quality control levels used by ODM are listed in Table 3.3.

3.4.4. Value Accuracy

Each observation stored in ODM can be attributed with an indication of the accuracy of the observation. This attribute is a numeric value that quantifies the total measurement accuracy defined as the nearness of a measurement to the true or standard value. The value accuracy quantifies the uncertainty of the measurement due to errors in both bias and precision. In practice, since the true value is not known, the value accuracy should be estimated based on knowledge of the instrument accuracy, measurement method, and operational environment. In some cases it is possible to quantify precision by statistical analysis of the scatter associated with repeated measurements and to quantify bias through comparison to specially designed unbiased measurements. Value accuracy can then be estimated by combining these using a root mean square sum. In other cases value accuracy will be a more subjective estimate.

Value accuracy is an observation level attribute because it can change with each measurement, dependent on the instrument or measurement protocol. For example, if streamflow is estimated using a V-notch weir, it is actually the stage that is measured, with accuracy limited by the precision and bias of the depth recording instrument. The conversion to discharge through the stage-discharge relationship results in greater absolute error for larger discharges. Inclusion of the value accuracy attribute, which will be unknown for many historic datasets because historically accuracy has not been recorded, adds to the size of data in ODM, but provides a way for factoring the accuracy associated with measurements into data analysis and interpretation, a practice that should be encouraged.

3.4.5. Groups and Derived from Associations

ODM provides the capability to associate observations into logical groups using the Groups and GroupDescriptions tables. Observation groups maintain association between related data values (e.g., all of the temperature observations from a single lake depth profile). Each observation group is identified by a group name and a list of all of the unique ValueIDs for the data values that make up the group. There is no limit to how many observation groups a data value may be associated with.

ODM also provides the capability to store derived quantities (e.g., discharge) and the observations (e.g., stage) from which they were derived. Raw observation values and values derived from raw observations are stored together in the central DataValues table, while the connection between each derived data value and its more primitive raw measurement is preserved in the DerivedFrom table. Derived values may be created by transforming data, for example transforming stage to discharge, or by simply creating a quality controlled data series from a raw data series. Derived values may be associated with one or many more primitive data values via the DerivedFrom table to, for example, identify the single gage height value used to estimate an instantaneous discharge value, or the 96 instantaneous discharge values at 15-minute intervals that go into an estimate of mean daily discharge. Preserving the relationships between data values and the values from which they were derived is important in maintaining the provenance of observations.

3.4.6. Qualifying Comments and Censored Data

Many observations are accompanied by comments that qualify how the data should be interpreted or used. These comments are important in stipulating the quality of the data or in flagging potential problems. For example, when sample holding times associated with a particular chemical analysis method are exceeded before a sample is analyzed, the resulting data may be suspect. Data qualifying comments are typically added to such observations by the laboratory that performs the analysis, and it is critical that these comments follow the data wherever they are used. To this end, each individual observation stored within ODM can be qualified by a text comment that describes limitations of, or information about, that observation that are required in interpreting its value and in evaluating its appropriateness for use.

Censored data, or data that are above or below a detection or quantitation limit, are another issue that must be dealt with in storing environmental observations. Within ODM, each individual observation can be qualified by a censor code that indicates whether the true value is greater than or less than the value that is reported. All other values are assumed to be not censored. ODM uses a convention similar to that used by the USGS of recording the censoring level (e.g., the detection limit or the quantitation limit) as the value, preserving this information for data analysis methods that require that the censoring level be known [e.g., *Helsel*, 1990].

3.4.7. Data Sources

Information about the organization responsible for collecting and analyzing the data is an important part of data provenance. ODM provides a link for each observation in the database to the Sources table that holds information about the organization that originally collected the data.

3.4.8. Controlled Vocabularies

A controlled vocabulary is a carefully selected list of words and phrases that is used to describe units of information or data. Each of the terms within a controlled vocabulary has a unique and unambiguous definition. ODM imposes controlled vocabularies on some fields within the data model for several reasons. First, the use of controlled vocabularies for elements such as variable and unit names eliminates the use of different terms for the same concept (e.g., “water temperature” vs. “temperature, water”) and resolves any associated ambiguity. Secondly, controlled vocabularies can improve the accuracy and performance of searches over fields that could otherwise contain repetitive or ambiguous terms. Additionally, controlled vocabularies form the basis of the metadata within ODM and provide specific language to describe characteristics of the data to aid in its identification, discovery, assessment, and management.

3.4.9. Data Series

In order to support common data discovery queries that identify which variables have been measured at which locations and for what time periods, we use the concept of a “data series” as an organizing principle within ODM. A data series is a set of observation values of a particular type (e.g., continuously measured water temperature or irregular, instantaneous observations of nitrate concentrations), measured at a single site by a single source using a single method. The ODM Series Catalog table maintains a list of all of the data series within the database and essentially performs for an ODM database what a card catalog does for a library. It enables users to search for the data they are looking for as well as providing them with enough information to retrieve the data from the database. This table was designed to satisfy many common data discovery queries such as “which variables have been collected at a particular site” or “which sites have data for a particular variable.” Evaluation of these common queries against the SeriesCatalog table rather than against the DataValues table, which holds all of the observation values, significantly simplifies and improves the performance of these queries and facilitates more efficient data discovery.

3.5. ODM Examples

The examples in the following sections demonstrate the capability of the ODM data model to store different types of point observations. The examples present selected fields and tables chosen to illustrate key capabilities of the data model. These examples are presented using table names and field names shown in Figure 3.1. For a more in depth listing of ODM examples and a data dictionary that describes in detail all of the tables and fields within ODM, readers are refereed to the ODM Design Specifications document [*Tarboton et al.*, 2007]. Additional resources, sample databases, and software applications for using ODM can be found on the CUAHSI HIS website (http://his.cuahsi.org).

3.5.1. Streamflow - Gage Height and Discharge

Figure 3.3 illustrates how both stream gage height measurements and the associated discharge estimates derived from the gage height measurements can be stored in ODM. Note that gage height in feet and discharge in cubic feet per second are both in the same data table but with different VariableIDs that reference the Variables table, which specifies the variable name, units, and other quantities associated with these data values. The link between VariableID in the DataValues table and Variables table is shown. In this example, discharge measurements are derived from gage height (stage) measurements through a rating curve. The MethodID associated with each discharge record references into the Methods table that describes this and provides a URL that contains metadata details for this method. The DerivedFromID in the DataValues table references into the DerivedFrom table that references back to the corresponding gage height in the DataValues table from which the discharge was derived.

3.5.2. Streamflow - Daily Average Discharge

Figure 3.4 shows excerpts from tables illustrating the population of ODM with both continuous discharge values and derived daily averages. Daily average streamflow is reported as an average of continuous 15 minute interval data values. The record giving the single daily average discharge with a value of 722 ft^3 s^{-1} in the DataValues table has a DerivedFromID of 100. This refers to multiple records in the DerivedFrom table, with associated ValueIDs 97, 98, 99, … 113 shown. These refer to the specific 15 minute discharge values in the DataValues table used to derive the average daily discharge. VariableID in the DataValues table identifies the appropriate record in the Variables table specifying that this is a daily average discharge with units of ft^3 s^{-1} from UnitsID referencing in to the Units table. MethodID in the DataValues table identifies the appropriate record in the Methods table specifying that the method used to obtain this data value was daily averaging.

3.5.3. Water Chemistry from a Profile in a Lake

Reservoir profile measurements provide an example of the logical grouping of data values and data values that have an offset in relationship to the location of the monitoring site. These measurements may be made simultaneously (by multiple instruments in the water column) or over a short time period (one instrument that is lowered from top to bottom). Figure 3.5 shows an example of how these data would be stored in ODM. The OffsetTypes table and OffsetValue attribute are used to quantify the depth offset associated with each measurement. Each of the data values shown has an OffsetTypeID that references into the OffsetTypes table. The OffsetTypes table indicates that for this OffsetType the offset is “Depth below water surface.” The OffsetTypes table references into the Units table indicating that the OffsetUnits are meters, so OffsetValue in the DataValues table is in units of meters depth below the water surface.

Each of the data values shown has a VariableID that in the Variables table indicates that the variable measured was dissolved oxygen concentration in units of mg liter^{-1}. Each of the data values shown also has a MethodID that in the Methods table indicates that dissolved oxygen was measured with a Hydrolab multiprobe. The combination of the variable name, units, and method are sufficiently general to describe what has been measured. Within the ODM controlled vocabularies, the convention is that the units remain generic, whereas the variable names are more specific. For example, “dissolved phosphorus as P” is a different variable name than “dissolved phosphorus as $PO_{4,}$” but the units of both are mg liter^{-1}.

Additionally, the data values shown are part of a logical group of data values representing the water chemistry profile in a lake. This is represented using the Groups table and GroupDescriptions table. The Groups table associates GroupID 1 with each of the ValueIDs of the data values belonging to the group. A description of this group is given in the GroupDescriptions table.

3.6. ODM Implementation

As part of the process of planning for a national network of environmental observatories, 11 Test Bed projects across the United States are focused on developing techniques and technologies for environmental observatories ranging from innovative application of environmental sensors to publishing observations data in common formats that can be accessed by investigators nationwide. The Test Bed sites are located in a range of environmental conditions from the high Sierra Nevada of California to urban Baltimore, Maryland. Investigators at each of the Test Beds are participating in the development and deployment of common hydrologic information system capability for publishing observations from each of the Test Beds. Because a common cyberinfrastructure is being adopted, it is enabling cross-domain analysis within individual Test Beds as well as cross-Test Bed sharing and analysis of data. More information about the Test Beds and the data being collected at each can be found at the following URL (http://www.watersnet.org/wtbs/index.html). The following sections describe how ODM is being used as the basis for the common cyberinfrastructure across the Test Bed sites and how the issues of heterogeneity in data syntax and semantics are being overcome.

3.6.1. Overcoming Syntactic Heterogeneity

Within each of the Test Beds, one barrier in publishing and making use of observational data has been heterogeneity in the syntax of the data. It has been observed, for example, that data downloaded from automated data loggers are formatted differently than data generated as a result of chemical analysis of water samples in a laboratory, and within the Test Beds, these are only two of a variety of data sources. In addition to these methodological inconsistencies, syntactic heterogeneity within the Test Beds has also been caused by a proliferation of different file types (e.g., ascii text files versus Microsoft Excel files), different file formats (e.g., cross-tab tables versus serial lists), as well as other differences that are, in general, a result of investigator preference. Individuals working at the Test Bed sites all have their own favorite software and file formats in which they choose to work.

ODM has overcome this syntactic heterogeneity by providing a common and encompassing database within which all of the observations, regardless of source, collection method, or original file type and format, can be stored along with their metadata. A variety of software tools have been developed for assisting with and automating the process of loading data into an ODM database. Once data have been loaded from their original format into an ODM database, they are syntactically similar and become available to analytical tools that exploit this format. For example, the WaterOneFlow web services are the main mechanism for publishing and exchanging observations between Test Beds. The WaterOneFlow web services, which have been built to extract data from an ODM database based on a user defined query and transmit it over the Internet, preserve the syntactic homogeneity achieved by loading data into ODM because the data are transmitted in a single format that is consistent across Test Beds.

3.6.2. Overcoming Semantic Heterogeneity

Semantic heterogeneity has been another barrier in the effective publishing and use of observational data that has been addressed within and across the Test Beds.

Semantic heterogeneity refers to the variety in language used to describe observations. Within the Test Beds, ODM has overcome two different types of semantic heterogeneity: (1) the language used to describe the names of observation attributes; and (2) the language used to encode observation attribute values. The first type is general, and is addressed through the standard table and field schema of ODM. For example, within ODM a monitoring location is called a “Site” and all Site attributes are stored in a table called “Sites.” In each ODM database, the table names and field/attribute names are consistent and so when investigator data are loaded into ODM they adopt a consistent language.

The second type of semantic heterogeneity is in the attribute values themselves. For example, within ODM, each variable has an attribute called “VariableName” that describes the variable that has been measured. Within the Test Beds, different investigators use different names for the same constituent (e.g., “water temperature” versus “temperature, water”). These differences are reconciled within ODM through the use of controlled vocabularies. Since the controlled vocabularies within ODM list the terms that are acceptable for use within many fields in the database, only one of the terms describing water temperature would be available in the ODM variable name controlled vocabulary and so when multiple datasets are added to an ODM database they are reconciled through the use of appropriate and consistent controlled vocabulary terms to describe the data. The ODM controlled vocabularies are dynamic and growing in that users can add new terms or edit existing terms by using the functionality on the ODM website (http://water.usu.edu/cuahsi/odm/).

3.6.3. A National Network of Consistent Data

By providing a new method for overcoming the syntactic and semantic heterogeneity in data being collected and published at each of the Test Bed Sites, ODM, along with the WaterOneFlow web services, has enabled a group of independent Test Bed investigators working on very different science problems to create a national network of published observational data that enables cross-domain and cross-Test Bed access to data. The advantages are clear: (1) consistent and fully specified data lead to higher quality analyses with less uncertainty; (2) the Test Bed network enabled by ODM is a new data resource for the scientific community; and (3) a standard method for publishing observational data means that the network can grow as more investigators publish their data.

3.7. Discussion and Conclusions

A data model for storing and managing environmental observations has been presented. The importance of metadata in describing environmental observations data cannot be overstated. It is critical that the data be carefully documented and annotated with metadata so that it can be unambiguously interpreted and used by investigators other than those that collected the data. The co-location of observational data and their associated metadata within a single, integrated ODM database enables easy and automated access.

The reliance of ODM on relational database technology provides several advantages. First, implementation of ODM within a relational database management system enables users to take advantage of the mature technology and advanced tools available in relational database systems. These include data import and export tools, a standardized, high level query language, and, more recently, tools for advanced data analysis and manipulation such as online analytical processing (OLAP), data mining, and data warehousing.

Next, ODM provides a framework in which data of different types and from disparate sources can be integrated. For example, data from multiple scientific disciplines can be assembled within a single ODM instance (e.g., hydrologic variables, water quality variables, climate variables, etc.). This has been the case at each site within a national network of environmental observatory Test Beds where publishing observational data using ODM and the WaterOneFlow web services has enabled both multi-disciplinary and cross-Test Bed access to a national network of consistent data.

The number of characteristics used to describe observations can potentially be large and different across data sources. One significant advantage of ODM is that, along with the observation values, it provides a place to store a standard set of the most commonly used attributes of environmental observations. As with any other model, this representation has some limitations. However, once assembled within ODM, observations can be presented in a consistent way – negating the need for users to learn the diverse data formats of multiple scientific communities. This can be useful when data from multiple disciplines need to be combined into a single analysis or simulation model.

Last, a consistent data model enables the standardization of software application development. These software tools include the WaterOneFlow web services, data loading and editing tools, and data visualization and retrieval tools. Readers are referred to the CUAHSI HIS website for details of these software applications (http://his.cuahsi.org). Thus, ODM supports a set of functions that are not available through simple file-based data publishing.

References

Ahern, T. (2004), Earth Scope US Array data management plan, report, Inc. Res. Inst. for Seismol. Data Manage. Cent., Seattle, Wash. (Available at http://www.iris.edu/USArray/publications/US_Data_Plan_Final-V7.pdf)

Bandaragoda, C. J., D. G. Tarboton, and D. R. Maidment (2005), User Needs Assessment, in *Hydrologic Information System Status Report, Version 1*, edited by D. R. Maidment, chap. 4, pp.48-87, Consorium of Univ. for the Adv. Of Hydrol. Sci., Washington, D. C. (Available at http://www.cuahsi.org/docs/HISStatusSept15.pdf)

Bandaragoda, C., D. G. Tarboton, and D. R. Maidment (2006), Hydrology's effort towards the Cyberfrontier, *EOS, Trans. AGU*, *87*(1), 2-6, doi:10.1029/2006EO010005.

Blöschl, G., and M. Sivapalan (1995), Scale Issues in hydrological modelling: A review, *Hydrol. Processes*, *9*, 251-290, doi:10.1002/hyp.3360090305.

Bose, R. (2002), A conceptual framework for composing and managing scientific data lineage, in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pp. 15-19, IEEE Press, Pascataway, N. J.

Bouganim, L., M. C. Cavalcanti, F. Fabret, M. L. Campos, F. Llirbat, M. Mattoso, R. Melo, A. M. Moura, E. Pacitti, F. Porto, M. Simoes, E. Simon, A. Tanaka, and P. Valduriez (2001), The Ecobase Project: Database and Web technologies for Environmental Information Systems, *SIGMOD Rec.*, *30*(3), 70-75, doi:10.1145/603867.603879.

Connolly, T., and C. Begg (2005), *Database Systems A Practical Approach to Design, Implementation, and Management*, 1374 pp., 4th ed., Addison-Wesley, Harlow, U. K.

Dewald, T. (2006), NHDPlus user guide. (Available at http://www.epa.gov/waters/NHDPlus%20Workshop/NHDPLUS_Documentation_20050822.pdf)

Environmental Data Standards Council (2006), Environmental sampling, analysis, and results data standards: overview of component data standards, *Stand. EX000001.1*, Environ. Data Stand. Counc. U. S. Environ. Rot. Agency, Washington, D.C. (Available at http://www.envdatastandards.net/files/693_file_ESAR_Overview_01_06_2006__Final_.pdf)

Goodall, J. L., J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky (2008), A first approach to Web services for the National Water Information System, *Environ. Model. & Software*, *23*(4), 404-411, doi:10.1016/j.envsoft.2007.01.005.

Gray, J., D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber (2005), Scientific data management in the coming decade, *SIGMOD Rec.*, *34*(4), 34-41, doi:10.1145/1107499.1107503.

Helsel, D. R. (1990), Less than obvious: Statistical treatment of data below the detection limit, *Environ. Sci. and Technol.*, *24*(12), 1766-1774, doi:10.1021/es00082a001.

Maidment, D. R., (Ed.) (2002), *Arc Hydro GIS for Water Resources*, 203 pp., ESRI Press, Redlands, Calif.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford (1997), Nongeospatial metadata for the ecological sciences, *Ecol. Appl.*, *7*(1), 330-342, doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

NASA, (2005), Committee on Data Management, Archiving, and computing (CODMAC) Data Level Definitions, http://science.hq.nasa.gov/research/earth_science_formats.html. [Last accessed January 23, 2008].

National Water Quality Monitoring Council (2006), Water quality data elements: A user guide, Tech. Rep. 3, Advis. Comm. On Water Inf., Washington D. C. (Available at http://acwi.gov/methods/pubs/wdqe_pubs/wqde_trno3.pdf)

Open Geospatial Consortium, Inc. (2006), Observations and measurements, OGC Best Practices Document, OGC 05-087r4, Version 0.14.7, Simon Cox, editor. (Available at http://www.opengeospatial.org/standards/bp).

Open Geospatial Consortium, Inc., (2007), CUAHSI WaterML, OGC Discussion Paper, OGC 07-041r1, Version 0.3.0, Zaslavsky, I., D. Valentine, and T. Whiteaker Editors. (Available at http://www.opengeospatial.org/standards/dp)

Pokorný, J. (2006), Database architectures: Current trends and their relationships to environmental data management, *Environ. Model. & Software*, *21*, 1579-1586, doi:10.1016/j.envsoft.2006.05.004.

Tarboton, D. G. (2005), Review of proposed CUAHSI Hydrologic Information System Hydrologic Observations Data Model, Utah State University, May 5, 2005,

(Available at http://www.engineering.usu.edu/cee/faculty/dtarb/HydroObsDataModelReview.pdf)

Tarboton, D. G., J. S. Horsburgh, and D. R. Maidment (2007), CUAHSI community Observations Data Model (ODM) design specifications document: Version 1.0, (Available at http://his.cuahsi.org/documents/ODM1.pdf)

Tomasic, A., and E. Simon (1997), Improving access to environmental data using context information, *SIGMOD Rec.*, *26*(1), 11-15, doi:10.1145/248603.248606.

Table 3.1. ODM Attributes Associated with an Observation

Attribute	Definition
Value	The observation value itself
Accuracy	Quantification of the measurement accuracy associated with the observation value
Date and Time	The date and time of the observation (including time zone offset relative to UTC and daylight savings time factor)
Variable Name	The name of the physical, chemical, or biological quantity that the value represents (e.g. streamflow, precipitation, water quality)
Location	The location at which the observation was made (e.g. latitude and longitude)
Units	The units (e.g. m or m^3/s) and unit type (e.g. length or volume/time) associated with the variable
Interval	The interval over which each observation was collected or implicitly averaged by the measurement method and whether the observations are regularly recorded on that interval
Offset	Distance from a reference point to the location at which the observation was made (e.g. 5 meters below water surface)
Offset Type/ Reference Point	The reference point from which the offset to the measurement location was measured (e.g. water surface, stream bank, snow surface)
Data Type	An indication of the kind of quantity being measured (e.g. an instantaneous or cumulative measurement)
Organization	The organization or entity providing the measurement
Censoring	An indication of whether the observation is censored or not
Data Qualifying Comments	Comments accompanying the data that can affect the way the data is used or interpreted (e.g. holding time exceeded, sample contaminated, provisional data subject to change, etc.)
Analysis Procedure	An indication of what method was used to collect the observation (e.g. dissolved oxygen by field probe or dissolved oxygen by Winkler Titration)
Source	Information on the original source of the observation (e.g. from a specific instrument or investigator 3rd party database)
Sample Medium	The medium in which the sample was collected (e.g. water, air, sediment, etc.)
Quality Control Level	An indication of the level of quality control the data has been subjected to (e.g., raw data, checked data, derived data)
Value Category	An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation

Table 3.2. Data Types that can be Represented Within ODM

Data Type	Description	Example
Continuous	The phenomenon, such as streamflow, Q(t) is specified at a particular instant in time and measured with sufficient frequency (small spacing) to be interpreted as a continuous record of the phenomenon.	Fifteen minute observations of discharge at a stream gage station.
Sporadic	The phenomenon is sampled at a particular instant in time but with a frequency that is too coarse for interpreting the record as continuous. This would be the case when the spacing is significantly larger than the support and the time scale of fluctuation of the phenomenon.	Infrequent water quality samples that characterize nutrient concentrations.
Cumulative	The data represents the cumulative value of a variable measured or calculated up to a given instant of time: $\mathrm{V(t)} = \int_0^t \mathrm{Q}(\tau)\mathrm{d}\tau$, where τ represents time in the integration over the interval [0,t].	Cumulative volume of flow or cumulative precipitation.
Incremental	The data value represents the incremental value of a variable over a time interval Δt: $\Delta V(t) = \int_t^{t+\Delta t} Q(\tau)d\tau$.	Incremental volume of flow or incremental precipitation.
Average	The data value represents the average over a time interval, such as daily mean discharge or daily mean temperature: $\overline{Q}(t) = \frac{\Delta V(t)}{\Delta t}$. The averaging interval is quantified by time support in the case of regular data and by the time interval from the previous data value at the same position for irregular data.	Daily mean discharge or daily mean air temperature.
Maximum	The data value is the maximum value occurring at some time during a time interval. ODM adopts the convention that the time interval is the time support for regular data and the time interval from the previous data value at the same position for irregular data.	Annual maximum discharge or daily maximum air temperature.
Minimum	The data value is the minimum value occurring at some time during a time interval. The time interval is defined similarly to Maximum data.	The 7-day low flow for a year or daily minimum air temperature.
Constant Over Interval	The data value is a quantity that can be interpreted as constant over the time interval from the previous measurement.	Discharge from a control structure that does not change unless a gate is moved or reset.
Categorical	The value stored is a numerical value that represents a categorical rather than continuous valued quantity. Each category is represented by a numeric value, and the mapping from numeric values to categories is stored in ODM.	Weather observations such as “Cloudy” or “Partly Cloudy.”

Table 3.3. Quality Control Levels in ODM

Level	Description	Example
0	Raw and unprocessed data and data products that have not undergone quality control. Depending on the variable, data type, and data transmission system, raw data may be available within seconds or minutes after the measurements have been made.	Real time precipitation, streamflow, and water quality measurements
1	Quality controlled data that have passed quality assurance procedures such as routine estimation of timing and sensor calibration or visual inspection and removal of obvious errors.	USGS published daily average discharge records following parsing through USGS quality control procedures.
2	Derived products that require scientific and technical interpretation and may include multiple-sensor data.	Basin average precipitation derived from rain gages using an interpolation procedure.
3	Interpreted products that require researcher driven analysis and interpretation, model-based interpretation using other data and/or strong prior assumptions.	Basin average precipitation derived from the combination of rain gages and radar return data.
4	Knowledge products that require researcher driven scientific interpretation and multidisciplinary data integration and include model-based interpretation using other data and/or strong prior assumptions.	Percentages of old or new water in a hydrograph inferred from an isotope analysis.

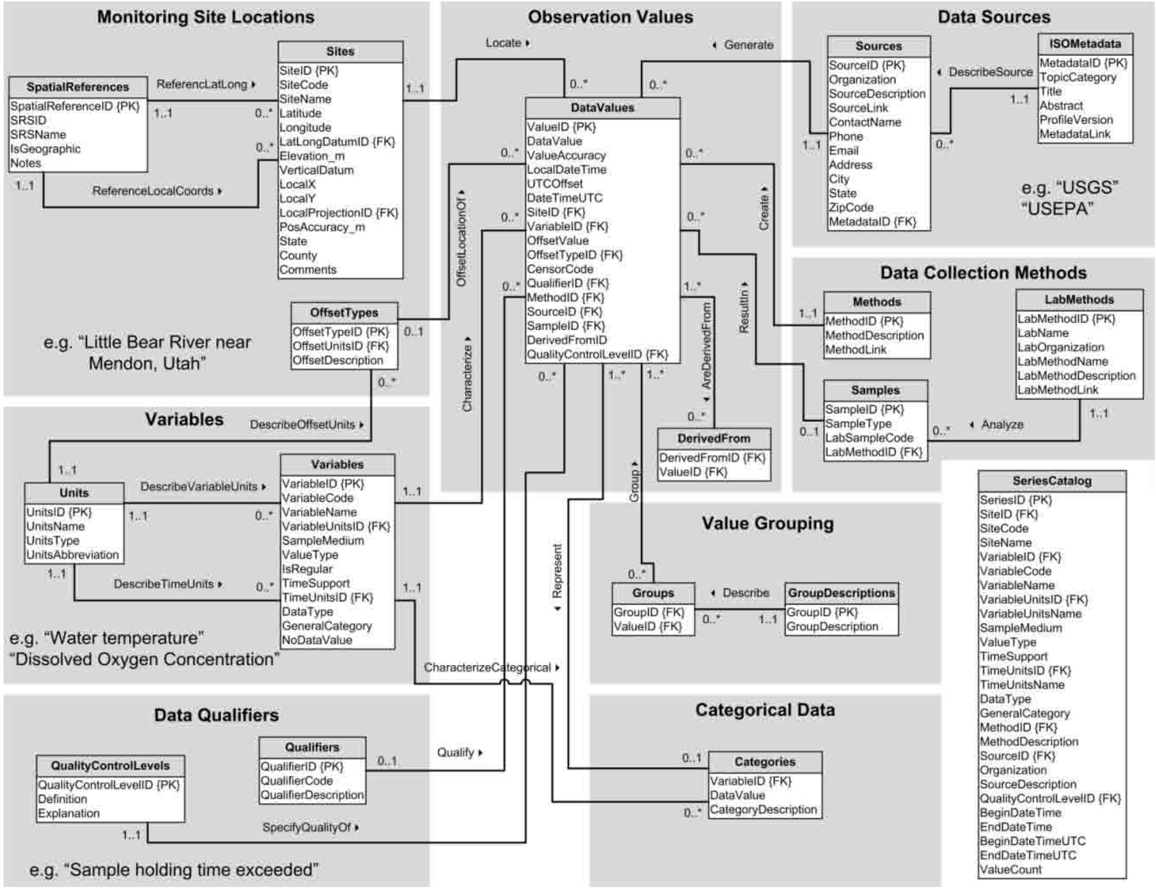


Figure 3.1. ODM logical data model. The primary key field for each table is designated with a {PK} label. Foreign keys are designated with a {FK} label. The lines between tables show relationships with cardinality indicated by numbers and labeled with the name and directionality of the relationship.

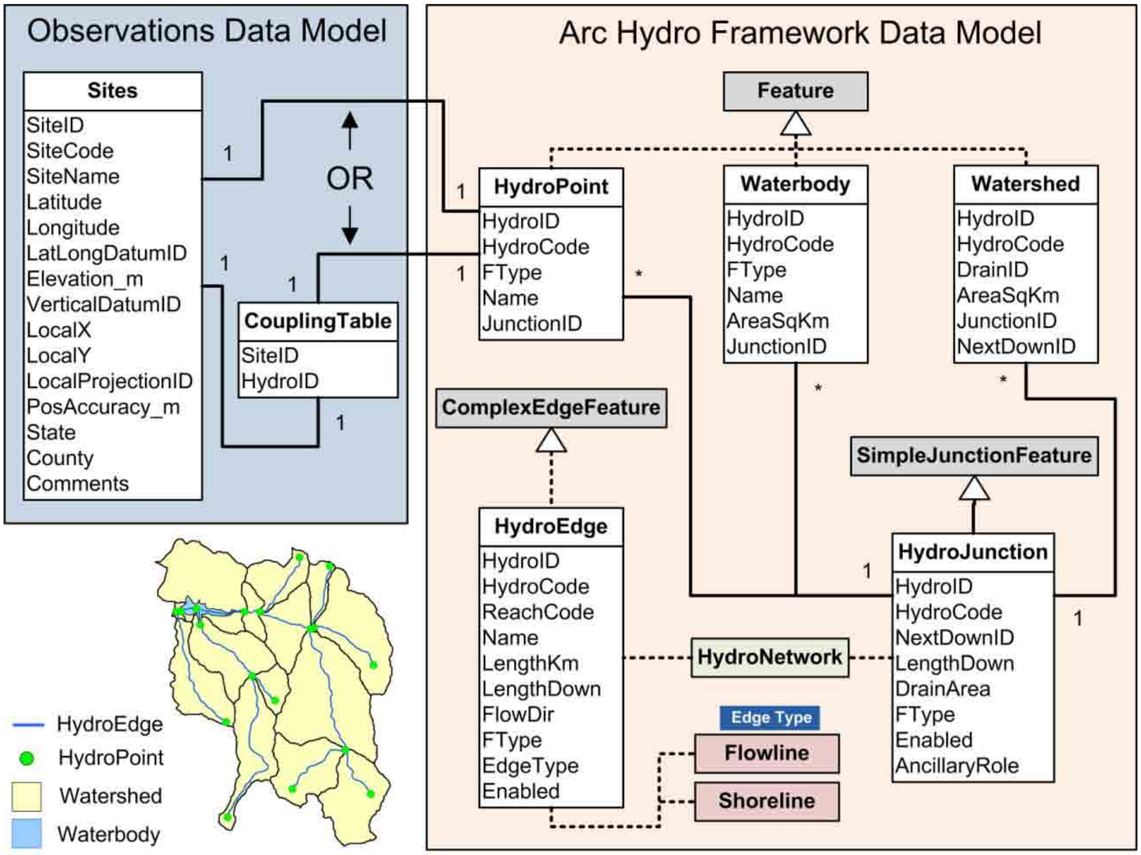
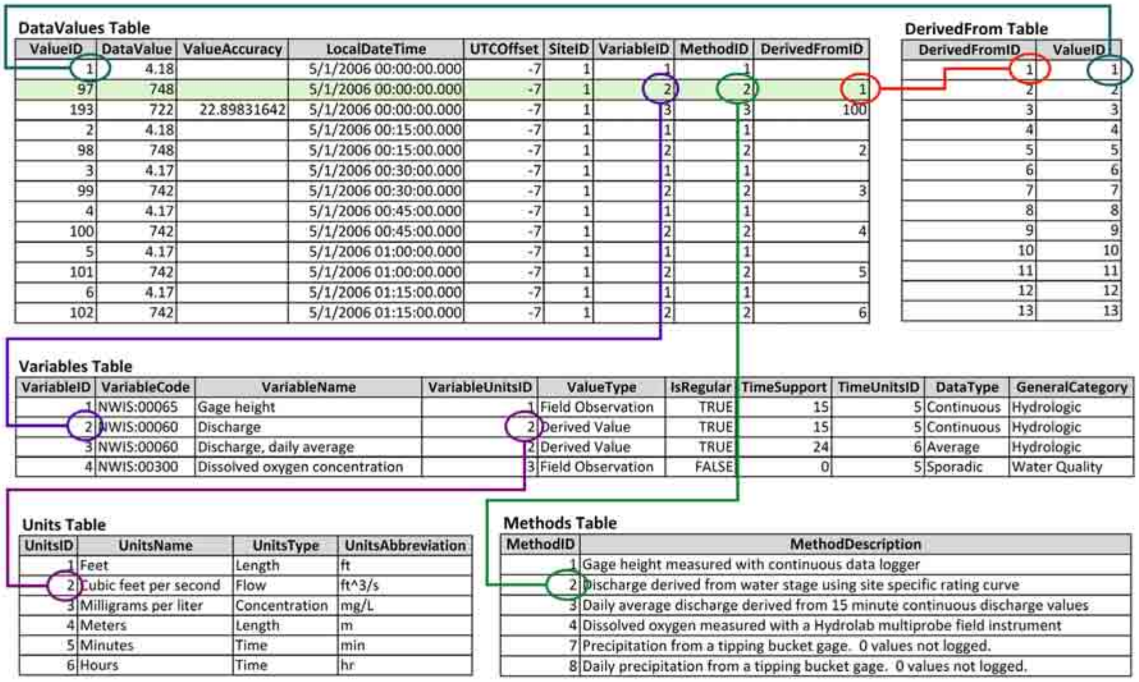


Figure 3.2. Arc Hydro Framework Data Model and Observations Data Model related through SiteID field in the Sites table.



DataValues Table

ValueID	DataValue	ValueAccuracy	LocalDateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		5/1/2006 00:00:00.000	-7	1	1	1	
97	748		5/1/2006 00:00:00.000	-7	1	2	2	1
193	722	22.89831642	5/1/2006 00:00:00.000	-7	1	3	3	100
2	4.18		5/1/2006 00:15:00.000	-7	1	1	1	
98	748		5/1/2006 00:15:00.000	-7	1	2	2	2
3	4.17		5/1/2006 00:30:00.000	-7	1	1	1	
99	742		5/1/2006 00:30:00.000	-7	1	2	2	3
4	4.17		5/1/2006 00:45:00.000	-7	1	1	1	
100	742		5/1/2006 00:45:00.000	-7	1	2	2	4
5	4.17		5/1/2006 01:00:00.000	-7	1	1	1	
101	742		5/1/2006 01:00:00.000	-7	1	2	2	5
6	4.17		5/1/2006 01:15:00.000	-7	1	1	1	
102	742		5/1/2006 01:15:00.000	-7	1	2	2	6

DerivedFrom Table

DerivedFromID	ValueID
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13

Variables Table

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
1	NWIS:00065	Gage height	1	Field Observation	TRUE	15	5	Continuous	Hydrologic
2	NWIS:00060	Discharge	2	Derived Value	TRUE	15	5	Continuous	Hydrologic
3	NWIS:00060	Discharge, daily average	2	Derived Value	TRUE	24	6	Average	Hydrologic
4	NWIS:00300	Dissolved oxygen concentration	3	Field Observation	FALSE	0	5	Sporadic	Water Quality

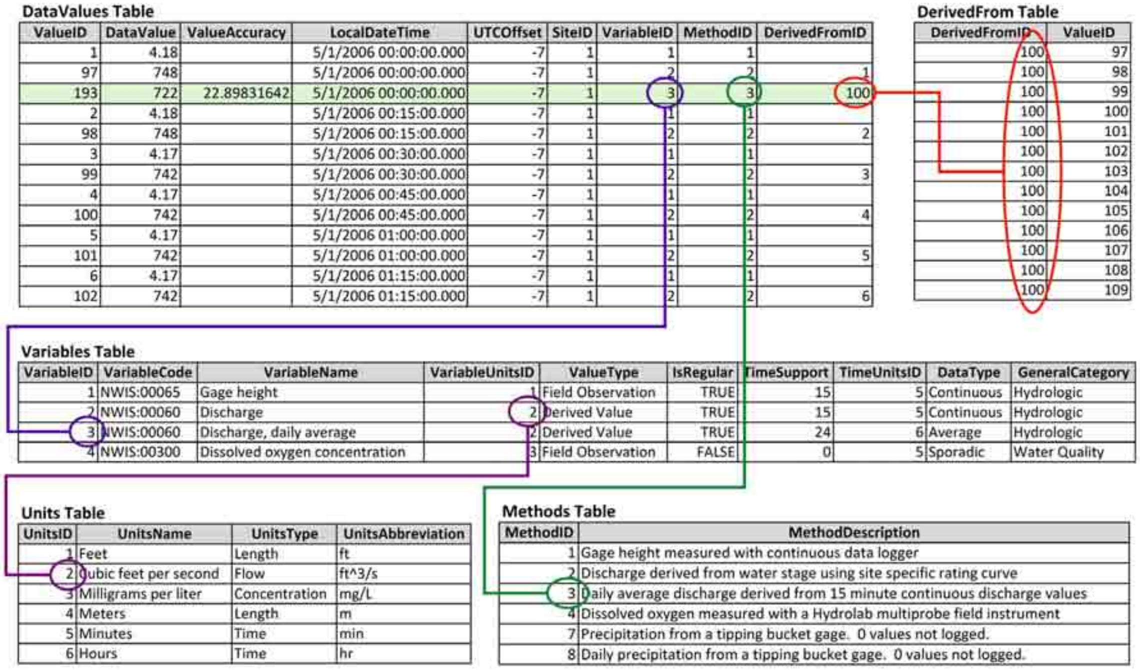
Units Table

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
1	Feet	Length	ft
2	Cubic feet per second	Flow	ft^3/s
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

Methods Table

MethodID	MethodDescription
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument
7	Precipitation from a tipping bucket gage. 0 values not logged.
8	Daily precipitation from a tipping bucket gage. 0 values not logged.

Figure 3.3. Excerpts from tables illustrating the population of ODM with streamflow gage height (stage) and discharge data.



DataValues Table

ValueID	DataValue	ValueAccuracy	LocalDateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		5/1/2006 00:00:00.000	-7	1	1	1	
97	748		5/1/2006 00:00:00.000	-7	1	2	2	1
193	722	22.89831642	5/1/2006 00:00:00.000	-7	1	3	3	100
2	4.18		5/1/2006 00:15:00.000	-7	1	1	1	
98	748		5/1/2006 00:15:00.000	-7	1	2	2	2
3	4.17		5/1/2006 00:30:00.000	-7	1	1	1	
99	742		5/1/2006 00:30:00.000	-7	1	2	2	3
4	4.17		5/1/2006 00:45:00.000	-7	1	1	1	
100	742		5/1/2006 00:45:00.000	-7	1	2	2	4
5	4.17		5/1/2006 01:00:00.000	-7	1	1	1	
101	742		5/1/2006 01:00:00.000	-7	1	2	2	5
6	4.17		5/1/2006 01:15:00.000	-7	1	1	1	
102	742		5/1/2006 01:15:00.000	-7	1	2	2	6

DerivedFrom Table

DerivedFromID	ValueID
100	97
100	98
100	99
100	100
100	101
100	102
100	103
100	104
100	105
100	106
100	107
100	108
100	109

Variables Table

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
1	NWIS:00065	Gage height	1	Field Observation	TRUE	15	5	Continuous	Hydrologic
2	NWIS:00060	Discharge	2	Derived Value	TRUE	15	5	Continuous	Hydrologic
3	NWIS:00060	Discharge, daily average	2	Derived Value	TRUE	24	6	Average	Hydrologic
4	NWIS:00300	Dissolved oxygen concentration	3	Field Observation	FALSE	0	5	Sporadic	Water Quality

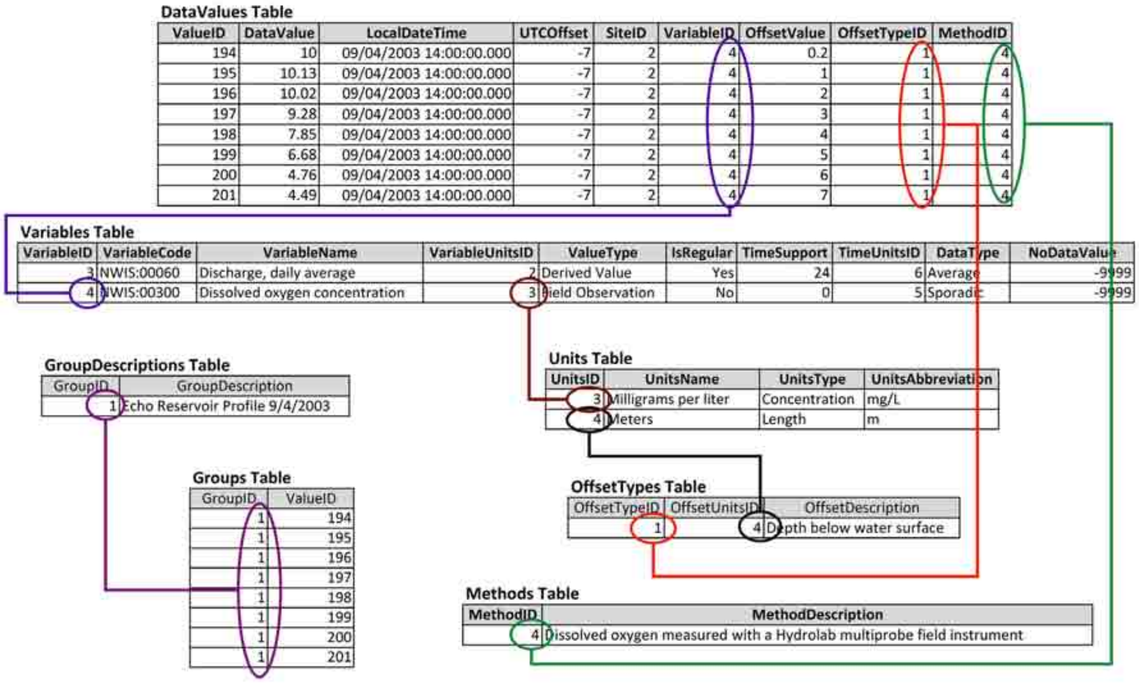
Units Table

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
1	Feet	Length	ft
2	Cubic feet per second	Flow	ft^3/s
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

Methods Table

MethodID	MethodDescription
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument
7	Precipitation from a tipping bucket gage. 0 values not logged.
8	Daily precipitation from a tipping bucket gage. 0 values not logged.

Figure 3.4. Excerpts from tables illustrating the population of ODM with daily average discharge derived from 15 minute discharge values.



DataValues Table

ValueID	DataValue	LocalDateTime	UTCOffset	SiteID	VariableID	OffsetValue	OffsetTypeID	MethodID
194	10	09/04/2003 14:00:00.000	-7	2	4	0.2	1	4
195	10.13	09/04/2003 14:00:00.000	-7	2	4	1	1	4
196	10.02	09/04/2003 14:00:00.000	-7	2	4	2	1	4
197	9.28	09/04/2003 14:00:00.000	-7	2	4	3	1	4
198	7.85	09/04/2003 14:00:00.000	-7	2	4	4	1	4
199	6.68	09/04/2003 14:00:00.000	-7	2	4	5	1	4
200	4.76	09/04/2003 14:00:00.000	-7	2	4	6	1	4
201	4.49	09/04/2003 14:00:00.000	-7	2	4	7	1	4

Variables Table

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	NoDataValue
3	NWIS:00060	Discharge, daily average	2	Derived Value	Yes	24	6	Average	-9999
4	NWIS:00300	Dissolved oxygen concentration	3	Field Observation	No	0	5	Sporadic	-9999

GroupDescriptions Table

GroupID	GroupDescription
1	Echo Reservoir Profile 9/4/2003

Units Table

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m

Groups Table

GroupID	ValueID
1	194
1	195
1	196
1	197
1	198
1	199
1	200
1	201

OffsetTypes Table

OffsetTypeID	OffsetUnitsID	OffsetDescription
1	4	Depth below water surface

Methods Table

MethodID	MethodDescription
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument

Figure 3.5. Excerpts from tables illustrating the population of ODM with water chemistry data from a profile in a lake.

CHAPTER 4

AN INTEGRATED SYSTEM FOR PUBLISHING ENVIRONMENTAL OBSERVATIONS DATA[1]

Abstract

Over the next decade, it is likely that science and engineering research will produce more scientific data than has been created over the whole of human history. The successful use of these data to achieve new scientific breakthroughs will depend on the ability to access, integrate, and analyze these large datasets. Robust data organization and publication methods are needed within the research community to enable data discovery and scientific analysis by researchers other than those that collected the data. We present a new method for publishing research datasets consisting of point observations that employs a standard observations data model populated using controlled vocabularies for environmental and water resources data along with web services for transmitting data to consumers. We describe how these components have reduced the syntactic and semantic heterogeneity in the data assembled within a national network of environmental observatory test beds and how this data publication system has been used to create a federated network of consistent research data out of a set of geographically decentralized and autonomous test bed databases.

[1] Coauthored by Jeffery S. Horsburgh, David G. Tarboton, Michael Piasecki, David R. Maidment, Ilya Zaslavsky, David Valentine, and Thomas Whitenack.

Submitted to Environmental Modelling & Software.

4.1. Introduction

New technology and data resources are often instrumental in the emergence of new scientific discoveries. Because results from local research projects can be aggregated across sites and times, in many cases by investigators other than those who originally collected the data, the potential exists to advance science and research significantly through the publication of research data [*Borgman et al.*, 2007; *Research Information Network*, 2008]. There is a need, therefore, for standardized and robust methods to organize and publish environmental observations data as resources that can be discovered and used for scientific analysis.

Indeed, environmental research and education have recently become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared, and analyzed [*National Science Foundation*, 2007]. Over the next decade, it is likely that science and engineering research will produce more scientific data than has been created over the whole of human history [*Cox et al.*, 2006]. Successfully using these data to achieve new scientific breakthroughs and increase understanding of the world around us, as well as in making sound and informed resource management decisions, will depend in large part on the ability to access, organize, integrate, and analyze these large datasets.

Comprehensive infrastructure that is being used to capitalize on dramatic advances in information technology has been termed “cyberinfrastructure” and integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools [*National Science Foundation*, 2007]. This paper describes new cyberinfrastructure that enables the publication of point observations (i.e., measurements made at a point in space such as a weather station or water quality monitoring site). This cyberinfrastructure has been developed as part of a Hydrologic Information System (HIS), which is a distributed network of data sources and functions that are integrated using web services and that provide access to data, tools, and models that enable synthesis, visualization, and evaluation of hydrologic system behavior (http://his.cuahsi.org). Although the data publication system described in this paper has been developed primarily to advance the information science knowledge base and available data resources for water resources research, the general system architecture could be extended to many other types of point observations.

The HIS consists of four major components: data publication, data curation, data discovery, and data delivery. Publication is the process by which data are made available to users other than those that collected the data. Curation is the long term preservation of data to ensure that they persist indefinitely. Discovery involves tools that allow users to find published data, and delivery involves the transmittal of data to users in formats that they can use. In this paper, we focus mainly on the data publication component, although we include some discussion of the other components to place data publication in the context of the overall HIS.

Publication of research data involves persistent storage, management, and communication of data to potential users. Within and across research sites, multiple investigators and organizations are involved in both collecting and consuming data. To be effective, data publication systems must facilitate interoperation and mediation among data sources and their consumers. One challenge that arises in the design of data publication systems is heterogeneity within the formats and vocabularies that support the data [*Sheth and Larson*, 1990; *Colomb*, 1997; *Morocho et al.*, 2003]. Additionally, data consumers may not have intimate knowledge of the data collection process, requiring that the data be published with sufficient metadata to enable unambiguous interpretation [*Gray et al.*, 2005]. These metadata should include information about the location at which the observations were made, the variable that was observed or measured, the source of or organization that created the data, the procedures used to create the data, data qualifying comments, quality assurance and quality control information, time support, spacing, and extent, and other important attributes [Chapter 3].

In this paper, we describe a data publication system that overcomes the challenges in publishing research data through the use of a standard observations data model populated using controlled vocabularies for environmental and water resources data along with web services for transmitting data to consumers. Section 4.2 describes existing data publication efforts for environmental and water resources data. Section 4.3 describes syntactic and semantic heterogeneity and their implications for the publication, search for, and interpretation of existing environmental and water resources data. Section 4.4 describes how this heterogeneity can be overcome. Sections 4.5 and 4.6 provide an implementation case study that describes the components of the data publication system and how it has been used to create a federated network of consistent research data out of a set of geographically decentralized and autonomous databases from 11 environmental observatory test beds, effectively creating a publically-available, community data resource from data that might otherwise have been confined to the private files of the individual investigators.

4.2. Existing Data Publication Methods

Within the United States, many organizations and individuals measure hydrologic variables such as streamflow, water quality, groundwater levels, soil moisture, and precipitation. Several national data collection and publication networks operated by government agencies have arisen over the years. These include the USGS WATer Data STOrage and REtrieval System (WATSTORE), which has been replaced by the National Water Information System (NWIS) (http://waterdata.usgs.gov/nwis), the USEPA STOrage and RETrieval (STORET) System (http://www.epa.gov/storet/), the USDA SNOpack TELemetry (SNOTEL) System (http://www.wcc.nrcs.usda.gov/snow/) and Soil Climate Analysis Network (SCAN) (http://www.wcc.nrcs.usda.gov/scan/), the NOAA National Climatic Data Center (NCDC) (http://www.ncdc.noaa.gov/oa/ncdc.html), and a host of others. These national data repositories contain a wealth of data, but, in general, they have different data storage systems and formats, different data retrieval systems, and different data publication formats. Synthesizing data from these disparate sources into a single analysis can be difficult because each one presents users with the task of navigating through pages, menus, and files to access the data and metadata that they contain.

Recent times have also seen a push in the publication of data from existing experimental watersheds such as Reynolds Creek [*Slaughter et al.*, 2001], the Little River [*Bosch et al.*, 2007], and Walnut Gulch [*Moran et al.*, 2008; *Nichols and Anson*, 2008]. The technical details and much of the metadata for these datasets have been described in journal publications, and the data themselves have been made available as files that can be retrieved from public websites. Similarly, the Long Term Ecological Research

(LTER) Network has made climatic and hydrologic data collected at LTER sites available through their ClimDB/HydroDB climate and hydrology database projects website (http://www.fsl.orst.edu/climhy/).

Although these efforts represent considerable progress, none of the data publication systems that have been developed have been embraced as a standard for the academic and scientific research communities. Because of this, data and metadata resulting from academic research in water resources continue to be published in peer-reviewed journals [*Helly*, 2006]. Interpretations and figures based on data are widely published and archived in libraries, while most of the primary data are confined to the research files of the investigators, making verification of research results difficult. More recently, however, the idea of publishing observational data along with analysis results is gaining ground within the research community as the technology for doing so becomes more generally accessible [*Research Information Network*, 2008].

4.3. Syntactic and Semantic Heterogeneity in Environmental and Water Resources Data

Syntactic heterogeneity refers to a difference in how data and metadata are organized (e.g., rows vs. columns) and encoded (e.g., text files versus Excel spreadsheets), while semantic heterogeneity refers to the variety in language and terminology used to describe observations. Syntactic heterogeneity arises where there are methodological inconsistencies. For example, data downloaded from automated data loggers are generally encoded as delimited text files, whereas data generated as a result of chemical analysis of water samples in a laboratory may be entered by hand from a hard-copy laboratory report into an Excel spreadsheet. In addition to these methodological differences, different software applications have given rise to the proliferation of different file types and formats.

Semantic heterogeneity occurs when there is disagreement about the meaning, interpretation, or intended use of the same or related data [*Sheth and Larsen*, 1990]. Among observational data, this heterogeneity can be generalized into two types: 1) structural – i.e. the language used to describe the names of observation attributes; and 2) contextual – i.e. the language used to encode observation attribute values. Structural heterogeneity begs the questions – what are the common attributes of environmental observations, and what should those attributes be called? For example, should the location at which an observation was made be called a “monitoring site” or a “station?” Should the measured quantity be called a “variable” or a “parameter?” This type of semantic heterogeneity is structural because it determines the structure of any model that is used to represent the data.

Contextual heterogeneity lies in the attribute values themselves. For example, one attribute of scientific observations is the name of the variable that was measured. It is common for different investigators to use different names for the same variable (e.g., “discharge” versus “streamflow”), or the same name for different variables (e.g., using a single term “temperature” to represent both air temperature and water temperature). Many of the semantic differences that arise in research datasets are a result of investigator preference and inconsistencies among scientific domains. Table 4.1 provides examples of semantic heterogeneity in data from two popular water resources data sources and demonstrates both structural and contextual semantic heterogeneity.

The implications of syntactic and semantic heterogeneity in publishing environmental observations data are threefold – first in users finding the data, second in decoding and organizing the data, and third in interpreting them. Within water resources research, data are available from many different sources that use different nomenclature, storage technologies, user interfaces, and even languages, making data discovery a difficult and time consuming task [*Beran and Piasecki*, 2008]. Data discovery is an important aspect of the cyberinfrastructure required to support publication of research data because scientists' ability to find, decode, and interpret available datasets will determine how or if the data are used for scientific analyses. Performance of queries and search mechanisms for data discovery can be significantly improved when syntactic and semantic heterogeneity among datasets is overcome [*Madin et al.*, 2007; *Beran and Piasecki*, 2008]. After data are discovered, much research time and effort (up to 50% or more) is spent decoding, manipulating, and organizing observational data into a format that is useful [*Bandaragoda et al.*, 2005; *Ramachandran et al.*, 2005; *Ruddell and Kumar*, 2006]. This process is also error prone. Specialized knowledge and expensive software may be required to handle files in different formats from disparate sources.

Serious errors in data use and interpretation can result from semantic heterogeneity in data from different sources. This was spectacularly demonstrated when navigators of NASA's $125 Million Mars Climate Orbiter sent the space craft off course to its eventual loss because they assumed that data used to compute the effects of thruster firings on the trajectory of the spacecraft were in metric units when they were in fact in English units [*Mars Climate Orbiter Mishap Investigation Board*, 1999]. *Madnick and Zhu* [2006] use this example as well as many others to describe how many perceived data quality problems are actually data misinterpretation problems that result from semantic heterogeneity. It is critical, therefore, that data are published with sufficient metadata so that they can be unambiguously interpreted.

4.4. Overcoming Heterogeneity

Reconciling heterogeneity in data from different sources, which may be required both within and across research sites, is a complex problem that has a long history in information science [*Colomb*, 1997; *Bergamaschi et al.*, 2001; *Cox et al.*, 2006]. This challenge is fueling much of the movement toward using standardized markup languages as self-describing, common data formats that can be used by data producers and data consumers. Examples include Earth Science Markup Language (ESML) [*Ramachandran et al.*, 2005], Ecological Metadata Language (EML) [*EML Project Members*, 2008], Water Markup Language (WaterML) [*Zaslavsky et al.*, 2007], and the Open Geospatial Consortium's (OGC) Observations and Measurements (O&M) [*Cox*, 2006]. Other methods that have been used for this task include the use of standard data models, controlled vocabularies, and ontologies. In evaluating these methods, an important distinction must be made between technologies for data communication (i.e., the formats and mechanisms used to transmit data to consumers) and technologies for persistent data storage and management (i.e., the formats and mechanisms used by the data source for long term storage and management). Approaches for handling heterogeneity within these two distinct data publication tasks can be quite different, but both should be addressed in the publication of research datasets.

Existing published data sources such as NWIS, NCDC, and STORET provide a good example of the data publication problem. Data stored within these systems hold much value for scientific research, but each has its own autonomous methods for storing, managing, and communicating its data. Providing consistent access to the datasets from each of these federal data providers is important in leveraging these data for scientific research, but it requires mediating across the different data formats and vocabularies of each of these systems. Overcoming heterogeneity in these existing data repositories is mainly an issue of data communication (i.e., can the data from each of these systems be provided to users in a format that is syntactically and semantically similar regardless of their source?) because the data sources do not have the same underlying persistent storage or data communication mechanisms.

Standardized markup languages such as ESML, EML, WaterML, O&M, and others provide a structured syntax for communicating data from multiple sources as eXtensible Markup Language (XML) documents. These markup languages can be used to transmit data in a format that resolves syntactic heterogeneity, but they generally do not place semantic constraints on the meanings of the document contents. Recognizing this, scientists have begun to use ontologies in concert with these markup languages to overcome semantic heterogeneity in scientific data [*Lin and Ludäscher*, 2003; *Madin et al.*, 2007; *Beran and Piasecki*, 2008]. A domain ontology defines the terms used to describe and represent an area of knowledge and that are used by people, databases, and applications that need to share domain information [*Heflin*, 2004]. Ontologies can be implemented as structured, machine-interpretable vocabularies that include definitions of basic concepts in a domain and the relationships among them, thus capturing the semantics of the data that they represent.

Within a scientific domain, ontologies can provide a conceptual view of data stored within a variety of databases, and, because they can be formalized into machine-interpretable forms, they are powerful tools for virtually integrating disparate data sources without replicating the data or changing its persistent storage mechanism. For example, *Beran and Piasecki* [2008] describe an ontology-aided search engine called Hydroseek (http://www.hydroseek.org) that was specifically designed to mediate across the disparate formats and vocabularies of several national hydrologic data providers and provide users with a single interface to query and retrieve consistently formatted data from each of these data repositories. Hydroseek does not replicate or store the data from each of these repositories; it simply retrieves data from its source and communicates it to a user in a consistent format. Hydroseek's data discovery mechanism is based on an ontology that stores the vocabulary terms (e.g., variable names) from each of the data sources and the relationships between them so that a search using a single term such as "discharge" can return results from multiple data sources, even if some of those data sources use a different but equivalent term such as "streamflow" to describe their data. One significant barrier in using this approach, however, is that constructing the ontology that mediates across the vocabularies used by each data source is a difficult task that is prone to error because the mapping of terms from one source to another must be done by people who know how to interpret both vocabularies and there isn't always a one-to-one translation or mapping of terms.

Because the underlying data formats, vocabularies, and communication mechanisms of existing national data sources are different for each source, tools such as standardized markup languages and ontologies are needed to mediate across the sources and provide consistent access to the data. Unlike existing national data networks, however, most research datasets have not been formally published, they have not adopted standard methods for either persistent data storage or for data communication, and they have not settled on a specific vocabulary or format that define the syntax and semantics of the data. The opportunity exists, therefore, for the community of scientists collecting environmental and water resources data to build and adopt common data models and common vocabularies to describe the observations data for both storage and management and communication of data that are collected. A standardized data publication system can be used to resolve heterogeneity in existing datasets, both at the storage and communication levels, and to prevent heterogeneity in data to be collected in the future. Obviously, the easiest way to resolve heterogeneity is for it to never exist in the first place.

In the following sections, we present a case study for publishing point observations data that have been collected at 11 environmental observatory test beds in the United States. The observatory test beds represent a specialized case of the more general research data publication problem. This case study demonstrates the components of the general data publication system, how they address persistent storage, management, and communication of the data, and how they have been used to resolve semantic and syntactic heterogeneity in data collected both within and across test beds.

4.5. A Case Study for Publishing Point Observations Data

Leaders within the science and engineering research communities believe that new data networks, field observations, and field experiments that recognize the spatial and temporal heterogeneity of hydrologic processes are needed to address complex and encompassing questions and advance the science of hydrology [*Woods et al.*, 2001; *Kirchner*, 2006; *Hart and Martinez*, 2006]. In order to address these needs, a network of environmental observatories, which are integrated real-time observing systems that seek to improve understanding of the earth's water and biogeochemical cycles across multiple spatial and temporal scales, has been proposed for the United States under the premise that knowledge of the physical, chemical, and biological mechanisms controlling water quantity and quality is limited by lack of observations at the necessary spatial density and temporal frequency needed to infer the controlling processes [*Montgomery et al.*, 2007].

As part of the process of planning for this network, 11 test bed projects, which are part of the Water and Environmental Research Systems (WATERS) Network (http://www.watersnet.org) and are located across the United States, are focused on developing techniques and technologies for environmental observatories ranging from innovative application of environmental sensors to publishing observations data in common formats that can be accessed by investigators nationwide. Investigators at each of the test beds are participating in the development and deployment of common hydrologic information system capability for publishing observational data from each of the test beds. A common cyberinfrastructure is being adopted, with goals of enabling cross-domain analysis within individual test beds as well as cross-test bed sharing and analysis of data. More information about the test beds and the data being collected at each can be found at the following URL (http://www.watersnet.org/wtbs/index.html).

Data collection within the test beds is occurring at a variety of spatial and temporal scales, spanning different scientific investigators and domains, and across a variety of different locations and watersheds. Because of this, heterogeneity has emerged within the datasets that have been collected, especially from one test bed to the next. The following sections describe the components of the data publication system for the test beds as well as how the heterogeneity within test bed datasets has been reduced. Figure 4.1 shows the general architecture of the test bed data publication system and describes the step-by-step process for publishing data. Data collected in the field using in situ sensors or other sampling techniques are stored in a variety of differently formatted files. Data from these files are loaded into a database with special attention given to populating the metadata using controlled vocabularies. Next, web services are implemented to make the data in the database available over the Internet. Last, the address of the web services is registered with a central registry, effectively announcing the availability of the data to the public and enabling data discovery tools like Hydroseek, which provide map and context based search capabilities, to consume the data.

4.5.1. A Data Model for Environmental and Water Resources Data

The test beds have adopted the Observations Data Model (ODM) [Chapter 3] as a common model for storing and managing their observational data. ODM is a relational model that is implemented within a Relational Database Management System (RDBMS) and that defines the persistent structure of the data, including the set of attributes that accompany the data, their names, their data type, and their context. Each of the test beds has created one or more ODM databases into which they have loaded their point observations data. Each ODM database contains observational data for a variety of different variables collected at a set of monitoring sites. The data being collected differs from one test bed to the next, but examples of data that are being loaded into ODM databases include: discharge and water quality variables such as water temperature, dissolved oxygen concentration, and turbidity; samples of water quality constituents such as nutrients and sediment; groundwater levels and quality; and meteorological variables such as precipitation, air temperature, and solar radiation. Additionally, some of the test bed investigators are publishing data collected by other local agencies and organizations.

The use of ODM as the persistent data storage mechanism has two significant advantages. First, ODM addresses the syntactic heterogeneity in the data (i.e., different file types, data formats, etc.) collected both within and across test bed sites. By loading data into an ODM database, data managers at each of the test beds ensure that their data are syntactically similar to the data at all of the other test beds. Second, because ODM defines the attributes that accompany the data and their context, loading the test bed data into ODM overcomes any structural semantic heterogeneity in the test bed data.

4.5.2. ODM Controlled Vocabularies

Contextual semantic heterogeneity within and across the test bed datasets has been reduced through the use of controlled vocabularies for many of the attributes within ODM. Multiple datasets added to an ODM database are reconciled through the use of appropriate and consistent controlled vocabulary terms to describe the data. Since the controlled vocabularies within ODM list the terms that are acceptable for use within many fields in the database, data managers choose from the list of acceptable terms when loading data into the database rather than using their own, potentially inconsistent terms. While this places a burden on the data managers to select the appropriate controlled vocabulary terms, the advantage is that the terms in the ODM controlled vocabularies are unique and devoid of ambiguity (i.e., only a single term exists in a controlled vocabulary for each concept described). Figure 4.2 provides an example of how contextual heterogeneity in attributes of datasets from multiple investigators is reconciled through the use of the ODM controlled vocabularies.

Resolving the contextual heterogeneity in datasets using the ODM controlled vocabularies ensures that datasets are consistently described within each ODM database. In addition, it assures that datasets are consistently described across ODM databases (i.e., across test beds). The controlled vocabularies form the basis of the metadata within ODM and provide specific language to describe characteristics of the data to aid in its identification, discovery, assessment, and management.

4.5.3. Controlled Vocabulary System Implementation

A master list of approved controlled vocabulary terms is maintained within a central database. This central repository represents a community vocabulary for describing environmental and water resources data in that it was developed by the community of researchers working within the test beds. It is dynamic and growing; users can add new terms or edit existing terms by using the functionality available through the HIS website (http://his.cuahsi.org). If a data manager cannot find an appropriate term to describe data that is being added to an ODM database, he or she can navigate to the HIS website and use an online form to request addition of an appropriate term to the master controlled vocabulary. The ODM controlled vocabulary submission system (Figure 4.3) is moderated to ensure that submitted terms are appropriate, unique, and unambiguous. Once a new term is accepted, it becomes part of the master database.

The ODM controlled vocabularies are duplicated within each ODM database to maintain the integrity of data and to ensure that data loaded into local databases are connected with the required metadata. Because of this, and because new terms are continually being added to the master list, local databases must be synchronized periodically with the master repository to ensure the availability of the controlled vocabulary terms within each local database. This is accomplished through a software application called ODM Tools and the ODM Controlled Vocabulary web services.

Web services are applications that provide the ability to pass information between computers over the Internet, usually formatted using a platform independent markup language such as XML [*Goodall et al.*, 2008]. The ODM Controlled Vocabulary web services are implemented on top of the master controlled vocabulary repository database and broadcast the terms within the master repository in XML format. Data managers at each of the test beds can use functionality within the ODM Tools application to compare their local controlled vocabulary with the master repository and download any updated or added terms. ODM Tools gets the controlled vocabulary terms from the local database, accesses the ODM Controlled Vocabulary web services and automatically parses the XML messages that are returned, and then presents a tabular, side-by-side comparison of local and master terms. Users can then compare the terms in their local database with those in the master list and add any new or updated terms to their local database. Figure 4.3 shows this interaction between the data manager, the ODM Tools application, and the ODM Controlled Vocabulary web services, and Figure 4.4 shows how the master ODM controlled vocabulary repository serves the ODM databases located at each of the test beds.

4.5.4. WaterOneFlow Web Services

The main mechanism for communicating test bed observational data to users is the WaterOneFlow web services. The WaterOneFlow web services respond to user queries and transmit data extracted from an ODM database encoded using WaterML [*Zaslavsky et al.*, 2007]. The WaterOneFlow web services preserve the semantic and syntactic homogeneity achieved by loading data into ODM because the data are transmitted over the Internet in a single format using a vocabulary that is consistent across test beds. They also promote the interoperability of the data through the use of standard web services protocols and XML formats that are platform and programming language independent.

User queries are performed by calling methods that are exposed by the web services, such as *GetSites* for returning a list of sites within an ODM database along with the metadata for each site, *GetVariableInfo* for returning a list of variables within an ODM database along with the metadata for each variable, *GetSiteInfo* for returning a list of variables with data at a site, and *GetValues* for returning the time series of data for a site and variable combination. The web service methods can be called from many different programming languages and other software applications, including Microsoft

Visual Basic, Microsoft Excel, MatLab, and others from anywhere an Internet connection is available. Using the web services, users can discover the data that they are interested in and then access it using the analysis software of their choice, rather than being forced to learn a new analysis system. The service oriented architecture used by the HIS and represented by the WaterOneFlow web services serves to get the browser out of the way for data acquisition, thus enhancing environmental analysis and modeling capabilities through direct access to remote data sources from a wide range of software environments.

The WaterOneFlow web services are designed to be implemented on top of individual ODM databases so that the web services for each ODM database can be uniquely addressable. Each set of web services implements the same set of methods and returns data in the same format, but receives a unique URL for accessing the data in its underlying database. Because of this, users need only change the URL when accessing data from multiple ODM databases via the WaterOneFlow web services. The WaterOneFlow web services for ODM are also consistent with WaterOneFlow web services that have been developed for the USGS NWIS system, the USEPA STORET system, and other national hydrologic data providers. This means that data consumers can access the test bed data and data from national providers using a consistent set of methods, and data are returned in the same format from all of these sources.

4.5.5. Central Web Services Registry

Once data have been loaded into an ODM database and the WaterOneFlow web services have been implemented on top of that database, the data can be accessed over the Internet. However, making the data available on the Internet does not necessarily mean that they are easily discoverable. Because of this, the data publication process is not complete until the address of the web services has been registered with a central repository that stores links to each of the web services that make up the research data network and some metadata about each. The central web services registry is essentially a digital card catalog – it stores enough information about each of the databases and web services to know what they contain and how to access them, but it does not contain the published data. Users can navigate to the central web services registry from http://his.cuahsi.org and browse through the list of registered web services to determine which data are available. They can then query individual web services to get more detailed metadata and download the data.

Registering web services with the central registry also ensures that the data are available to centralized discovery, delivery, visualization, and analysis tools that have been developed as part of the HIS. For example, the Hydroseek application that was described previously has the capability to discover and deliver all of the data within databases and web services registered with the central registry. Simple keyword searches within Hydroseek return results from test bed databases alongside data from other national data providers, and the data from all of these sources is delivered in a consistent and easy to use format.

4.6. A National Research Data Network

ODM, the ODM controlled vocabulary system (i.e., the ODM CV website, ODM CV web services, and ODM Tools), the WaterOneFlow web services, and the central web services registry together form a data publication system that has enabled a group of independent test bed investigators working on very different science problems to publish their data within a network of syntactically and semantically similar scientific data. Not only are the data from each test bed available as a resource for the scientific community, but they are published in a way that cross-test bed access to and analysis of data is possible.

A snap-shot summary of the data published within the research data network, which now includes data from the test beds and other external data sources that have joined the network, is provided in Table 4.2 and Figure 4.5. The statistics for the research data network were compiled using Visual Basic code that was written to call each of the published web services and compile an overall list of sites and variables, along with a summary of the observations for each site and variable combination. Table 4.2 lists statistics for the entire network of research sites, and Figure 4.5 shows the number of monitoring sites, variables, and data values collected at each research site that has been added to the network. In Figure 4.5, each dot on the map represents an ODM database with a corresponding set of WaterOneFlow web services. The dots are plotted at the location of the average latitude and longitude of all of the monitoring sites stored in the ODM database.

The numbers in Table 4.2 and Figure 4.5 represent a snap-shot in time because new sites, variables, and data values are continually being added to the research data network. The following definitions apply for Table 4.2 and Figure 4.5: a data source is the organization that collected the data; a monitoring site is a location at which data are collected and is identified by its latitude and longitude coordinates; a variable is characterized by the combination of its name (e.g., temperature), the medium in which it was sampled (e.g., surface water), how the measurement was obtained (e.g., field observation), the time support interval over which the observation was made (e.g., hourly), its data type (e.g., average), and the method used to make the measurements (e.g., the type of temperature sensor used); and a data value is a single observation of a single variable at a single site on a particular date and time (e.g., the dissolved oxygen concentration at site x was 8.3 mg L^{-1} on April 7, 2008 at 3:00 PM).

4.7. Discussion and Conclusions

A standard method for publishing environmental and water resources point observations data has been presented. It provides a framework in which data of different types and from disparate sources can be integrated, while overcoming the syntactic and semantic heterogeneity in the data from each source. This has been the case at each site within a network of environmental observatory test beds in the United States, where publishing observational data using this system has enabled a group of independent test bed investigators working on very different science problems to create a network of syntactically and semantically similar scientific data. The research data network now contains over 3,700 data collection sites, nearly 800 measured variables, and nearly 42 million individual data values. The data publication system's flexibility in storing and enabling public access to similarly formatted data and metadata from multiple scientific domains and research sites has created a community data resource from data that might otherwise have been confined to the private files of the individual investigators.

Much of the success of the data publication system can be attributed to the federation of the individual databases. Each of the test beds maintains their own databases, and each is ultimately in charge of which data get published. Some have chosen to publish raw sensor data as it streams into their ODM database from field based sensors. Some have chosen to publish only data that have undergone quality control procedures. ODM stores data qualifying comments and information about the level of quality control data have been subjected to, and the WaterOneFlow web services transmit this information to ensure that users are aware of the quality and limitations of the data. Issues of data editing and cleansing, metadata population, data aggregation, and derived data generation are left to the data collectors who are most familiar with their datasets.

A significant challenge associated with this distributed data storage approach is that resources and expertise are required to implement the publication tools at each local research site. The data publication system requires a server on which an ODM database and a set of WaterOneFlow web services has been implemented. The server must be capable of hosting web applications, but does not have to be an expensive machine. Expertise with server administration, relational database management systems, and installing and configuring Internet applications is helpful for data managers; however, instructions for implementing ODM databases and the WaterOneFlow web services are contained in documentation available via the CUAHIS HIS website (http://his.cuahsi.org). Data managers with varying levels of expertise at the 11 test beds were able to successfully publish data using the system after having received a pre-configured server. Once the ODM database and web services are set up, they require little maintenance apart from loading new data if and when it becomes available. Personnel (i.e., data manager) resources required to implement the system depend on the amount and complexity of the data to be published. The degree to which data acquisition is automated and the level of manual quality control to which the data are subjected are also drivers in the required personnel costs..

One advantage of this data publication system is that a standard, robust data model and controlled vocabularies ensure consistent and fully specified data and metadata, leading to higher quality analysis with less uncertainty and fewer data interpretation errors. The value of fully specified metadata cannot be overstated. Federation of individual databases (i.e., test bed or observatory databases) is also simplified because each of the databases has the same format and uses the same vocabulary. This simplifies the design of applications that facilitate data discovery across the entire network of published data. Additionally, because a consistent data model and vocabulary are used across sites, software application development can also be standardized and components reused at each site.

The ODM controlled vocabulary system provides a community resource for building a common vocabulary for environmental and water resources data and is a good example of how common systems can support a larger community. Other software tools include the WaterOneFlow web services, data loading and editing tools for ODM, and data visualization and retrieval tools that interact with the WaterOneFlow web services. Readers are referred to the CUAHSI HIS website for details of these software applications (http://his.cuahsi.org). The free availability of these software tools is a significant asset to investigators who cannot afford or do not have the expertise to develop sophisticated and interactive data publication websites on their own.

The data publication system described in this paper is not limited to test beds or environmental observatories, and, because of this, the network of available data is expected to grow. Data from several research sites outside of the original 11 test beds have already been published using this system. Investigators working outside of the environmental observatory community can adopt the methods and available software tools to publish their own data. By doing so, the network of observatories and other data sources that adopt the same infrastructure, although separated in space, will become an integrated network of consistent data like NWIS, STORET, and other national repositories. Sophisticated tools such as ontologies may still be needed to integrate research datasets with those from other national data providers, but one level of complexity (i.e., semantic and syntactic heterogeneity among the network of research datasets) can be avoided through the adoption of a common data publication system and common vocabulary.

Last, the conceptual framework of the data publication system presented in this paper (i.e., a common data model, a centralized controlled vocabulary system, web services for communicating data from federated data sources, and a central registry for web services) can be applied within any domain in which a community of diverse investigators is collecting data.

4.8. Software and Data Availability

The software components described in this paper, including ODM, ODM Tools, the ODM controlled vocabulary system, the WaterOneFlow web services, and the central web services registry can be accessed through the CUAHSI HIS website http://his.cuahsi.org. The test bed data described in this paper can be accessed through the individual web services for each test bed, which are listed in the central web services registry, also available through the HIS website.

References

Bandaragoda, C. J., D. G. Tarboton, and D. R. Maidment (2005), User Needs Assessment, in *Hydrologic Information System Status Report, Version 1*, edited by D. R. Maidment, chap. 4, pp.48-87, Consortium of Univ. for the Adv. Of Hydrol. Sci., Washington, D. C. (Available at http://www.cuahsi.org/docs/HISStatusSept15.pdf)

Beran, B., and M. Piasecki (2008), Engineering new paths to water data, *Computers and Geosciences*, In press. doi:10.1016/j.cageo.2008.02.017.

Bergamaschi, S., S. Castano, M. Vincini, and D. Beneventano (2001), Semantic integration of heterogeneous information sources, *Data & Knowledge Eng.*, *36*(3), 215-249, doi:10.1016/S0169-023X(00)00047-1.

Borgman, C.L., J. C. Wallis, and N. Enyedy (2007), Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries, *International J. Digital Libraries*, *7*(1), 17-30, doi:10.1007/s00799-007-0022-9.

Bosch, D.D., J. M. Sheridan, R. R. Lowrance, R. K. Hubbard, T. C. Strickland, G. W. Feyereisen, and G. W. Sullivan (2007), Little River Experimental Watershed database, *Water Res. Res.*, *43*, W09470, doi:10.1029/2006WR005844.

Colomb, R.L. (1997), Impact of semantic heterogeneity on federating databases, *The Computer J., 40*(5), 235-244, doi:10.1093/comjnl/40.5.235.

Cox, S. (Ed.) (2006), Observations and Measurements, OGC Best Practices Document, OGC 05-087r4, Version 0.14.7. (Available at http://www.opengeospatial.org/standards/bp)

Cox, S., R. Jones, B. Lawrence, N. Milic-Frayling, and L. Moreau (2006), Interoperability issues in scientific data management (Version 1.0). Technical report, The Technical Computing Initiative, Microsoft Corporation. (Available at http://download.microsoft.com/download/f/b/3/fb3d02b8-2210-4d0d-a747-9519eafae6c1/ScientificDataManagement4.18.07.pdf)

EML Project Members (2008), Ecological Metadata Language (EML). http://knb.ecoinformatics.org/software/eml/. [Last accessed February 26, 2008.]

Goodall, J. L., J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky (2008), A first approach to Web services for the National Water Information System, *Environ. Model. & Software*, *23*(4), 404-411, doi:10.1016/j.envsoft.2007.01.005.

Gray, J., D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber (2005), Scientific data management in the coming decade, *SIGMOD Record, 34*(4), 34-41, doi:10.1145/1107499.1107503.

Hart, J. K., and K. Martinez (2006), Environmental sensor networks: A revolution in earth system science?, *Earth-Science Reviews*, *78*, 177-191, doi:10.1016/j.earscirev.2006.05.001.

Heflin, J. (Ed.) (2004), OWL Web Ontology Language use cases and requirements, W3C Recommendation 10 February 2004. (Available at http://www.w3.org/TR/webont-req/)

Helly, J. J. (2006), Digital Library Technology for Hydrology, in *Hydroinformatics Data Integrative Approaches in Computation, Analysis, and Modeling*, chap. 3, pp. 21-37, edited by P. K. Kumar, J. Alameda, P. Bajcsy, M. Folk, and M. Markus, CRC Press, Boca Raton, Fla.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Res. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.

Lin, K., and B. Ludäscher (2003), A system for semantic integration of geologic maps via ontologies, in *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data (SCISW)*, Sanibel Island, Florida, October 20, 2003. (Available at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/sia_2.pdf)

Madin, J., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa (2007), An ontology for describing and synthesizing ecological observation data, *Ecological Informatics*, *2*(3), 279-296, doi:10.1016/j.ecoinf.2007.05.004.

Madnick, S., and H. Zhu (2006), Improving data quality through effective use of data semantics, *Data & Knowledge Eng.*, *59*(2), 460-475, doi:10.1016/j.datak.2005.10.001.

Mars Climate Orbiter Mishap Investigation Board (1999), Mars Climate Orbiter Mishap Investigation Board Phase I Report, November 10, 1999. (Available at ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf)

Montgomery, J. L., T. Harmon, W. Kaiser, A. Sanderson, C. N. Haas, R. Hooper, B. Minsker, J. Schnoor, N. L. Clesceri, W. Graham, and P. Brezonik (2007), The WATERS Network: an integrated environmental observatory network for water research, *Environ. Sci. and Technology*, *41*(19), 6642-6647. (Available at http://pubs.acs.org/subscribe/journals/esthag/41/i19/pdf/100107feature_waters.pdf)

Moran, S. M., W.E. Emmerich, D. C. Goodrich, P. Heilman, C. D. Holifield Collins, T. O. Keefer, M. A. Nearing, M.H. Nichols, K. G. Renard, R. L. Scott, J. R. Smith, J. J. Stone, C. L. Unkrich, and J. Wong (2008), Preface to special section on Fifty Years of Research and Data Collection: U.S. Department of Agriculture Walnut Gulch Experimental Watershed, *Water Res. Res., 44*, W05S01, doi:10.1029/2007WR006083.

Morocho, V., F. Saltor, and L. Perez-Vidal (2003), Ontologies: Solving semantic heterogeneity in federated spatial database system, in *Proceedings of 5th International Conference on Enterprise Information System*, pp. 347-352, Angers, France, April, 2003. (Available at http://citeseer.ist.psu.edu/morocho03ontologies.html)

National Science Foundation (2007), Cyberinfrastructure vision for 21st century discovery, National Science Foundation Cyberinfrastructure Council, NSF 07-28. (Available at http://www.nsf.gov/pubs/2007/nsf0728/index.jsp)

Nichols, M. H., and E. Anson (2008), Southwest Watershed Research Center Data Access Project, *Water Res. Res.*, *44*, W05S03, doi:10.1029/2006WR005665.

Ramachandran, R., S. A. Christopher, S. Movva, X. Li, H. T. Conover, K. R. Keiser, S. J. Graves, and R. T. McNider (2005), Earth Science Markup Language: A solution to address dataformat heterogeneity problems in atmospheric sciences, *Bull. Am. Meteorological Soc.*, *86*(6), 791-794, doi:10.1175/BAMS-86-6-791.

Research Information Network (2008), To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN). (Available at http://www.rin.ac.uk/data-publication)

Ruddell, B. L., and P. Kumar (2006), Hydrologic data models, in *Hydroinformatics Data Integrative Approaches in Computation, Analysis, and Modeling*, chap. 5, pp. 61-79, edited by P. K. Kumar, J. Alameda, P. Bajcsy, M. Folk, and M. Markus, CRC Press, Boca Raton, Fla.

Sheth, A. P., and J. A. Larson (1990), Federated database systems for managing distributed, heterogeneous, and autonomous databases, *Computing Surveys*, *22*(3), doi:10.1145/96602.96604.

Slaughter, C. W., D. Marks, G. N. Flerchinger, S. S. Van Vactor, and M. Burgess (2001), Thirty-five years of research data collection at the Reynolds Creek Experimental Watershed, Idaho, United States, *Water Res. Res.*, *37*(11), doi:10.1029/2001WR000413.

Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental design and initial results from the Mahurangi River Variability Experiment: MARVEX, in *Observations and Modelling of Land Surface Hydrological Processes*, edited by V. Lakshmi, J. D. Albertson and J. Schaake, pp. 201-213, Water Resources Monographs, American Geophysical Union, Washington, D. C.

Zaslavsky, I., D. Valentine, and T. Whiteaker (Eds.) (2007), CUAHSI WaterML, OGC Discussion Paper, OGC 07-041r1, Version 0.3.0. (Available at http://www.opengeospatial.org/standards/dp)

Table 4.1. Examples of Semantic Heterogeneity in Two Popular Water Resources Datasets Demonstrating Both Structural and Contextual Semantic Heterogeneity

General Description of Attribute	USGS NWIS[a]	EPA STORET[b]
Structural Semantic Heterogeneity		
Code for location at which data are collected	"site_no"	"Station ID"
Name of location at which data are collected	"Site" OR "Gage"	"Station Name"
Code for measured variable	"Parameter"	?[c]
Name of measured variable	"Description"	"Characteristic Name"
Time at which the observation was made	"datetime"	"Activity Start"
Code that identifies the agency that collected the data	"agency_cd"	"Org ID"
Contextual Semantic Heterogeneity		
Name of measured variable	"Discharge"	"Flow"
Units of measured variable	"cubic feet per second"	"cfs"
Time at which the observation was made	"2008-01-01"	"2006-04-04 00:00:00"
Latitude of location at which data are collected	"41°44'36"	"41.7188889"
Type of monitoring site	"Spring, Estuary, Lake, Surface Water"	"River/Stream"

[a] United States Geological Survey National Water Information System (http://waterdata.usgs.gov/nwis/).
[b] United States Environmental Protection Agency Storage and Retrieval System (http://www.epa.gov/storet/).
[c] An equivalent to the USGS parameter code does not exist in data retrieved from EPA STORET.

Table 4.2. Test Bed Data Network Summary as of June 17, 2008

Item	Total Number
ODM Databases	31
Data Sources	41
Monitoring Sites	3,767
Variables	793
Measurement Methods	99
Data Values	41,651,095

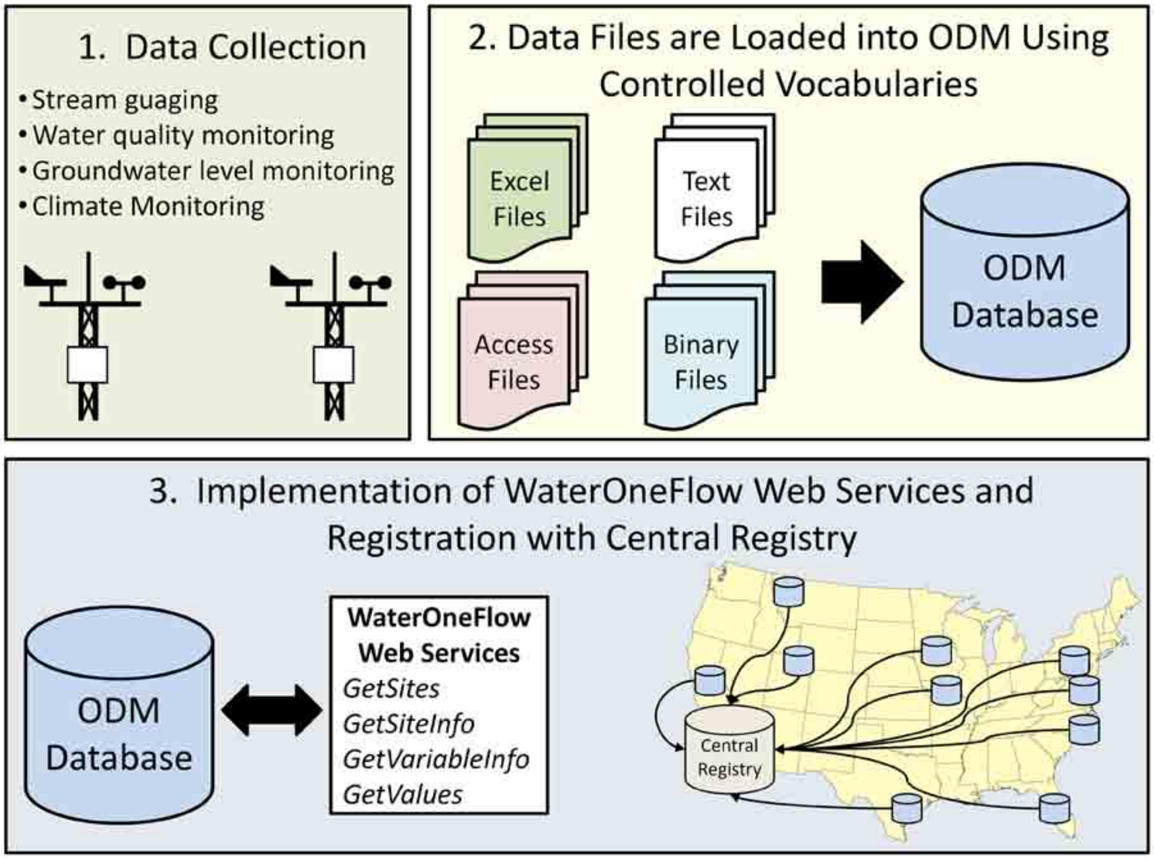

1. Data Collection
•Stream guaging
•Water quality monitoring
•Groundwater level monitoring
•Climate Monitoring
2. Data Files are Loaded into ODM Using Controlled Vocabularies
Excel Files
Text Files
Access Files
Binary Files
ODM Database
3. Implementation of WaterOneFlow Web Services and Registration with Central Registry
ODM Database
WaterOneFlow Web Services
GetSites
GetSiteInfo
GetVariableInfo
GetValues
Central Registry

Figure 4.1. General architecture of the test bed data publication system. Data are collected using field sensors and other observational procedures. Observational data with multiple formats are combined within a single ODM database where they are annotated with appropriate metadata using the ODM controlled vocabularies. The ODM web services are then implemented on top of the ODM database and are registered with the central web services registry.

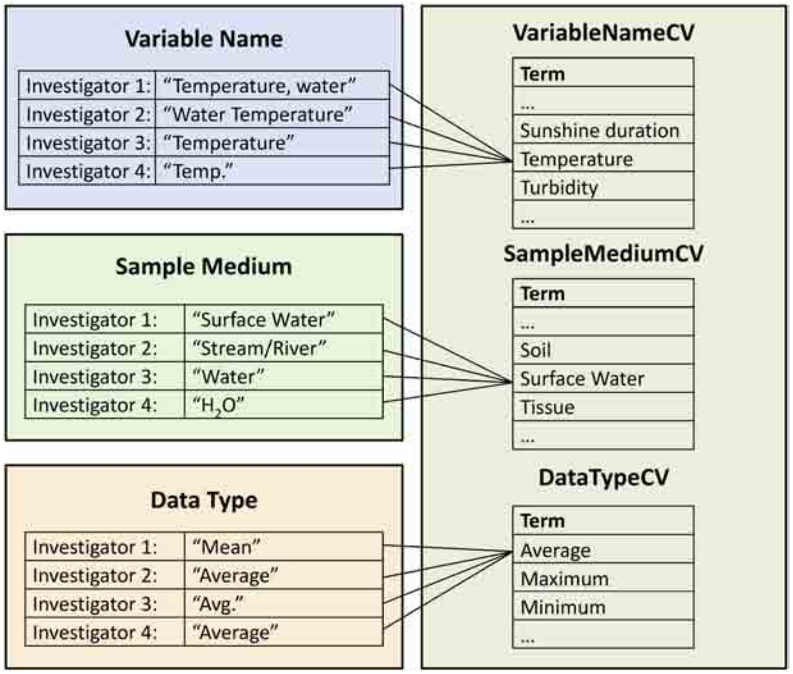


Figure 4.2. Example of how contextual heterogeneity in the attributes of similar datasets from several different investigators can be reconciled through the use of the ODM controlled vocabularies.

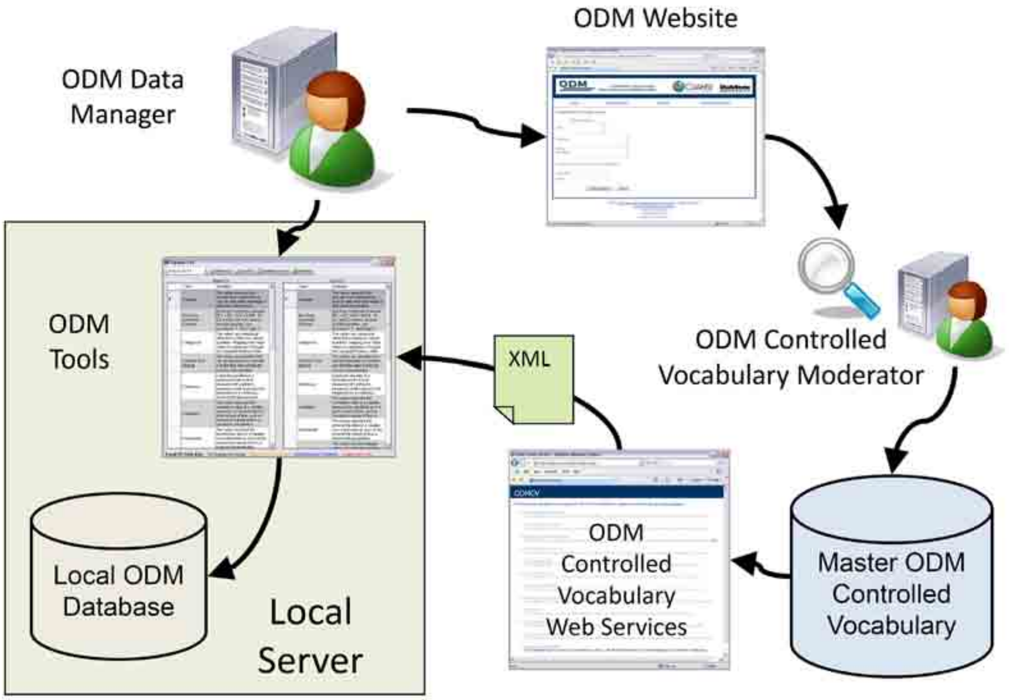

ODM Website
ODM Data Manager
ODM Tools
ODM Controlled Vocabulary Moderator
XML
Local ODM Database
Local Server
ODM Controlled Vocabulary Web Services
Master ODM Controlled Vocabulary

Figure 4.3. The ODM controlled vocabulary system.

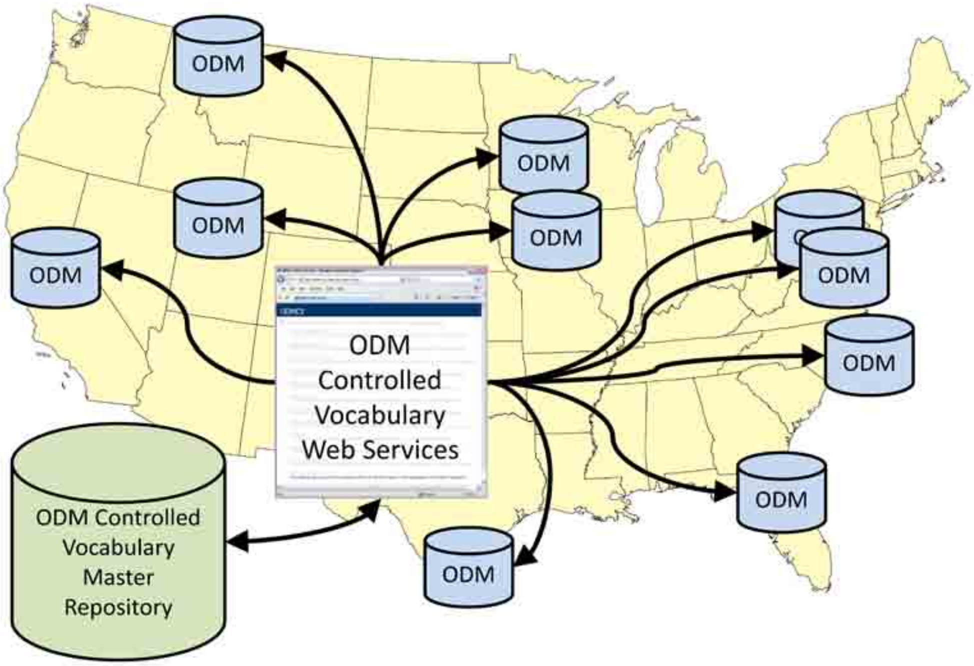
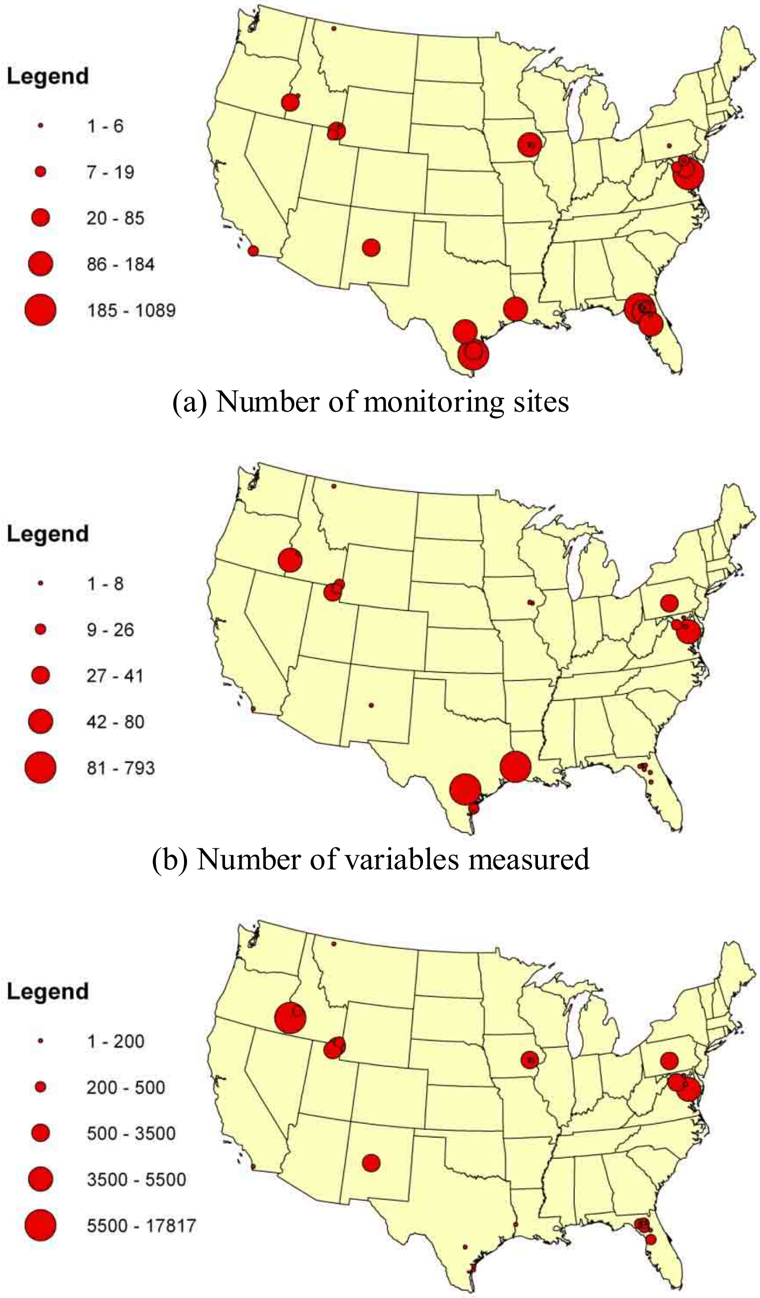

ODM
ODM
ODM
ODM
ODM
ODM
ODM
ODM
ODM
ODM Controlled Vocabulary Web Services
ODM Controlled Vocabulary Master Repository

Figure 4.4. The central ODM controlled vocabulary repository serves the ODM databases located at each of the test beds.



(a) Number of monitoring sites

(b) Number of variables measured

(c) Number of data values (X 10^3)

Figure 4.5. Distribution of monitoring sites (a), variables (b), and data values (c) across the U.S. in the research data publication network as of June 17, 2008.

CHAPTER 5

COMPONENTS OF AN INTEGRATED ENVIRONMENTAL OBSERVATORY INFORMATION SYSTEM[1]

Abstract

Recently, community initiatives have emerged for the establishment of cooperative large-scale environmental observatories. Cyberinfrastructure is the backbone upon which these observatories will be built, and scientists' ability to access and use the data collected within observatories to address broad research questions will depend on the successful implementation of cyberinfrastructure. The research described in this paper advances the cyberinfrastructure available for supporting environmental observatories. We describe the general components of an environmental observatory information system for collecting, storing, and publishing point observations data. We then describe the implementation of prototypes for each of the generalized components within the Little Bear River environmental observatory test bed, as well as across a nationwide network of 11 observatory test bed sites. Together, these components comprise an integrated environmental observatory information system that has enabled us to not only analyze and synthesize our data to advance our understanding of the Little Bear River watershed but also manage and publish all of the observational data that we are collecting on the Internet in simple to use formats that are easily accessible and discoverable by others. Enhancements to the infrastructure for research and education that are enabled by this research will impact a diverse community, including the community of researchers

[1] Coauthored by Jeffery S. Horsburgh and David G. Tarboton

involved with prospective CUAHSI/CLEANER/WATERS environmental observatories as well as other observatory efforts, research watersheds, and test beds.

5.1. Introduction

Many researchers within the science and engineering research communities have suggested that new data networks, field observations, and field experiments that recognize the spatial and temporal heterogeneity of hydrologic processes will be needed to address complex and encompassing questions and advance the science of hydrology [*Woods et al.*, 2001; *Hart and Martinez*, 2006; *Kirchner*, 2006; *Montgomery et al.*, 2007]. This knowledge that current understanding is constrained by a lack of observations at appropriate spatial and temporal scales has motivated community initiatives (e.g., http://www.cuahsi.org, http://cleaner.ncsa.uiuc.edu, http://www.watersnet.org/) towards the establishment of large-scale environmental observatories, which aim to overcome this limitation through the collection of data at unprecedented spatial and temporal resolution.

To what extent is current understanding constrained by the tools and methods that have heretofore been used to organize, manage, publish, visualize, and analyze data? This question is important in an observatory context because as the amount and complexity of data grows, it becomes increasingly difficult, if not impossible, for data analysts to identify trends and relationships in the data and to derive information that enhances understanding using simple query and reporting tools [*Connolly and Begg*, 2005]. Combining multiple lines of evidence (e.g., using data streams from multiple sensors or from multiple sites) into a single analysis becomes much more difficult when they consist of thousands or even tens or hundreds of thousands of observations. Thus, even if the data are available, without the tools to manage and manipulate the data their utility in fostering process understanding is limited.

Additionally, it is difficult for the broader scientific community beyond individuals who collected the data to use them for scientific analyses if they are never published or if semantic and syntactic differences among datasets preclude their use in common analyses. Recently, these questions of data availability, organization, publication, visualization, and analysis have come to the forefront within many scientific communities (e.g., hydrology, environmental engineering, etc.). With advances in observing, computing, and information technology, it is becoming increasingly important and feasible to develop systems and models that answer these questions. Hydrologic Information Systems are emerging as technology to address these questions in the area of Hydrology and Water Resources.

Observatory initiatives will require enormous investments in both capital and in information technology infrastructure to manage and enable the observing systems. According to the *National Research Council* [2008], advanced information technology infrastructure will be required as a central component in the planning and design of observatories to help manage, understand, and use diverse datasets. Comprehensive infrastructure that is being used to capitalize on advances in information technology has been termed “cyberinfrastructure” and integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools [*National Science Foundation*, 2007].

Cyberinfrastructure is the backbone upon which environmental observatories will be built. Scientists' ability to access and use the data collected within observatories to address research questions will depend on the successful implementation of cyberinfrastructure. The overall cyberinfrastructure platform for environmental observatories is expected to include high-performance computing tools and intensive database management for the collection, storage, and dissemination of environmental data; advanced data visualization tools; community-vetted models for system and process synthesis that can be used in near-real time; and collaboration and knowledge networking tools that will help multidisciplinary and geographically dispersed teams of researchers to work together effectively [*Montgomery et al.*, 2007].

The research described in this paper seeks to advance the cyberinfrastructure available for supporting environmental observatories. We focus mainly on the very practical aspects of data management within observatories and the software components required to establish seamless linkages between sensors in the field, a centralized data storage and management system, applications that publish the data in easy to use formats on the Internet, and applications that support data discovery and unambiguous interpretation. We first articulate what the necessary cyberinfrastructure components are that are required to support this functionality and describe the functional requirements of each. We then discuss emerging technologies that are being used to build and implement these components. We present specific implementations in the form of a case study for the Little Bear River environmental observatory test bed (LBRTB), where instances of the generalized components have been developed and implemented. These methods and tools are applicable not only to proposed environmental observatories, but to all data-intensive studies and experimental sites where management and publication of large quantities of observational data is required.

The focus of this paper is on a single, yet very important, class of water resources data – observational data measured at a point (e.g., time series data collected at a stream monitoring site or weather station located at a fixed point in space). It is anticipated that the enhancements to the infrastructure for research and education that are enabled by the methods and software described in this paper will impact a diverse community, including researchers involved with prospective CUAHSI/CLEANER/WATERS environmental observatories, as well as other observatory efforts, research watersheds, and test beds.

5.2. Existing Cyberinfrastructure for Environmental Observations

There are currently several large-scale cyberinfrastructure activities underway. These include: the National Ecological Observatory Network (NEON), which is planning the deployment of networked sensors and cyberinfrastructure to gather data on the nation's most compelling ecological challenges (http://www.neoninc.org); the Long Term Ecological Research Network (LTER), which is a network of research sites that promotes synthesis and comparative research across sites and ecosystems (http://www.lternet.edu/); the Geosciences Network (GEON), which has developed infrastructure for discovering, accessing and integrating earth sciences data and tools (http://www.geongrid.org/); EarthScope, which is a national earth science program to explore the structure and evolution of the North American Continent and understand processes controlling earthquakes and volcanoes (http://www.earthscope.org/); and many others. Although these initiatives have similar goals, which include creating and sharing multidisciplinary datasets, facilitating collaborative and interdisciplinary research, and creating infrastructure to enable scientific discoveries, the cyberinfrastructure being developed for each is driven by the needs and requirements of each of the specific communities. The types of data being collected within each of these communities can be quite diverse, and, because of this, there have been relatively few efforts to date aimed at using common cyberinfrastructure across observatory initiatives.

Within the hydrologic science community, the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) has been developing cyberinfrastructure aimed at providing a single portal to access hydrologic data from a variety of federal, state, and local agencies and, more importantly in the context of this paper, as a support structure for the development of cooperative large-scale environmental observatories [*Maidment*, 2005, 2008]. The CUAHSI Hydrologic Information Systems (HIS) project has produced a variety of technologies under this effort that are advancing the way hydrologists, engineers, and scientists are storing, accessing, and analyzing environmental data. Some of these technologies, including an Observations Data Model (ODM) that provides a persistent storage mechanism for observatory data [Chapter 3], a Data Access System for Hydrology (DASH) that provides Internet map based access to data stored in a central observations database [*Whitenack et al.*, 2007], and web services that provide remote programmatic access to data stored within a central observations database and other national data sources [*Valentine et al.*, 2007], have progressed to the point that they can be implemented as the support structure for an environmental observatory.

With respect to providing support for environmental observatories, the proving grounds for the CUAHSI HIS tools has been a national network of 11 environmental observatory test beds that are part of the Water and Environmental Research Systems (WATERS) network (http://watersnet.org/wtbs/index.html). Investigators at each of the WATERS network test beds participated with the CUAHSI HIS Team in the development and deployment of common hydrologic information system capability for publishing observations from each of the test beds. The goal in implementing a common set of cyberinfrastructure was to create a national network of consistent data and to enable cross-domain analysis within individual test beds as well as cross-test bed sharing and analysis of data.

Although much progress has been made by the CUAHSI HIS project in providing better access to national datasets and in supporting environmental observatories, significant work remains to better define the components required for integrated observatory information systems, the functionality that each of these components should have, and the specific technologies that are available for creating or implementing these components.

5.3. Functionality of an Integrated Observatory Information System

Environmental observations are fundamental to hydrology and water resources, and the manner in which data are collected, organized, and presented either enables or inhibits their scientific analysis [Chapter 3]. When scientists and engineers are looking for environmental observations data, they may face the following problems: (1) data do not exist or are not sufficient; (2) data are not published; (3) data are not easy to access;

(4) data are inconsistent; and (5) data are not adequately documented [*Tomasic and Simon*, 1997]. The first item is one of the main drivers for establishment of environmental observatories and will be addressed through the establishment of sensor networks and collection of high frequency data. Items 2 – 5 pose significant challenges for the cyberinfrastructure components that will support systematic data organization and publication within observatories. Indeed, the data management and publication tools will be every bit as important as the data itself in establishing observatories as community resources. The following sections present the overall conceptual architecture for an environmental observatory information system. We describe each of the components and the functionality that is required to support observational data collected within environmental observatories.

5.3.1. Data Collection and Communication Infrastructure

The fundamental premise of environmental observatories is that knowledge of the physical, chemical, and biological mechanisms controlling water quantity and quality is limited by lack of observations at the necessary spatial density and temporal frequency needed to infer the controlling processes [*Montgomery et al.*, 2007]. The goal, then, is to create a network of heavily instrumented sites where data are collected with unprecedented spatial and temporal resolution, aiming at creating greater understanding of the earth's water and related biogeochemical cycles and enabling improved forecasting and management of critical water processes.

Environmental sensors and network communications infrastructure will play a major part in proposed environmental observatories. An environmental sensor network is an array of sensor nodes and a communications system that allows their data to reach a server [*Hart and Martinez*, 2006]. Dynamic variables measured at sensor nodes within observatories will include microclimate variables, precipitation chemistry variables, soil variables, stream physical and chemical variables, groundwater variables, snow variables, and many others [*WATERS Network*, 2008]. Many of these variables will be measured and reported in near real time, enabling researchers to conduct dynamic, predictive modeling for water, sediment, and water quality, and enabling feedback within the monitoring systems to adjust operation in response to events [*Montgomery et al.*, 2007].

Real time or near-real time reporting of data requires robust communications infrastructure. Currently available telemetry options include both hard wired connections (e.g., telephone land lines or Internet connections) and wireless solutions (e.g., cellular phone, radio, satellite). The choice of which type of communication technology to use is dependent on the following factors: (1) the required data collection and reporting frequency; (2) the location and characteristics of the monitoring site; (3) power requirements and availability at remote locations; and (4) equipment and service costs. Each of these factors present challenges for the design and implementation of environmental observatories, and in current practice, communications networks may be made up of a combination of the available technologies to overcome the challenges listed above.

5.3.2. Persistent Data Storage

Once observational data are delivered from sensor nodes to a centralized server, they must be parsed into a persistent data storage structure. This has been done in a number of different ways, ranging from file- and directory-based data structures to complex relational databases implemented using diverse data models in advanced database management systems. The key functionality that must be supported by the persistent data store includes storage and retrieval and transaction management (i.e., loading the data, querying the data, and editing the data).

Environmental observations are not self describing, and, because of this, interpretation of a particular set of observations requires contextual information, or metadata. Metadata is the descriptive information about data that explains the measurement attributes, their names, units, precision, accuracy, and data layout, as well as the data lineage describing how the data was measured, acquired, or computed [*Gray et al.*, 2005]. The importance of recording fundamental metadata to help others discover and access data products is well recognized [*Michener et al.*, 1997; *Bose*, 2002; *Gray et al.*, 2005]. The persistent data store must capture not only the observation values, but all of their supporting metadata as well, providing traceable heritage from raw measurements to usable information and allowing observations to be unambiguously interpreted [Chapter 3].

5.3.3. Quality Assurance, Quality Control, and Data Provenance

In-situ environmental sensors often operate under harsh conditions, and, because of this, they often malfunction. Many sensors are prone to drift, and the data they collect can also become corrupt when they are transmitted over communication networks. Uncorrected errors can significantly affect the data's value for scientific applications, especially if they are to be used by investigators that did not collect the data and are not intimately familiar with the data collection methods and environmental conditions that may have caused the anomalies. Several studies have investigated automated methods for detecting anomalies in sensor data streams, which is particularly important in real time applications of the data and in detecting instrument malfunctions [*Mourad and Bertrand-Krajewski*, 2002; *Hill et al.*, 2007; *Liu et al.*, 2007]. Although these methods are good at detecting and flagging potentially bad sensor values, they are not always good at fixing them.

Producing high quality, continuous data streams from environmental sensors requires correcting raw sensor data for instrument drift, filling missing values where appropriate, and correcting other spurious values. It also involves maintaining the linkages between raw data values and quality controlled data values so that the provenance of the data can be maintained. The process of correcting raw sensor data can be time and labor intensive, and tools that facilitate this process are needed.

5.3.4. Data Publication and Interoperability

Environmental observatories may be operated as cooperative community resources. To become so, the data collected within observatories must be published in a way that investigators working both within and across observatories and scientific domains can easily access and unambiguously interpret the data. One of the biggest challenges in achieving this is heterogeneity within both data formats and the vocabularies used to describe the data [*Sheth and Larson*, 1990; *Colomb*, 1997]. The data publication systems used in environmental observatories must not only transmit data to users, but they must do it in a way that overcomes semantic and syntactic heterogeneity in datasets [Chapter 4].

Web services are applications that provide the ability to pass information between computers over the Internet, usually formatted using a platform independent markup language such as extensible markup language (XML) [*Goodall et al.*, 2008]. Many large-scale cyberinfrastructure initiatives are now using web service-oriented architectures [*Droegemeier et al.*, 2005; *Youn et al.*, 2007; *Maidment*, 2008]. Service-oriented architectures rely on a collection of loosely coupled, self-contained services that communicate with each other through the Internet and that can be called from multiple clients (e.g., Excel, Matlab, Visual Studio, etc.) in a standard fashion [*Maidment*, 2008]. Web services can be used to accomplish both data publication (by making data available over the Internet) and interoperability (by transmitting data in a platform independent format like XML using a standard schema like Water Markup Language, or WaterML), making them powerful tools in the development of cyberinfrastructure for environmental observatories.

The distributed nature of the proposed network of environmental observatories will require distributed cyberinfrastructure. According the *National Research Council* [2008], a robust cyberinfrastructure will provide common frameworks, components, modules, and interface models that can be used in multiple observatories or applications. Standardization upon a service-oriented architecture is the key. Each observatory can publish data using a common set of web services that transmit data using a common language, and all of the underlying processing and complexity (which may be different from one observatory to the next) is hidden from data consumers. In addition, by standardizing the data transmission services and formats, others outside of the observatory community can also publish their data using the same tools.

5.3.5. Data Discovery, Visualization, and Analysis

Data discovery is an important aspect of the cyberinfrastructure required to support publication of research data because scientists' ability to find, decode, and interpret available datasets will determine how or if the data are used for scientific analyses [Chapter 4]. In most cases, scientists want to download data and work with them in their own analysis environment. To do this, they need simple screening level tools to assist them in deciding which data will be useful for their analyses. Map-based, point-and-click access to observational data can be a powerful tool for providing users with data discovery capabilities. *Beran and Piasecki* [2008] describe a map-based search engine called Hydroseek (http://www.hydroseek.org) that was specifically designed to provide users with a single interface to query and retrieve consistently formatted data from several national hydrologic data providers. Users don't always know exactly what they are looking for, and the ability to see the layout of monitoring sites superimposed upon a map provides them with the spatial context that they need to select the data that they are interested in. Juxtaposition of spatial data and time series of environmental observations also provides important spatial reference for interpreting the data. For example, knowing the land use distribution or terrain above a stream monitoring site is important in assessing nutrient and sediment concentrations.

Simple data visualization tools can also assist users in discovering data that they are interested in. Many users prefer to visualize multi-dimensional datasets so that they have a better understanding of the quality and characteristics of the data before downloading them [*Jeong et al.*, 2006]. Tools that enable users to query data and then generate simple plots and descriptive statistics are generally adequate for this purpose and can also be useful for users that do not have the technical expertise to extract the data, load it into data analysis software, and then develop useful visualizations or analyses of the data. By providing users with tools that manipulate the data automatically and that do not require any specialized software expertise other than knowing how to operate an Internet browser, an observatory information system can extend the reach of the data to less technical users.

5.4. The Little Bear River Environmental Observatory Test Bed: A Case Study

As part of the planning process for a network of large-scale environmental observatories, a network of 11 environmental observatory test bed projects was created in 2006. The test beds are located throughout the United States, and each was established to demonstrate techniques and technologies that could be used in the design and implementation of a national network of large-scale environmental observatories. Research within the test beds has targeted the innovative application of environmental sensors to achieve a better understanding of the fluxes, flow paths, and stores of environmental constituents and the development of software components for publishing observations data in common formats that can be accessed by investigators throughout the scientific community. More information about the test beds and the data that have been or are being collected at each can be found at the following URL (http://www.watersnet.org/wtbs/index.html).

The Little Bear River of northern Utah was established as one of the WATERS test beds with the overarching goal of improving the observing infrastructure and cyberinfrastructure available for the design and implementation of environmental observatories. The primary hypotheses that have been tested in the LBRTB are: 1) that high-frequency estimates of streamflow and constituent concentrations based on surrogate sensor data (e.g., using turbidity as a surrogate for total suspended solids and total phosphorus) collected at multiple sites can significantly improve understanding of the spatial and temporal patterns in constituent fluxes within the watershed, especially for constituents that cannot be logistically or economically measure with high-frequency, and 2) that high-frequency streamflow and hydrochemistry data (i.e., temperature, dissolved oxygen, pH, and specific conductance) can improve our understanding of hydrologic and hydrochemical response to both natural and human induced changes in the environment.

The data intensive nature of the ongoing research within the LBRTB required the development of prototypes for many of the components of an integrated observatory information system to provide tools for managing the data that are being collected. In addition, components of the CUAHSI HIS were adopted for publishing the LBRTB data in a way that is consistent with all of the other observatory test beds. In the following sections we describe each of the components, the role that they have served, and how the combination of these components has led to an integrated observatory information system for the LBRTB.

5.4.1. Data Collection and Communication Infrastructure: The LBRTB Sensor Network

In order to generate the necessary data to enable the investigation of the hypotheses listed above, a sensor network was established that includes seven continuous streamflow and water quality monitoring sites and 2 continuous weather stations. At each monitoring site, a suite of sensors was connected to a Campbell Scientific, Inc. datalogger, and the data are transmitted in near real time to the Utah Water Research Laboratory (UWRL) via a telemetry network. Table 5.1 lists the monitoring sites in the Little Bear River test bed. Table 5.2 lists the variables measured at each type of monitoring site, the data collection frequency, and the sensors used. Figure 5.1 shows the locations of each of the monitoring sites within the Little Bear River watershed.

The LBRTB telemetry system was designed to use a combination of 900 MHz spread spectrum radio links and TCP/IP Internet links to manage transmission of data from each of the monitoring sites to the UWRL. This system was chosen because it eliminated monthly service costs, it had relatively low power requirements, and it maximized the flexibility of the system for accepting new monitoring sites onto the existing network. Establishment of the radio network enabled remote connections to each site for monitoring site status and for retrieving data.

Terrain and vegetation were major challenges that had to be overcome in the design of the radio telemetry network. Digital elevation model (DEM) based viewshed analysis using a Geographic Information System (GIS) was used to identify appropriate locations for radio repeaters so that data from the river monitoring locations, which are typically located at lower elevations with poor line of sight, could be transmitted to one of two remote base stations located at local schools located within the watershed. Figure 5.2 shows the network map for the LBRTB sensor network and identifies pathways, distances, and link types between each of the remote monitoring sites and the UWRL.

Communications with the remote monitoring sites are managed using Campbell Scientific's LoggerNet software (http://www.campbellsci.com). LoggerNet has enabled the setup and configuration of the radio linkages within the telemetry network, the encoding of data collection logic into programs for the dataloggers, and monitoring of the status of the communications links within the sensor network. In the LBRTB implementation, the LoggerNet server at the UWRL is programmed to connect hourly to each remote site and download the most recent data to delimited text files, which are then stored in a location accessible on the local Intranet.

5.4.2. Persistent Data Storage: The LBRTB Observations Database

Once the sensor data are transmitted to the UWRL, they are parsed into an instance of the CUAHSI HIS Observations Data Model (ODM) [Chapter 3]. ODM is a relational model that was designed to be implemented within a relational database management system (RDBMS) and that defines the persistent structure of the data, including the set of attributes that accompany the data, their names, their data type, and their context. ODM also includes a set of controlled vocabularies for many of the data attributes, which are used to ensure that data stored within and across ODM instances are semantically similar. The Little Bear River ODM database serves as the persistent storage mechanism for the LBRTB information system and was implemented in the Microsoft SQL Server 2005 software.

Because there is opportunity for error each time the sensor data are handled, automation is critical to avoiding errors in parsing the datalogger files into the database. Because of this, we developed the ODM Streaming Data Loader application, which allows users to map individual table-based datalogger files to the ODM schema and then run the data loading task periodically as new data are received. Through a wizard-based graphical user interface (GUI), users define the location of the datalogger file(s) on disk (or on a network shared folder or website) and then create all of the necessary metadata records within the ODM database so that the data can be loaded. Figure 5.3 shows the GUI for the ODM Streaming Data Loader. The ODM Streaming Data Loader can then be run manually or on a user defined schedule, and, upon execution, checks each datalogger file that has been mapped for new observations and automatically loads them into the database without user intervention. The combination of the LoggerNet server, which manages the retrieval of data from the remote sensor nodes, and the ODM Streaming Data Loader, which automatically parses the data into an ODM database, demonstrates seamless, automated integration between sensors in the field and a central observations database that persistently stores the data and its metadata.

5.4.3. Quality Assurance, Quality Control, and Data Provenance: ODM Tools

The data loaded into the ODM database from the datalogger files are raw sensor data. Before the data can be used for most applications and analyses they have to be passed through a set of quality assurance and quality control procedures [*Mourad and Bertrand-Krajewski*, 2002]. For this purpose, we developed a software application called ODM Tools that enables data managers who are administrating ODM databases to query, visualize, and edit data stored within an ODM database. ODM Tools provides a suite of functionality for editing data series (i.e., the time series of observations from a single sensor at a single monitoring site) to remove obvious errors, sensor malfunctions, and instrument drift. Users can insert data values, delete data values, adjust data values by multiplying by or adding a constant value, interpolate data values, and perform linear drift corrections over ranges of data. Users can also flag data values with qualifying comments, which are then stored with the data in the database.

Data editing is performed within a form that has both graphical and tabular views of the data. Figure 5.4 shows the ODM Tools data editing interface. Several data filters are available for finding and selecting data values that may need to be edited. Specific filters include selecting data values above or below a threshold, selecting data values where gaps occur, selecting data where the change from one observation to the next is greater than some value, and selecting data occurring within a particular time interval. The ODM Tools application adopts the business rules (i.e., the relationships and constraints) of ODM. Primary instrument data streams are preserved, while any edits are performed on copies derived from these data. ODM and ODM Tools preserve the provenance of the data by maintaining the linkages between derived or quality controlled observations and the raw observations that they were derived from. Figure 5.5 shows a portion of a specific conductance time series before and after quality control editing using ODM Tools.

5.4.4. Data Publication and Interoperability: The LBRTB Web Services

The LBRTB information system has adopted the WaterOneFlow web services of the CUAHSI HIS as the main mechanism for communicating the observational data to users. The WaterOneFlow web services respond to user queries using a standard set of web service methods, and transmit data extracted from the LBRTB observations database encoded using WaterML [*Zaslavsky et al.*, 2007]. WaterOneFlow methods include *GetSites* for returning a list of sites within the database along with the metadata for each site, *GetVariableInfo* for returning a list of variables within the database along with the metadata for each variable, *GetSiteInfo* for returning a list of variables with data at a site, and *GetValues* for returning the time series of data for a site and variable combination. The web service methods can be called from many different programming languages and other software applications, including Microsoft Visual Basic, Microsoft Excel, MatLab, and others from anywhere an Internet connection is available.

By adopting the WaterOneFlow web services and WaterML, the LBRTB data are published in a format that is consistent with all of the other WATERS observatory test beds (which have also adopted the WaterOneFlow web services), creating a network of consistently published scientific data. WaterML serves as a standard data transmission language, ensuring that data retrieved from all of the test beds is syntactically similar and promoting interoperability of the data through the use of standard web services protocols and an XML schema that is platform, application, and programming language independent. The use of ODM as the underlying data model with its controlled vocabularies ensures that when the data from each test bed are encoded using WaterML they are consistently described and semantically similar.

One additional advantage to using the WaterOneFlow web services is that high level search tools like Hydroseek, which is part of CUAHSI's Central HIS system and is capable of consuming WaterOneFlow web services, can find and present data to potential users. Simple keyword searches in Hydroseek are now capable of returning observational data from each of the test beds' web services as well as from national data providers such as the United States Geological Survey and the U.S. Environmental Protection Agency. The significance of this is not just the linkage with Hydroseek, but that through the adoption of a common service oriented architecture, any application developer can now program against any of the test bed web services as if the data that they present were located on their own machine.

5.4.5. Data Discovery, Visualization, and Analysis: The LBRTB Map Server and Time Series Analyst

A website was developed for the LBRTB that provides information about the ongoing research and links to several applications that present the LBRTB data (http://littlebearriver.usu.edu). Included is a listing of monitoring sites along with photographs, site descriptions, and information about the variables being measured and monitoring equipment installed at each one. Links are provided to launch the location of each site in a Google Maps interface. Also included in the website is a listing of the current conditions within the watershed. This listing shows the latest observation of each variable at each site and is invaluable in determining the status of the monitoring and telemetry system.

In addition to these information items, two separate Web applications were developed to provide access the LBRTB data. The first is the LBRTB map server, which is a light weight, map-based tool that plots the locations of the monitoring sites. It enables simple spatial queries by allowing users to select a variable from a drop down list, which then redraws the map showing only monitoring sites with data for the selected variable. The LBRTB map server was implemented using Google Maps and so benefits from the Google Maps base map data and the Google Maps JavaScript Application Programmer Interface (API) that enables customization of the mapping components. The LBRTB map server is available at the Little Bear River test bed website (http://littlebearriver.usu.edu).

When a user clicks on a monitoring site in the LBRTB map server, a balloon pops up that provides information about the selected site. The balloon also provides a hyperlink to the Time Series Analyst, which is the other application that was developed for visualization and analysis of the LBRTB data. The Time Series Analyst provides a simple, Internet-based interface to the LBRTB observations database. Users can select a site and variable combination and a date range and then generate a variety of plots and summary statistics for the selected data series directly in their Web browser. They can also save the plots as images and download the data used to generate the plots. The LBRTB Time Series Analyst application is available at the Little Bear River test bed website (http://littlebearriver.usu.edu).

Both of these applications were designed to use a direct SQL connection to an ODM database. However, they were also developed to be generic and reusable – i.e., they can be connected to multiple ODM databases. Each one has a simple query interface that allows query parameters to be passed to the application through the URL string. This is useful for launching the application in a specific state (e.g., launching the Time Series Analyst from the map server with a monitoring site pre-selected based on which site the user clicked on in the map).

Figure 5.6 shows the specific architecture of the LBRTB observatory information system. It illustrates how users can interact with the LBRTB observations database directly through the WaterOneFlow web services, through high level search applications like Hydroseek, and through the specific tools that we have built for data discovery, visualization, and analysis, including the LBRTB map server and Time Series Analyst. The flexibility of this system can appeal to a broad range of users, from programmers that want to call the web services to get data for scientific analyses (effectively getting the browser out of the way) to more casual users that simply want to examine a plot of the data on the Internet.

5.5. Discussion and Conclusions

Collection and management of large volumes of high frequency data present challenges for the community of scientists working toward the establishment of large-scale environmental observatories. In this paper, we have presented the general architecture and functional requirements of an environmental observatory information system for collecting, storing, and publishing point observations data. The LBRTB observatory information system is made up of a set of hardware and software components that together demonstrate a specific implementation of the general architecture and advance the cyberinfrastructure available for environmental observatories. The LBRTB information system has enabled the storage and management of all of our test bed data and open and free distribution of the data via simple to use, Internet-based tools. The components of the LBRTB information system are also transferrable, and some of them have already been used at other sites within the WATERS network of environmental observatory test beds.

The use of ODM and the Streaming Data Loader has enabled seamless, automated integration between sensors in the field and a central observations database that persistently stores the data and its metadata. Automation of the data loading task eliminates potential errors and ensures that the database always contains the most recent data. ODM Tools provides graphical tools for transitioning data from raw sensor streams to higher level, QA/QC checked data series that can be confidently used for scientific analyses. ODM Tools adopts the business rules of ODM and preserves the provenance of the data through the editing process.

The WaterOneFlow web services and WaterML serve as a data publication mechanism for the LBRTB and promote interoperability among all of the WATERS environmental observatory test beds. WaterML serves as a data transmission standard that is platform, application, and programming language independent, ensuring that data retrieved from all of the test beds is syntactically and semantically similar. Through adoption of a service oriented architecture, the test beds have created a national network of consistently published scientific data, and application programmers can program against their web services as if the data were located on their own machine. This is the type of functionality that must be supported within the proposed network of large-scale environmental observatories if they are to be community resources.

Data discovery and visualization tools such as the LBRTB Map Server and the Time Series Analyst provide potential data users with the ability to quickly screen data to find what they are most interested in. The linkage of the two and their accessibility within a Web browser makes the data more user-friendly to individuals who are not familiar with the Little Bear River watershed and also extends the reach of the data to individuals that may lack the skills to successfully use the web services.

The focus of this paper and the cyberinfrastructure components presented herein is on observational data measured at a point. The case study that we have presented provides an example of the types of software applications that are needed to manage the collection and publication of point observations data, and in particular observations made by in-situ environmental sensors. However, although important, point observations are only one class of water resources data that will be important for establishing environmental observatories. Spatially distributed data such as radar rainfall data and other remote sensing products are examples of data that are not addressed by the tools described in this paper. Like point measurements, these datasets represent observations at a point in time but across a spatial field, and future work is needed to provide infrastructure for storing, visualizing, analyzing, and publishing these data.

Other important cyberinfrastructure for environmental observatories will include applications that support advanced data analysis and modeling. Our current ability to predict hydrologic and water quality responses is constrained by our inability to test many of the concepts and assumptions that are the basis of our current understanding of hydrological processes (as embodied in the currently available suite of models) [*Grayson and Blöschl*, 2000; *Woods et al.*, 2001]. This is in part due to the lack of data collected at spatial and temporal scales that are consistent with these processes. Many believe that the next generation of hydrologic and water quality models will be driven by high-frequency data generated by coordinated, extensive field studies (such as those that will be conducted within proposed environmental observatories) [*Woods et al.*, 2001; *Kirchner*, 2006; *Hart and Martinez*, 2006]. Data collection and publication are the first steps toward making these types of data available to the community, and, although not specifically addressed by this paper, the use of observatory data to support modeling and advanced data analysis applications will be enhanced by the publication of data using standard exchange formats.

5.6. Software Availability

ODM, the ODM Streaming Data Loader, ODM Tools, and the WaterOneFlow web services were developed under the Berkeley Software Distribution License. Installation files, source code, and documentation can be accessed free of charge through the CUAHSI HIS website http://his.cuahsi.org. Source code for the LBRTB Map Server and the Time Series Analyst can be acquired by contacting the corresponding author at jeff.horsburgh@usu.edu.

References

Beran, B., and M. Piasecki (2008), Engineering new paths to water data, *Computers and Geosciences*, In press. doi:10.1016/j.cageo.2008.02.017.

Bose, R. (2002), A conceptual framework for composing and managing scientific data lineage, in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pp. 15-19, IEEE Press, Pascataway, N. J.

Colomb, R. L. (1997), Impact of semantic heterogeneity on federating databases, *The Computer Journal, 40*(5), 235-244, doi:10.1093/comjnl/40.5.235.

Connolly, T., and C. Begg (2005), *Database Systems A Practical Approach to Design, Implementation, and Management*, 1374 pp., 4th ed., Addison-Wesley, Harlow, U. K.

Droegemeier, K. K., K. Brewster, M. Xue, D. Weber, D. Gannon, B. Plale, D. Reed, L. Ramakrishnan, J. Alameda, R. Wilhelmson, T. Baltzer, B. Domenico, D. Murray, M. Ramamurthy, A. Wilson, R. Clark, S. Yalda, S. Graves, R. Ramachandran, J. Rushing, and E. Joseph (2005), Service-oriented environments for dynamically interacting with mesoscale weather, *Computing in Science & Engineering*, *7*(6), 12-29, doi:10.1109/MCSE.2005.124.

Goodall, J. L., J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky (2008), A first approach to Web services for the National Water Information System, *Environ. Model. & Software*, *23*(4), 404-411, doi:10.1016/j.envsoft.2007.01.005.

Gray, J., D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber (2005), Scientific data management in the coming decade, *SIGMOD Rec.*, *34*(4), 34-41, doi:10.1145/1107499.1107503.

Grayson, R. B., and G. Blöschl (Eds.) (2000), *Spatial Patterns in Catchment Hydrology: Observations and Modeling*, 423 pp., Cambridge University Press, Cambridge, U. K.

Hart, J. K., and K. Martinez (2006), Environmental sensor networks: A revolution in earth system science?, *Earth-Science Reviews*, *78*, 177-191, doi:10.1016/j.earscirev.2006.05.001.

Hill, D. J., B. Minsker, and E. Amir (2007), Real-time Bayesian anomaly detection for environmental sensor data, in *Proceedings: 32nd Congress of the International Association of Hydraulic Engineering and Research (IAHR 2007)*, Venice, Italy. (Available at http://reason.cs.uiuc.edu/eyal/papers/Bayesian-anomaly-sensor-IAHR07.pdf)

Jeong, S., Y. Liang, and X. Liang (2006), Design of an integrated data retrieval, analysis, and visualization system: Application in the hydrology domain, *Environmental Modelling & Software*, *21*, 1722-1740, doi:10.1016/j.envsoft.2005.09.007.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resources Research*, *42*, W03S04, doi:10.1029/2005WR004362.

Liu, Y., B. Minsker, and D. Hill (2007), Cyberinfrastructure technologies to support QA/QC and event-driven analysis of distributed sensing data, in *Proceedings: International Workshop on Advances in Hydroinformatics*, Niagara Falls, Canada. (Available at http://colab.ncsa.uiuc.edu/EESHG1/Documents/repository/conference/CyberinfrastructureTechQAQC.pdf)

Maidment, D. R. (Ed.) (2005), *Hydrologic Information System Status Report, Version 1*, 224 pp., Consortium of Univ. for the Adv. of Hydrol. Sci., Washington, D. C. (Available at http://www.cuahsi.org/docs/HISStatusSept15.pdf)

Maidment, D. R. (Ed.) (2008), *CUAHSI Hydrologic Information System: Overview of Version 1.1*, 92 pp., Consortium of Univ. for the Adv. of Hydrol. Sci., Washington, D. C. (Available at http://his.cuahsi.org/documents/HISOverview.pdf)

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford (1997), Nongeospatial metadata for the ecological sciences, *Ecol. Appl.*, *7*(1), 330-342, doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

Montgomery, J. L., T. Harmon, W. Kaiser, A. Sanderson, C. N. Haas, R. Hooper, B. Minsker, J. Schnoor, N. L. Clesceri, W. Graham, and P. Brezonik (2007), The WATERS Network: An integrated environmental observatory network for water research, *Environmental Science and Technology*, *41*(19), 6642-6647. (Available at http://pubs.acs.org/subscribe/journals/esthag/41/i19/pdf/100107feature_waters.pdf)

Mourad, M., and J. L. Bertrand-Krajewski (2002), A method for automatic validation of long time series of data in urban hydrology, *Water Sci. & Technology*, *45*(4-5), 263-270.

National Research Council (2008), *Integrating Multiscale Observations of U.S. Waters*, Committee on Integrated Observations for Hydrologic and Related Sciences, Water Science and Technology Board, Division on Earth and Life Studies, The National Academies Press, Washington, D. C.

National Science Foundation (2007), Cyberinfrastructure vision for 21st century discovery, National Science Foundation Cyberinfrastructure Council, NSF 07-28. (Available at http://www.nsf.gov/pubs/2007/nsf0728/index.jsp)

Sheth, A. P., and J. A. Larson (1990), Federated database systems for managing distributed, heterogeneous, and autonomous databases, *Computing Surveys*, *22*(3), doi:10.1145/96602.96604.

Tomasic, A., and E. Simon (1997), Improving access to environmental data using context information, *SIGMOD Rec.*, *26*(1), 11-15, doi:10.1145/248603.248606.

Valentine, D., I. Zaslavsky, T. Whitenack, and D. R. Maidment (2007), Design and implementation of CUAHSI WATERML and WaterOneFlow web services, Extended Abstract, Published Collection: Geoinformatics 2007 Bibliography. (Available at http://gsa.confex.com/gsa/2007GE/finalprogram/abstract_122329.htm)

WATERS Network (2008), Draft science, education, and design strategy for the WATer and Environmental Research Systems Network, WATERS Network Project Office. (Available at http://www.watersnet.org/docs/SEDS-20080227-draft.pdf)

Whitenack, T., I. Zaslavsky, D. Valentine, and D. Djokic (2007), Data Access System for Hydrology, Abstract, Published Bibliography: *Eos Trans. AGU*, *88*(52), Fall Meet. Suppl., Abstract H13H-1685.

Woods, R. A., R. B. Grayson, A. W. Western, M. J. Duncan, D. J. Wilson, R. I. Young, R. P. Ibbitt, R. D. Henderson, and T. A. McMahon (2001), Experimental design and initial results from the Mahurangi River Variability Experiment: MARVEX, in *Observations and Modelling of Land Surface Hydrological Processes*, edited by V. Lakshmi, J. D. Albertson and J. Schaake, pp. 201-213, Water Resources Monographs, American Geophysical Union, Washington, D. C.

Youn, C., C. Baru, K. Bhatia, S. Chandra, K. Lin, A. Memon, G. Memon, and D. Seber (2007), GEONGrid portal: design and implementations, *Concurrency and Computation: Practice and Experience*, *19*(12), 1597-1607, doi:10.1002/cpe.1129.

Zaslavsky, I., D. Valentine, and T. Whiteaker (Eds.) (2007), CUAHSI WaterML, OGC Discussion Paper, OGC 07-041r1, Version 0.3.0. (Available at http://www.opengeospatial.org/standards/dp)

Table 5.1. Monitoring Sites in the LBRTB

Site Number	Site Name	Latitude	Longitude	Site Type
1	Upper South Fork Little Bear River	41.4954	-111.818	Stream
2	Lower South Fork Little Bear River	41.5065	-111.8151	Stream
3	East Fork Little Bear River	41.5292	-111.7993	Stream
4	Little Bear River below Confluence of East and South Forks	41.5361	-111.8305	Stream
5	Little Bear River near Paradise	41.5756	-111.8552	Stream
6	Little Bear River near Wellsville	41.6435	-111.9176	Stream
7	Little Bear River near Mendon	41.7185	-111.9464	Stream
8	Lower Watershed Weather Station	41.667	-111.8906	Weather
9	Upper Watershed Weather Station	41.5355	-111.8059	Weather

Table 5.2. Sensor Specifications for the LBRTB Monitoring Sites

Site Type	Data Collection Frequency	Variable	Sensor
Stream	30 minutes	Water Temperature	Hydrolab MiniSonde 5 thermistor
		Dissolved Oxygen	Hydrolab MiniSonde 5 optical dissolved oxygen sensor
		pH	Hydrolab MiniSonde 5 reference electrode
		Specific Conductance	Hydrolab MiniSonde 5 four electrode conductivity sensor
		Turbidity	Forest Technology Systems DTS-12 Turbidity Sensor
		Stage	KWK Technologies SPXD-600 Pressure Transducer
Weather	1 hour	Air Temperature	Campbell Scientific CS215 temperature and relative humidity sensor
		Relative Humidity	Campbell Scientific CS215 temperature and relative humidity sensor
		Solar Radiation	Apogee PYR-P silicon pyranometer
		Precipitation	Texas Electronics TE25 tipping bucket
		Barometric Pressure	Setra 278 barometric pressure sensor
		Wind Speed	R. M. Young Wind Sentry Set
		Wind Direction	R. M. Young Wind Sentry Set

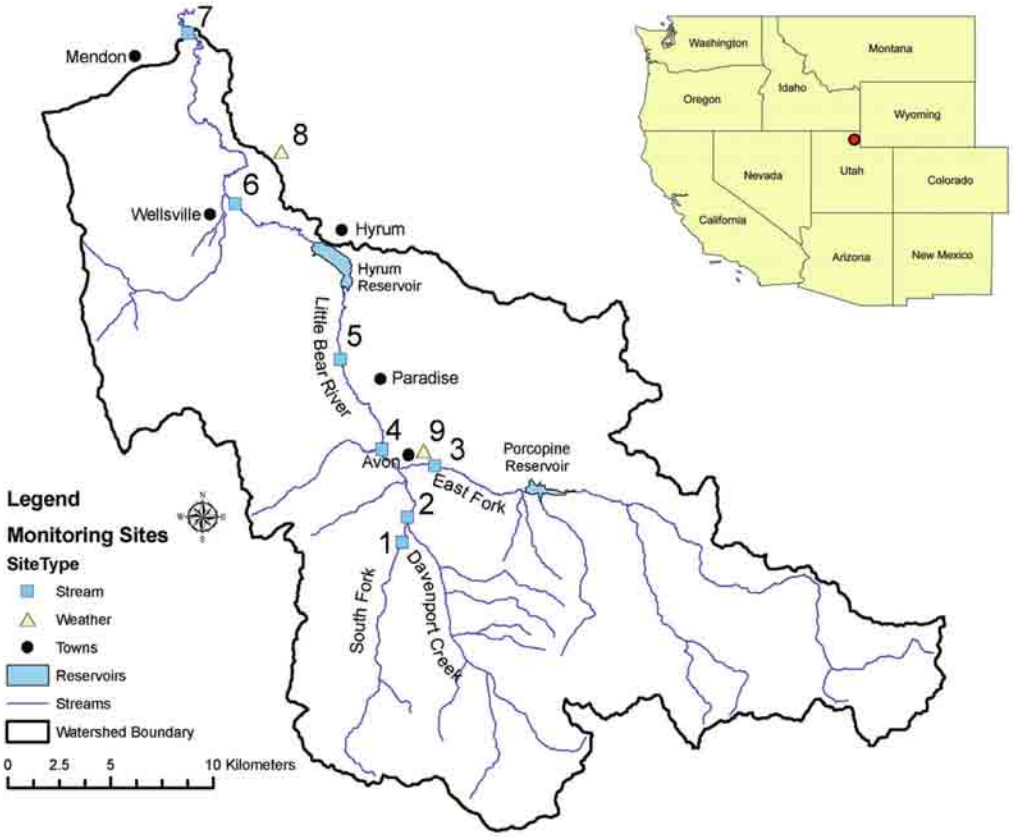


Figure 5.1. Little Bear River test bed monitoring site locations.

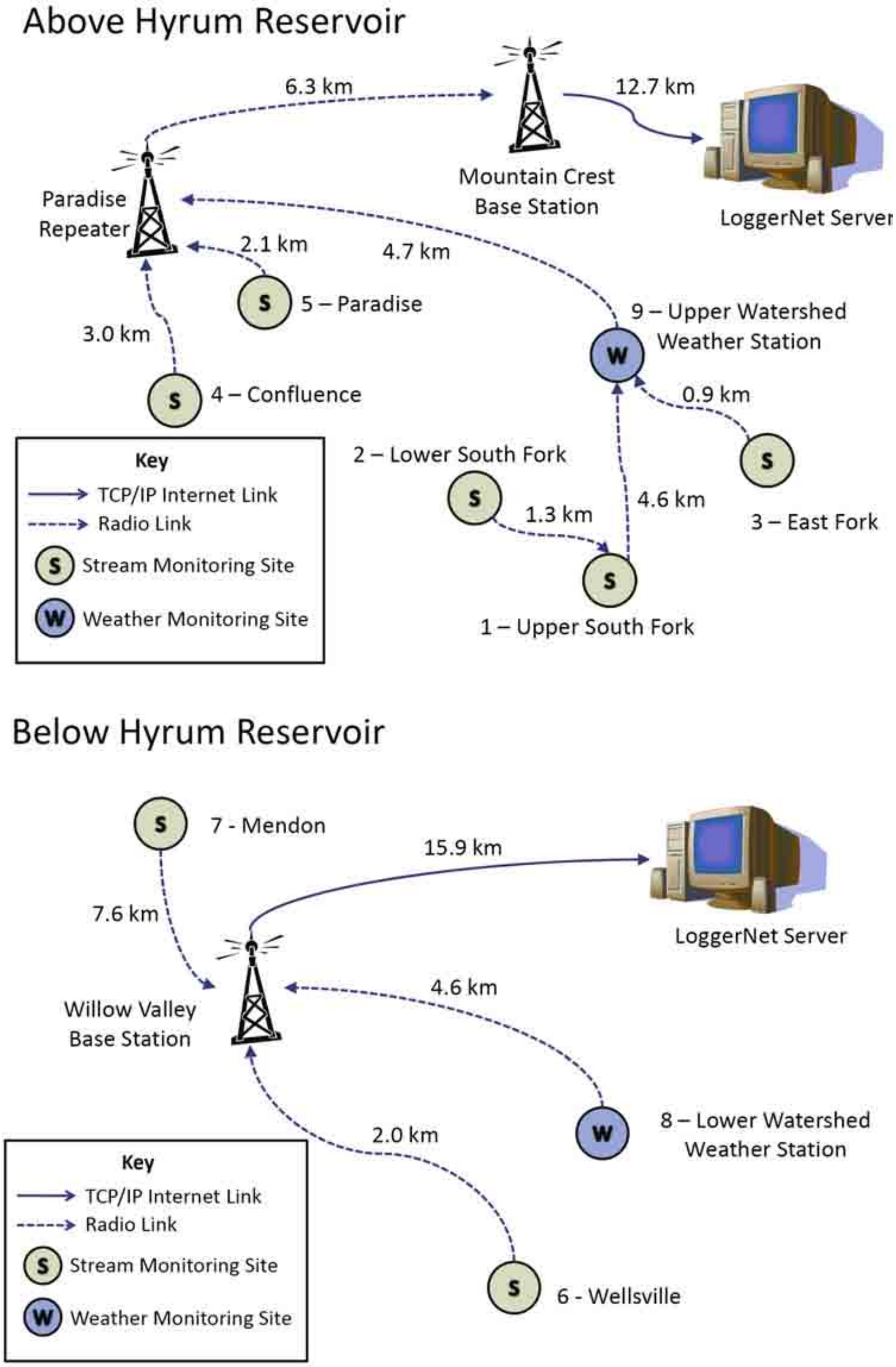
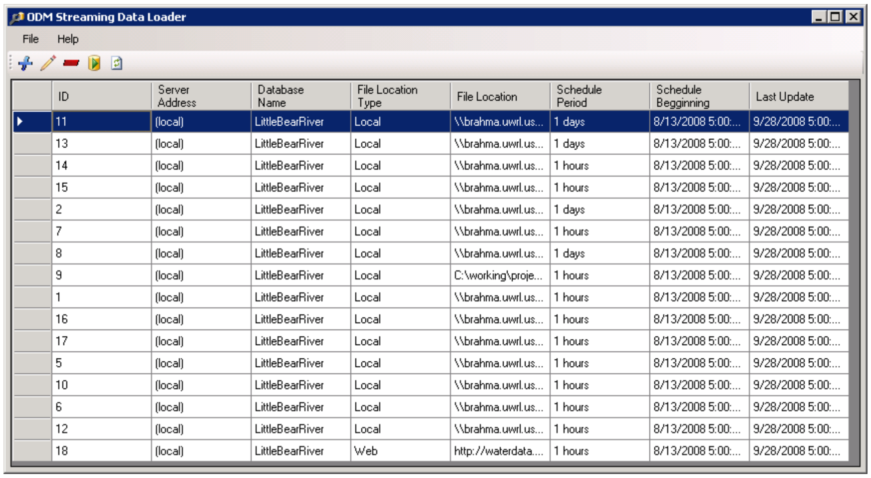


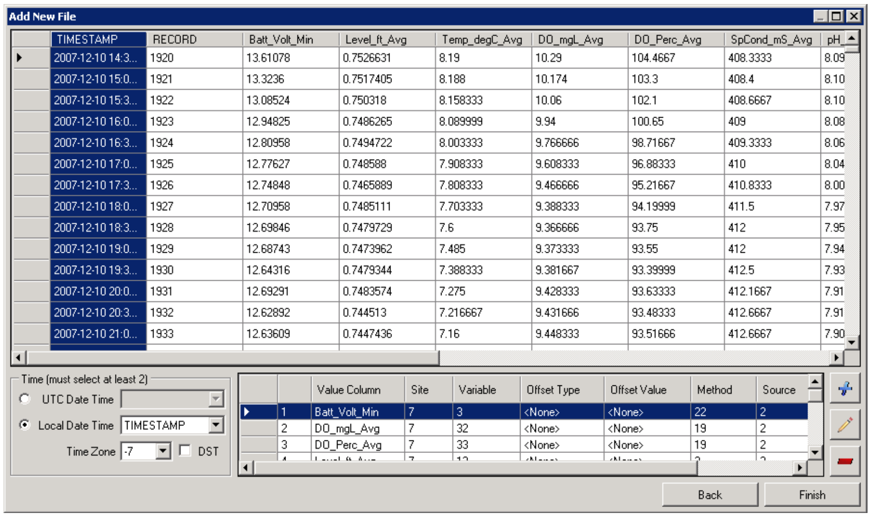
Figure 5.2. Little Bear River sensor network map.



ODM Streaming Data Loader

ID	Server Address	Database Name	File Location Type	File Location	Schedule Period	Schedule Beggining	Last Update
11	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 days	8/13/2008 5:00:...	9/28/2008 5:00:...
13	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 days	8/13/2008 5:00:...	9/28/2008 5:00:...
14	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
15	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
2	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 days	8/13/2008 5:00:...	9/28/2008 5:00:...
7	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
8	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 days	8/13/2008 5:00:...	9/28/2008 5:00:...
9	(local)	LittleBearRiver	Local	C:\working\proje...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
1	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
16	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
17	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
5	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
10	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
6	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
12	(local)	LittleBearRiver	Local	\\brahma.uwrl.us...	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...
18	(local)	LittleBearRiver	Web	http://waterdata....	1 hours	8/13/2008 5:00:...	9/28/2008 5:00:...

(a)



Add New File

TIMESTAMP	RECORD	Batt_Volt_Min	Level_ft_Avg	Temp_degC_Avg	DO_mgL_Avg	DO_Perc_Avg	SpCond_mS_Avg	pH_
2007-12-10 14:3...	1920	13.61078	0.7526631	8.19	10.29	104.4667	408.3333	8.09
2007-12-10 15:0...	1921	13.3236	0.7517405	8.188	10.174	103.3	408.4	8.10
2007-12-10 15:3...	1922	13.08524	0.750318	8.158333	10.06	102.1	408.6667	8.10
2007-12-10 16:0...	1923	12.94825	0.7486265	8.089999	9.94	100.65	409	8.08
2007-12-10 16:3...	1924	12.80958	0.7494722	8.003333	9.766666	98.71667	409.3333	8.06
2007-12-10 17:0...	1925	12.77627	0.748588	7.908333	9.608333	96.88333	410	8.04
2007-12-10 17:3...	1926	12.74848	0.7465889	7.808333	9.466666	95.21667	410.8333	8.00
2007-12-10 18:0...	1927	12.70958	0.7485111	7.703333	9.388333	94.19999	411.5	7.97
2007-12-10 18:3...	1928	12.69846	0.7479729	7.6	9.366666	93.75	412	7.95
2007-12-10 19:0...	1929	12.68743	0.7473962	7.485	9.373333	93.55	412	7.94
2007-12-10 19:3...	1930	12.64316	0.7479344	7.388333	9.381667	93.39999	412.5	7.93
2007-12-10 20:0...	1931	12.69291	0.7483574	7.275	9.428333	93.63333	412.1667	7.91
2007-12-10 20:3...	1932	12.62892	0.744513	7.216667	9.431666	93.48333	412.6667	7.91
2007-12-10 21:0...	1933	12.63609	0.7447436	7.16	9.448333	93.51666	412.6667	7.90

Time (must select at least 2)

○ UTC Date Time

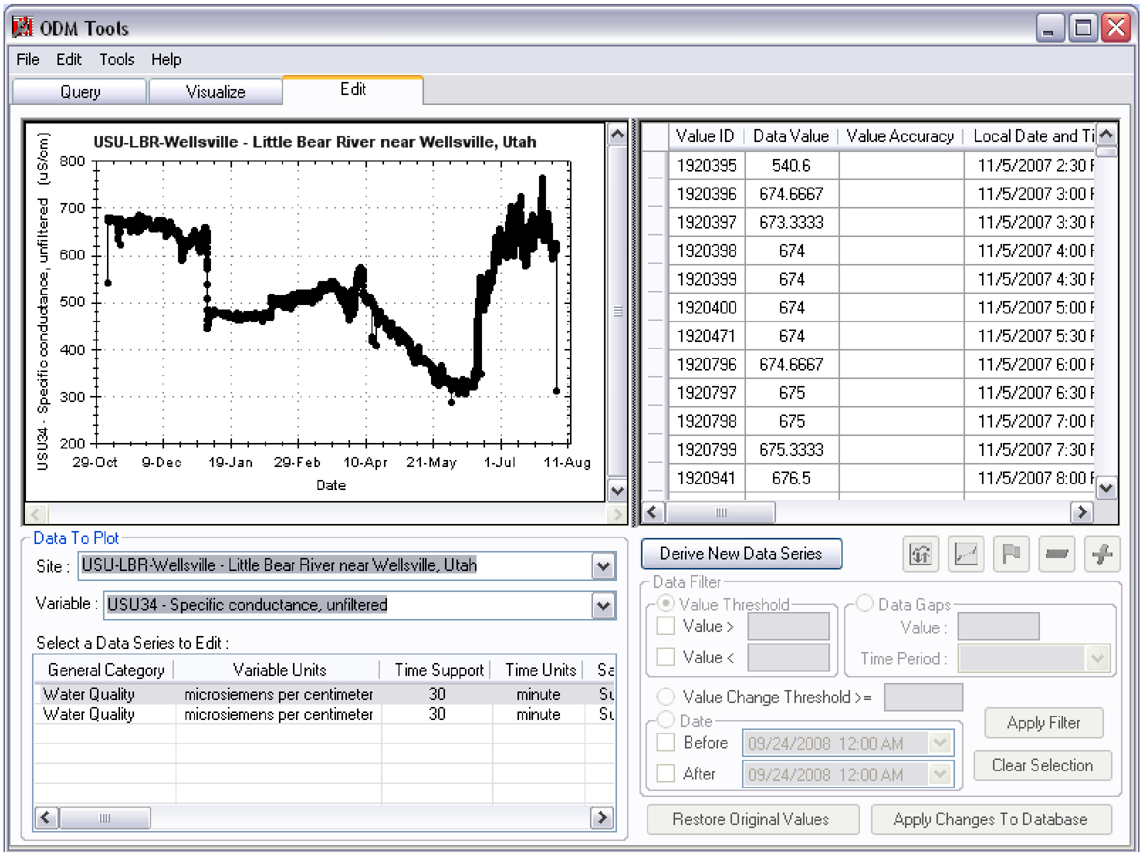
● Local Date Time: TIMESTAMP

Time Zone: -7 ☐ DST

	Value Column	Site	Variable	Offset Type	Offset Value	Method	Source
1	Batt_Volt_Min	7	3	<None>	<None>	22	2
2	DO_mgL_Avg	7	32	<None>	<None>	19	2
3	DO_Perc_Avg	7	33	<None>	<None>	19	2

(b)

Figure 5.3. The ODM Streaming Data Loader wizard-based graphical user interface. Panel (a) shows the listing of datalogger files that have been mapped and scheduled to be loaded into the LBRTB ODM database. Panel (b) shows the interface for mapping the individual columns in a single datalogger file to the ODM schema.



Value ID	Data Value	Value Accuracy	Local Date and Ti
1920395	540.6		11/5/2007 2:30 F
1920396	674.6667		11/5/2007 3:00 F
1920397	673.3333		11/5/2007 3:30 F
1920398	674		11/5/2007 4:00 F
1920399	674		11/5/2007 4:30 F
1920400	674		11/5/2007 5:00 F
1920471	674		11/5/2007 5:30 F
1920796	674.6667		11/5/2007 6:00 F
1920797	675		11/5/2007 6:30 F
1920798	675		11/5/2007 7:00 F
1920799	675.3333		11/5/2007 7:30 F
1920941	676.5		11/5/2007 8:00 F

Site : USU-LBR-Wellsville - Little Bear River near Wellsville, Utah

Variable : USU34 - Specific conductance, unfiltered

Select a Data Series to Edit :

General Category	Variable Units	Time Support	Time Units	Sa
Water Quality	microsiemens per centimeter	30	minute	Su
Water Quality	microsiemens per centimeter	30	minute	Su

● Value Threshold
□ Value >
□ Value <

○ Data Gaps
Value :
Time Period :

○ Value Change Threshold >=

○ Date
□ Before 09/24/2008 12:00 AM
□ After 09/24/2008 12:00 AM

Figure 5.4. ODM Tools data editing interface.

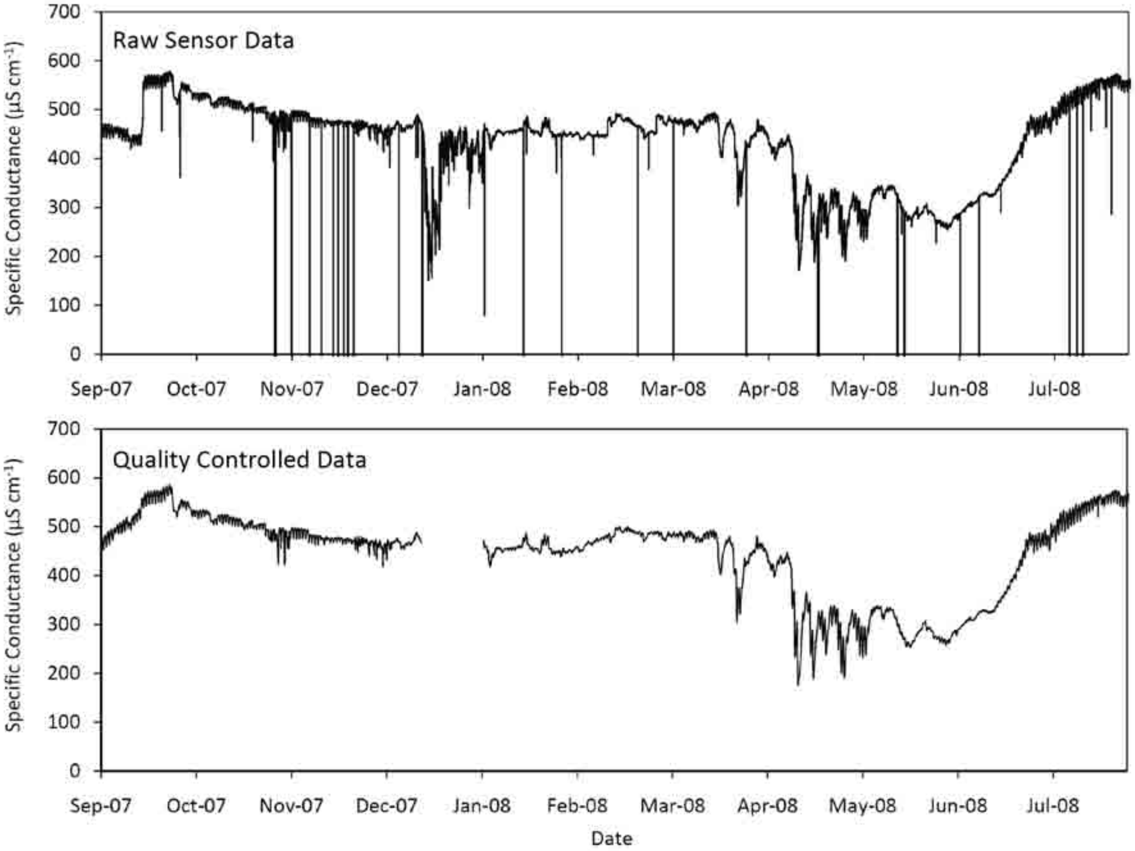


Figure 5.5. Example specific conductance data series from the Paradise monitoring site before and after quality control editing using ODM Tools.

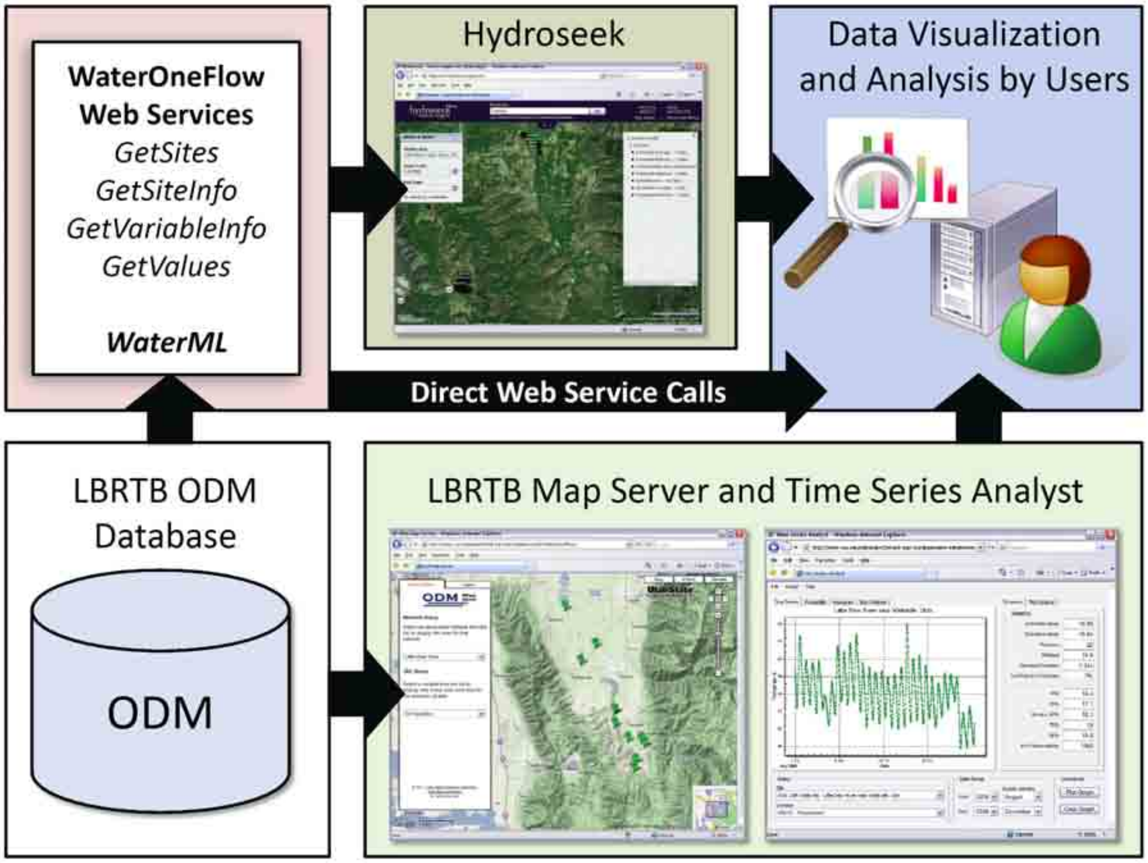


Figure 5.6. Data discovery, visualization, and analysis components of the LBRTB observatory information system.

CHAPTER 6

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The research described in this dissertation aims to address the growing need within the hydrologic and environmental engineering communities for cyberinfrastructure that supports coordinated, intensive field studies that are generating vast quantities of observational data. This need is primarily driven by the realization that we have moved beyond the time when all of our data fit in a simple spreadsheet that we could email to our colleagues and when we could visualize all of our data in a few simple summary plots. As the amount and complexity of data grows, resulting from the increasing use of sensor networks as general tools in scientific research and as large scale environmental observatories come online, advanced methods for data management, visualization, analysis, and publication are required to support and enable the use of the growing volume of observational data.

The results of our analyses of sensor data collected in the Little Bear River demonstrate the need for and value of high-frequency, continuous time series of discharge and hydrochemical variables. The observing system, surrogate methods, and cyberinfrastructure that we have demonstrated are advances to the infrastructure available for the design and implementation of environmental observatories and together have enabled us to gain insights into the importance and relative magnitude of water and constituent fluxes from different hydrologic pathways that are only possible through high-frequency data. Data and analyses such as these, as well as the cyberinfrastructure that enabled them, make it possible for us to better understand the way that water and water-borne constituents move through a watershed.

The software resulting from this research has also enhanced the available infrastructure for supporting environmental observatories and has already impacted a diverse community. The cyberinfrastructure components described in this dissertation show how the syntactic and semantic heterogeneity in data from different experimental sites and sources can be overcome and how data collectors can publish their observations so that they can easily be accessed and interpreted by others. Indeed, the tools and methods described represent a new opportunity for many within the water resources community to approach the organization, management, publication, and analysis of their data systematically. In most cases this will likely mean moving from collections of ASCII text or spreadsheet files to a system that enables better organization, better management, better documentation, and better distribution of research data.

The engineering significance of this work lies not only in the development of software tools and methods capable of handling large quantities of observational data, but also in the fact that the availability of these tools has sparked a concerted effort within the hydrologic science and environmental engineering communities to publish academic research data. The capabilities developed provide researchers with a standard method for publishing observational data that enables interoperability among published datasets. With the methods developed, results from projects spanning multiple research sites and collected at different times can be combined to perform analyses that may lead to better understanding of hydrologic processes than would be obtained from the individual sites alone. Thus the potential now exists to advance environmental and earth sciences significantly through the publication and subsequent reanalysis or recombination of research data.

Chapters 2 through 5 of this dissertation present the main results of this research and are focused on three objectives that guided this work. These objectives were chosen to address three very high level categories of cyberinfrastructure functionality required to support environmental observatories, namely: 1) data collection; 2) persistent data storage and management; and 3) data publication. The specific developments and case studies under each of these objectives were framed around creating cyberinfrastructure to support research in the Little Bear River environmental observatory test bed (LBRTB). The first objective was to establish a wireless sensor network that would provide high-frequency data for generating estimates of water quality constituent fluxes and investigating the hydrologic and hydrochemical responses within the LBRTB. The second objective was to design a generic data model for point observations, and the third objective was to create an integrated observatory information system for the LBRTB.

In Chapter 2 we describe the physical setting of the LBRTB, the experimental and sensor network design, and the data collection procedures that were implemented to enable generation of high frequency estimates of total phosphorus (TP) and total suspended solids (TSS) concentrations as well as investigation of the hydrologic and hydrochemical responses in the Little Bear River. This research has demonstrated the observing system and cyberinfrastructure required to more effectively quantify spatial and temporal variability in water quality constituent fluxes (especially for water quality constituents for which no in situ sensors exist) and demonstrates how high-frequency sensor data collected at multiple sites can provide multiple lines of evidence to improve hydrologic and hydrochemical process understanding.

Our analyses of the variability in hydrology and hydrochemistry in the Little Bear using the high-frequency data speak to the heart of the issues that are driving the push toward establishment of large scale environmental observatories. Coupled with generation of surrogate relationships, hydrograph separations, and evaluation of stream metabolism, the high-frequency data collected in the LBRTB suggest first that the spring snowmelt period is the dominant TSS and TP load generation period in the Little Bear River watershed and that the period of early snowmelt generates the majority of the annual TSS and TP load. Second, water quality constituent loads estimated using weekly or monthly data are not representative of the high variability in discharge and constituent concentrations, tend to more frequently under predict the true loading because of the high probability that peaks in discharge and concentration are missed, and should be considered as order of magnitude estimates of the true loading. Third, the contribution of slow subsurface pathways (i.e., baseflow) are relatively constant throughout the year and do not extend to a great degree into the peaks of the spring snowmelt hydrograph. Fast pathways (i.e., quickflow that primarily results from snowmelt) contribute more than half of the annual discharge and dominate the spring snowmelt hydrograph. The chemical signatures of baseflow and quickflow appear to be distinct, suggesting that the two flow paths have very different residence times within the system. These general characteristics may be true of many snowmelt driven watersheds that are similar to the Little Bear River. Fourth, estimates of photosynthesis and respiration rates are useful indicators of instream metabolism that can provide information about the physical, chemical, and biological differences between sites and may provide a useful indicator of the degree to which they have been affected by human disturbance.

Beyond the analyses that we have completed, we are now looking for answers to questions (e.g., Why are there diurnal fluctuations in turbidity?, What causes diurnal variability in discharge and specific conductance during the summer low flow conditions?, and Why do some sites follow the assumptions of a conceptual model for dissolved oxygen while others don't?) that we might not have even thought of before we started collecting continuous data. Data such as the ones we have collected in the Little Bear River, when collected within a variety of different catchments, as is planned for the network of large scale environmental observatories, will enable us to better understand the processes that control the fluxes, flow paths, and stores of both water and water-borne constituents.

In the Little Bear, we were able to test our conceptual model of discharge and quantify its make up as a combination of baseflow and quickflow primarily from snowmelt. We were able to test our understanding and assumptions about the temporal distribution of TSS and TP loading and draw inferences about the sources and pathways carrying these constituents to the stream. We were also able to quantify the magnitude of stream photosynthesis and respiration rates and compare them across monitoring sites. Additional data collection and analyses will enable us to extend these analyses to examine and better quantify the effects of human modification and land use change on hydrologic and hydrochemical response. For example, an additional research question might be: what is the magnitude of changes in photosynthesis and respiration rates as we move from the top of the Little Bear River, which is relatively pristine, to the bottom, where the river is highly influenced by agricultural lands, and how are these changes influenced by agricultural diversions, reservoir releases, and agricultural runoff and return flows. This question requires quantification of physical, chemical, and biological characteristics of the Little Bear River, and without the types of data that we have collected our ability to answer these types of questions and to better predict what those changes will be in the future remains constrained.

Motivated by the task of storing and managing the large quantities of data generated by the sensor network deployed in the LBRTB, and realizing that this was a general problem related to the use of large quantities of observational data, we created a generic Observations Data Model (ODM) that can be used for persistently storing both the observational data and its supporting metadata in a relational database. Indeed, our experience with the data collection aspects of the LBRTB provided critical experience that informed our work in developing the cyberinfrastructure components described in this dissertation. Chapter 3 presents the logical design for ODM and describes its features and functionality. ODM is a relational data model that preserves the context and provenance of data, the importance of which cannot be overstated if data are to be published.

ODM provides a framework in which data of different types and from disparate sources can be integrated. For example, data from multiple scientific disciplines can be assembled within a single ODM instance (e.g., hydrologic variables, water quality variables, climate variables, etc.), which can greatly facilitate their use within common analyses. Not only can the data be standardized and appropriately qualified with metadata, but applications that interact with ODM can be harmonized, leading to greater cooperation, sharing (of both data and application code), and interoperability.

The LBRTB is one of 11 environmental observatory test beds located across the United States that are part of the Water and Environmental Research Systems (WATERS) network. Data managers within each of the test beds were charged with publishing their data in a consistent format, thereby creating a consistent and interoperable network of scientific data. Chapter 4 describes the major components of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) data publication system, which was adopted by the test beds to accomplish the goal of consistently published data. Consisting of ODM, the ODM controlled vocabulary system, the WaterOneFlow web services, and a central web services registry, the HIS data publication system has provided the test bed data managers with the infrastructure needed to integrate their data, while overcoming the considerable syntactic and semantic heterogeneity in data from each source.

The HIS data publication system’s flexibility in storing and enabling public access to similarly formatted data and metadata from the test beds has created a community data resource from data that might otherwise have been confined to the private files of the individual investigators and serves as a prototype for the infrastructure that will be required to support a network of large scale environmental observatories. The current data publication network enabled by ODM and the data publication system described in Chapters 3 and 4 already consists of more than 3,700 data collection sites, nearly 800 measured variables, and nearly 42 million individual data values, many of which have been contributed by investigators outside of the original network of 11 test beds.

Last, in Chapter 5 we discuss the overall framework for an observatory information system and describe components in addition to those described in Chapter 4 that complete the observatory information system for the LBRTB. These are value added tools that make the job of data publication easier and provide Internet-based tools for accessing, visualizing, and analyzing the data. The Little Bear River observatory information system demonstrates the tools needed to create automated integration between sensors in a sensor network and a central observations database for storing and managing the resulting data. It also demonstrates mechanisms for and technology that enables the storage and archival of observations data and the open and free distribution of the data via simple to use, Internet-based tools.

Although the tools and methods described in this dissertation address many of the challenges associated with developing cyberinfrastructure for environmental observatories, many challenges still remain. The design and setup of appropriate sensor networks is certainly one challenge, especially since sensors do not currently exist to measure all of variables in which we are interested. In the short term, the absence of in-situ sensors for important constituents presents both challenges and opportunities for using existing sensors as surrogates, as we have done with turbidity for TSS and TP. The cyberinfrastructure challenge is in designing appropriate models and relationships between constituents and then integrating these with the monitoring and data publication systems. Additional challenges lie in quantifying the uncertainty in estimates of constituents based on surrogate relationships.

In the longer term, research is needed to create new and robust sensor technology for constituents that we cannot currently measure in-situ. This is happening as the market for them and the importance for quantifying these constituents grows (examples include newer ion specific electrodes and optical techniques), but more collaboration is needed between domain scientists and sensor manufacturers to speed up this process that seems to have been slow relative to technological advances realized in other fields (e.g., consumer electronics). New sensors should also consider technologies for increasing reliability, extending deployment periods, and reducing maintenance requirements for deployments in harsh environments (e.g., streams, soil, etc.). New sensors need to use less power, be less prone to drift and calibration issues, and be more resistant to bio-fouling or other environmental conditions.

Designing and implementing the communications infrastructure to support sensor networks and ensure the timely reporting of data is something that currently requires a great deal of expertise, and there is also a need to simplify and standardize aspects of the design, setup, configuration, programming, deployment, and maintenance of sensor network components. Improvements in the software and hardware that support sensor networks could also benefit from collaboration between domain scientists and the manufacturers of the hardware to make these components more user-friendly. However, some of the complexity in monitoring systems is caused by the environment being monitored and may not be easily overcome with better hardware or software. Establishing a line-of-sight radio communication network in complex terrain where multiple hops are needed to transmit data from remote sites is one example of this. The need for monitoring and communication systems is well established, but for these systems to reach their full potential it is likely that the next generation of scientists will need much more extensive training in developing observational infrastructure than students have received in the past. New courses will be needed that address emerging methods and technologies for acquisition of environmental observations data and the tools and techniques available for using and managing data to give students the foundation needed to implement the next generation of environmental studies and instrumented research sites.

Throughout the development of the tools described in this dissertation, we have promoted open and free access to data. However, there are inevitably datasets for which public access must be limited for security or sensitivity reasons. For the tools described in this dissertation to be generally applicable, they must address the issues of authentication and access constraints. It is likely that this will be required at multiple levels (i.e., as access restrictions within ODM databases and as required authentication and access levels within the WaterOneFlow web services that deliver data from ODM) so that users both within and outside of the organization publishing the data have appropriate access. This will require development of mechanisms within ODM to which access restrictions can be tied (e.g., fields in the database that contain appropriate access levels) and recognition of these within the WaterOneFlow web services so that queries only return data that are consistent with a user's permission level. This will also require development of mechanisms that support system administrators in implementing access policies that control which users have access to which levels of data (i.e., raw data or quality controlled data, or sensitive data restricted to particular users – e.g. human subjects data).

The focus of this research has been on observational data measured at a point (e.g., time series data collected at a stream monitoring site or weather station located at a fixed point in space). Although very important, point observations are not the only class of water resources data that will require development of infrastructure to support environmental observatories. Spatially distributed data such as radar rainfall data and other remote sensing products are examples of data that are not addressed by the tools described in this dissertation. Observations from moving platforms and sampling locations (i.e., measurements made from a moving boat) are another example of data that are not handled well. Like point measurements, these datasets represent observations at a point in time but across a spatial field or along a track or transect, and future work is needed to provide infrastructure for storing visualizing, analyzing, and publishing these data. Appropriate data models for representing these data will be required to enable their storage and manipulation, as will appropriate mechanisms to ensure that these data can be delivered to users. It may be that simple extensions to ODM and the WaterOneFlow web services will suffice, but it is likely that additional data models and different web service paradigms (e.g., the use of geospatial web services for representing feature geometry) will be required to support publication of spatially distributed data.

More research is also needed to provide context for point observations within a structured framework referred to as a “digital watershed.” According to *Maidment* [2005], a Digital Watershed is a fusion of point observation data, geographic information systems (GIS) data, remote sensing images, and weather and climate grid information linked to hydrologic simulation models. We envision a digital watershed as a structured collection of hydrologic objects (e.g., stream reaches, reservoirs, hillslopes, etc.) on which measurements are made. The relationships between the objects provide the context to facilitate integrated modeling and analysis. A digital watershed could be implemented as an object oriented data model that enables the relationships between important hydrologic objects to be expressed. For example, a hillslope object could be related to the stream reach into which it contributes flow or a stream reach could be associated with a downstream reservoir into which it flows. Conceptual models of the flow pathways between hydrologic objects and their associated fluxes of both water and water-borne constituents could then be applied and tested, creating an integrated representation of the behavior of a watershed and enabling tracking of the movement of water and water-borne constituents through the watershed.

In addition, observational data could be linked to the hydrologic objects that they represent (i.e., a stream gage could be related to the stream reach on which it is located, or a weather station could be related to the catchment in which it is located). Observations could then be used either to directly quantify fluxes between hydrologic objects or as inputs to models that define the fluxes. For example, where a stream reach flows into a reservoir, a stream gage may record the discharge. The data collected at the stream gage directly quantifies the stream discharge to the reservoir.

A digital watershed could encapsulate portions of the cyberinfrastructure described in this dissertation. In the example above, the stream gage would be represented as a point with a geographic location within the digital watershed. The digital watershed could store the observational data for the gage within an ODM database, or it might just store the information required to retrieve published data for the gage using web services.

The design of such a system will require careful consideration and identification of the important physical features of a watershed (e.g., hillslopes, stream reaches, vadose zone, groundwater), their geographic representation (e.g., points, lines, polygons, raster fields, or 3-dimensional volumes), the important flow paths and processes that link them together (e.g., surface runoff from hillslopes to streams, or exchange between the stream and the hyporheic zone), how observational data are associated with the objects that they represent (e.g., stream gages as points on a stream reach, or radar rainfall grids as rasters over a watershed area), and how they can be used to represent fluxes between hydrologic objects. It will also require careful consideration of how the information that defines the hydrologic objects, relationships between objects, and observations that are associated with objects are stored and manipulated. GIS technology is an obvious mechanism for implementation of a digital watershed because of its capability to juxtapose spatial representation of real world features with data that characterize those features in related databases.

Finally, the cultural challenge of getting investigators to participate in data publication efforts remains. Our current system of reward for publishing research results is heavily weighted toward archiving papers that include interpretations of data (i.e., condensed tables and figures) in peer reviewed journals, as opposed to archiving the data themselves, which in most cases remain in the private files of the investigators. Few research proposals are written that articulate a plan for long term management of the data that will be generated. Realization of scientific advancements resulting from reanalysis or reinterpretation of existing data will likely require a cultural change toward the publication of research data, but this will certainly be facilitated by the availability of cyberinfrastructure tools.

References

Maidment, D. R. (Ed.) (2005), *Hydrologic Information System Status Report, Version 1*, 224 pp., Consorium of Univ. for the Adv. Of Hydrol. Sci., Washington, D. C. (Available at http://www.cuahsi.org/docs/HISStatusSept15.pdf)

APPENDICES

Appendix A

Coauthor Approval Letters

UtahState
University

Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

10-30-2008

Amber Spackman Jones
Utah Water Research Laboratory
Utah State University
8200 Old Main Hill
Logan, UT 84322-8200

Dear Amber,

I am in the process of preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008.

I am requesting your permission to include the attached paper, of which you are a coauthor, as a chapter in my dissertation. I will include acknowledgments to your contributions as indicated. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you have any questions, please contact me.

Thank you,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to use and reprint all of the material that I have contributed to Chapter 2 of his dissertation.

____
Amber Spackman Jones

UtahState
University

Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

10-30-2008

Nancy Mesner
College of Natural Resources
Utah State University
5200 Old Main Hill
Logan, UT 84322-5200

Dear Nancy,

I am in the process of preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008.

I am requesting your permission to include the attached paper, of which you are a coauthor, as a chapter in my dissertation. I will include acknowledgments to your contributions as indicated. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you have any questions, please contact me.

Thank you,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to use and reprint all of the material that I have contributed to Chapter 2 of his dissertation.

____

Nancy Mesner

UtahState University

Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

10-30-2008

Ilya Zaslavsky
University of California, San Diego
San Diego Supercomputer Center, MC 0505
9500 Gilman Drive
La Jolla, CA 92093-0505

Dear Ilya,

I am in the process of preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008.

I am requesting your permission to include the attached papers, of which you are a coauthor, as chapters in my dissertation. I will include acknowledgments to your contributions as indicated. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you have any questions, please contact me.

Thank you,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to use and reprint all of the material that I have contributed to Chapters 3 and 4 of his dissertation.

____
Ilya Zaslavsky

UtahState
University

Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

10-30-2008

Michael Piasecki
Drexel University
Department of Civil, Architectural & Environmental Engineering
3141 Chestnut Street,
Philadelphia, PA 19104

Dear Michael,

I am in the process of preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008.

I am requesting your permission to include the attached paper, of which you are a coauthor, as a chapter in my dissertation. I will include acknowledgments to your contributions as indicated. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you have any questions, please contact me.

Thank you,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to use and reprint all of the material that I have contributed to Chapter 4 of his dissertation.

____
Michael Piasecki

UtahState University

Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

10-30-2008

David Valentine
University of California, San Diego
San Diego Supercomputer Center, MC 0505
9500 Gilman Drive
La Jolla, CA 92093-0505

Dear David,

I am in the process of preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008.

I am requesting your permission to include the attached paper, of which you are a coauthor, as a chapter in my dissertation. I will include acknowledgments to your contributions as indicated. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you have any questions, please contact me.

Thank you,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to use and reprint all of the material that I have contributed to Chapter 4 of his dissertation.

____

David Valentine

UtahState
University

Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

10-30-2008

Thomas Whitenack
University of California, San Diego
San Diego Supercomputer Center, MC 0505
9500 Gilman Drive
La Jolla, CA 92093-0505

Dear Thomas,

I am in the process of preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008.

I am requesting your permission to include the attached paper, of which you are a coauthor, as a chapter in my dissertation. I will include acknowledgments to your contributions as indicated. Please advise me of any changes you require.

Please indicate your approval of this request by signing in the space provided, attaching any other form or instruction necessary to confirm permission. If you have any questions, please contact me.

Thank you,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to use and reprint all of the material that I have contributed to Chapter 4 of his dissertation.

____
Thomas Whitenack

Appendix B

Permission to Reprint Chapter 3

UtahState
University
Department of Civil and Environmental Engineering
4110 Old Main Hill
Logan, UT 84322-4110
Telephone: (435) 797-2932
Fax: (435) 797-1185

Jeffery S. Horsburgh
8200 Old Main Hill
Logan, UT 84322-8200
Phone: (435) 797-2946
Fax: (435) 797-3663
jeff.horsburgh@usu.edu

10-30-2008

Water Resources Research
American Geophysical Union
2000 Florida Avenue, N. W.
Washington, DC 20009

To Permissions Editor:

I am preparing my dissertation in the Civil and Environmental Engineering Department at Utah State University. I hope to complete my degree in December of 2008. A paper titled *A relational model for environmental and water resources data*, of which I am first author, and which appeared in your journal Water Resources Research, reports an essential part of my dissertation research. I would like permission to reprint it as a chapter in my dissertation. Reprinting the chapter may necessitate some revision. Please note that Utah State University sends dissertations to Bell & Howell Dissertation Services to be made available for reproduction.

I will include an acknowledgment to the article on the first page of the chapter, as shown below. Copyright and permission information will be included in the form of this letter in a special appendix to the dissertation. If you would like a different acknowledgement, please so indicate.

Please indicate your approval of this request by signing in the space provided and attach any other form necessary to confirm permission. If you charge a reprint fee for use of an article by the author, please indicate that as well. If you have any questions, please call me at the number above or send me an email to the above address. Thank you for your assistance.

Sincerely,

Jeffery S. Horsburgh

I hereby give permission to Jeffery S. Horsburgh to reprint the requested article in his dissertation, with the following citation and acknowledgment:

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, Water Resour. Res., 44, W05406, doi:10.1029/2007WR006392.

Reprinted from *Water Resources Research*
Copyright 2008 by the American Geophysical Union
2000 Florida Avenue, N.W., Washington, DC 20009, USA

____
Signature

____
Date

____
Fee

We are pleased to grant permission for the use of material requested for inclusion in your thesis. The following non-exclusive rights are granted to AGU authors:

- All proprietary rights other than copyright (such as patent rights).
- The right to present the material orally.
- The right to reproduce figures, tables, and extracts, appropriately cited.
- The right to make hard paper copies of all or part of the paper for classroom use.
- The right to deny subsequent commercial use of the paper.

Further reproduction or distribution is not permitted beyond that stipulated. The copyright credit line should appear on the first page of the article or book chapter. The following must also be included, "Reproduced by permission of American Geophysical Union." To ensure that credit is given to the original source(s) and that authors receive full credit through appropriate citation to their papers, we recommend that the full bibliographic reference be cited in the reference list. The standard credit line for journal articles is: "Author(s), title of work, publication title, volume number, issue number, citation number (or page number(s) prior to 2002), year. Copyright [year] American Geophysical Union."

If an article was placed in the public domain, in which case the words "Not subject to U.S. copyright" appear on the bottom of the first page or screen of the article, please substitute "published" for the word "copyright" in the credit line mentioned above.

Copyright information is provided on the inside cover of our journals. For permission for any other use, please contact the AGU Publications Office at AGU, 2000 Florida Ave., N.W., Washington, DC 20009.

CURRICULUM VITAE

Jeffery S. Horsburgh
Research Engineer
Environmental Management Research Group
Utah Water Research Laboratory
Utah State University
8200 Old Main Hill
Logan, UT 84322-8200
(435) 797-2946
jeff.horsburgh@usu.edu

Education

Ph.D. Civil and Environmental Engineering, Utah State University, Logan, UT, Expected December 2008. Dissertation: Hydrologic Information Systems: Advancing Cyberinfrastructure for Environmental Observatories. Advisor: David G. Tarboton.

M.S. Civil and Environmental Engineering, Utah State University, Logan, UT, 2001. Thesis: Statistical Analysis of Regional Geographic Information Systems (GIS) Data to Predict Water Quality in Streams. Advisor: David K. Stevens.

B.S. Environmental Engineering, Utah State University, Logan, UT, 1999.

Professional Experience

Research Engineer, Director – Environmental Management Research Group: 2001-Present, Utah Water Research Laboratory, Utah State University, Logan, UT.

Research Technician: 2000-2001, Utah Water Research Laboratory, Utah State University, Logan, UT.

Research Assistant: 1999-2000, Utah Water Research Laboratory, Utah State University, Logan, UT.

Summer Fellow: Summers 1997, 1998, Idaho National Engineering and Environmental Laboratory (INEEL), Idaho Falls, ID.

Expertise

My research and training are in the areas of surface water quality, Environmental Engineering, and environmental information systems. I focus on the development of new technology to increase the understanding of environmental processes. This includes the use of Geographic Information Systems for data analysis and dissemination, observation systems and sensor networks, data models, development of information systems that support environmental observations, and modeling techniques for surface water hydrology and water quality. I have contributed to advances in the cyberinfrastructure available for hydrologic and environmental observatories. Software developed includes the Observations Data Model (ODM) and supporting applications for loading, visualizing, summarizing, querying, and editing observations data.

Professional Society Memberships

American Geophysical Union (2008 – present).

American Water Resources Association (2007 – Present).

Research Grants

State of Utah, Division of Forestry, Fire, and State Lands, An Internet Based Great Salt Lake Information System (GSLIS), **J. S. Horsburgh**, D. G. Tarboton, $67,897, 10/2008 – 9/2010.

Utah State University Water Initiative, Assessing a Variable Effective Source Area (VESA) Modification of TOPMODEL for Predicting Soil Moisture, B. T. Neilson, **J. S. Horsburgh**, N. O. Mesner, $40,884, 2008 – 2010.

National Science Foundation, CUAHSI Hydrologic Information System, D. R. Maidment, D. G. Tarboton, I. Zaslavsky, J. Goodall, M. Piasecki, $4,500,000 to the University of Texas at Austin for the Period 1/15/2007 to 12/31/2011. $80,000 per year subcontract to Utah State University.

Utah State University Water Initiative, State of Idaho Department of Environmental Quality, Continuous Water Quality Monitoring of Mud Lake to Support Evaluation of Effects of Bear River Water Diverted into Bear Lake, **J. S. Horsburgh**, D. K. Stevens, N. O. Mesner, $51,524 ($27,524 from USU and $26,000 from IDEQ), 2006 – 2008.

United States Bureau of Reclamation, Software Development to Support the Trinity River Integrated Information Management System, D. K. Stevens, T. Hardy, **J. S. Horsburgh**, $852,559, 9/2006-12/2009.

National Science Foundation, Tools for Environmental Observatory Design and Implementation: Sensor Networks, Dynamic Bayesian Nutrient Flux Modeling, and Cyberinfrastructure Advancement, D. S. Stevens, D. G. Tarboton, **J. S. Horsburgh**, N. O. Mesner, $350,000, 11/06-10/08.

Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), Development of Data Visualization, Management, and Processing Tools for the CUAHSI HIS Observations Data Model, **J. S. Horsburgh**, D. K. Stevens, $19,484, 9/2006 – 2/2007.

Cirrus Ecological Solutions (State of Utah, Department of Environmental Quality), Technical Support for TMDL Development: Upper Bear River Watershed, **J. S. Horsburgh**, $21,188, 3/2005 – 6/2006.

National Science Foundation, Development of Informatics Infrastructure for the Hydrologic Sciences, P.I. David Maidment, University of Texas at Austin. Subaward to Utah State University, $86,325, 3/04-10/07.

USDA CSREES, Evaluation of the Effects of Conservation Practices on Water Quality within the Biophysical Setting of a Watershed: Little Bear River Watershed, D. K. Stevens, **J. S. Horsburgh**, N. O. Mesner, D. Jackson-Smith, D. L. Sorensen, $645,000, 9/2004 – 9/2008.

Bear River Commission, USEPA Targeted Watersheds Program, Dynamic Water Quality Modeling to Support Water Quality Trading in the Bear River Basin, D. K. Stevens, **J. S. Horsburgh**, N. O. Mesner, $169,216, 6/2005 – 9/2008.

Bear River Commission, USEPA Targeted Watersheds Program, Development of an Internet Based Watershed Information System and Water Quality Trading Program in the Bear River Basin, David K. Stevens, **J. S. Horsburgh**, N. O. Mesner, T. Glover, $452,587, 10/2004 – 9/2008.

University of Utah, Long Term Hydrologic Observatory Website for the Great Salt Lake Basin, **J. S. Horsburgh**, $15,767, 2004.

Utah State University Water Initiative, Development of an Internet Based Laboratory Watershed Information System for the Bear River Basin, **J. S. Horsburgh**, $18,869, 3/2004 – 6/2006.

Utah State University Water Initiative, Real Time Water Quality Monitoring in the Logan River, **J. S. Horsburgh**, $14,160, 3/2004 – 6/2006.

Psomas (State of Utah, Department of Environmental Quality), Technical Support for TMDL Development: Strawberry Reservoir Watershed, D. K. Stevens, **J. S. Horsburgh**, $14,998, 3/2004 – 8/2004.

Cirrus Ecological Solutions (State of Utah, Department of Environmental Quality), Technical Support for TMDL Development: Echo Reservoir Watershed, D. K. Stevens, **J. S. Horsburgh**, D. P. Ames, $35,505, 7/2003 – 6/2005.

Cirrus Ecological Solutions (State of Utah, Department of Environmental Quality), Technical Support for TMDL Development: Otter Creek and East Fork Sevier River Watershed, **J. S. Horsburgh**, D. P. Ames, $42,946, 12/2002 – 12/2005.

Cirrus Ecological Solutions (State of Utah, Department of Environmental Quality), Technical Support for TMDL Development: Newton Reservoir and Clarkston Creek Watershed, **J. S. Horsburgh**, D. P. Ames, $23,174, 9/2002 – 3/2004.

Whatcom County, WRIA 1 Watershed Management Project, T. Hardy, M. McKee, D. Stevens, D. G. Tarboton, M. Kemblowski, J. Kaluarachchi, D. Sorenson, $2,319,446, 2000 - 2005.

State of Idaho, Department of Environmental Quality, Technical Support for TMDL Development: South Fork Payette River, D. K. Stevens, $16,446, 6/2000 – 6/2002.

USEPA, Better Assessment Science Integrating Point and Nonpoint Sources (BASINS) Training, D. K. Stevens, $225,000, 7/1999 – 6/2002.

Idaho National Engineering and Environmental Laboratory, Development of a User Driven Decision Support System for Water Availability and Quality Management, U. Lall, D. Stevens, R. Price, D. G. Tarboton, J. Kaluarachchi, Q. Weninger, T. Glover, G. Urroz, $2,275,000, 1997 - 2000.

Refereed Publications

Horsburgh, J. S., D. G. Tarboton, M. Piasecki, D. R. Maidment, I. Zaslavsky, D. Valentine, and T. Whitenack (2008), An integrated system for publishing environmental observations data, *Environmental Modeling and Software*, (In Review).

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resources Research*, 44, W05406, doi:10.1029/2007WR006392.

Goodall, J. L., J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky, (2008), A first approach to web services for the National Water Information System, *Environmental Modelling & Software*, 23(4), 404-411, doi:10.1016/j.envsoft.2007.01.005.

Maidment, D. R., Zaslavsky, I. and J. S. Horsburgh (2006), Hydrologic data access using web services, *Southwest Hydrology*, 5(3). (Available at http://www.swhydro.arizona.edu/archive/V5_N3/feature1.pdf)

Neilson, B. T., Stevens, D. K., and J. S. Horsburgh (2005), TMDL development approaches, in *Total Maximum Daily Load: Approaches and Challenges*, Edited by Tamim Younos, PenWell Corporation, Tulsa, OK.

Other Publications, Reports, and Conference Proceedings

Tarboton, D. G., and J. S. Horsburgh (2008), Observations Data Model, Chap. 3, in CUAHSI Hydrologic Information System: Overview of Version 1.1, Consortium of Universities for the Advancement of Hydrologic Science, Inc., Washington, D. C. (Available at http://his.cuahsi.org/documents/HISOverview.pdf)

Tarboton, D. G., J. S. Horsburgh, and D. R. Maidment (2008), CUAHSI Community Observations Data Model (ODM) Version 1.1 Design Specifications, (Available at http://his.cuahsi.org/)

Horsburgh, J. S., D. G. Tarboton and D. R. Maidment (2005), A Community Data Model for Hydrologic Observations, Chapter 6, in *Hydrologic Information System Status Report, Version 1*, Edited by D. R. Maidment, p.102-135, (Available at http://www.cuahsi.org/docs/HISStatusSept15.pdf)

Neilson, B. T., J. S. Horsburgh, D. K. Stevens, M. R. Matassa, J. N. Brogdon, and A. Spackman (2004), Comparison of Complex Watershed Models' Predictive Capabilities: EPRI's Watershed Analysis Risk Management Framework (WARMF) vs. USEPA's Better Assessment Science Integrating Point and Nonpoint Sources (BASINS/WinHSPF), Utah Water Research Laboratory, Utah State University, Logan, UT.

Stevens, D. K., and J. S. Horsburgh (2003), GIS-Based Watershed Information System Including Water Quality and Streamflow Data Analysts, pp. 420-427 in Total Maximum Daily Load (TMDL) Environmental Regulations II, Conference Proceedings, 8-12 November 2003, Albuquerque, New Mexico, USA, ed. Ali Saleh., 8 November 2003. ASAE Pub #701P1503.

Stevens, D. K., J. S. Horsburgh, B. T. Neilson, and B. Lunt (2002), GIS-based Watershed Data Viewer and Water Quality Data Analyst, in Proceedings of the WEF, National TMDL Science and Policy 2002 Specialty Conference, Phoenix, AZ, November 13-16.

Baldwin, C. K. and J. S. Horsburgh (2001), A Bayesian Method for Environmental Prediction, in Proceedings of the AWRA Summer Specialty Conference on Decision Support Systems for Water Resources Management, Snowbird, UT. June 27-30.

Horsburgh, J. S. and C. K. Baldwin (2001), KNNBN - Regional Prediction of Water Quality in Data Poor Watersheds, in Proceedings of the AWRA Summer

Specialty Conference on Decision Support Systems for Water Resources Management, Snowbird, UT, June 27-30.

Horsburgh, J. S. (2000), Water Quality Estimation from Regional Characteristics, In Proceedings of the Twentieth Annual ESRI International User Conference, San Diego, CA, June 26-30.

Theses

Horsburgh, J. S., (2008), Hydrologic Information Systems: Advancing Cyberinfrastructure for Environmental Observatories, PhD Dissertation, Utah State University, Logan, UT, (In preparation).

Horsburgh, J. S., (2001), Statistical Analysis of Regional Geographic Information Systems (GIS) Data to Predict Water Quality in Streams, M.S. Thesis, Utah State University, Logan, UT, 144 pp.

Conference Presentations, Posters, and Abstracts

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), Using the Observations Data Model in Hydrologic Information Systems, Presented at the CUAHSI Biennial Colloquium on Hydrologic Science and Engineering, Boulder, CO, July 14-16.

Horsburgh, J. S., A. Spackman, D. K. Stevens, D. G. Tarboton, and N. O. Mesner (2008), Using GIS in Creating an End-to-End System for Publishing Environmental Observations Data, Presented at the AWRA Spring Specialty Conference on GIS and Water Resources V, San Mateo, CA, March 17-19.

Tarboton, D. G., J. S. Horsburgh, D. R. Maidment, and I. Zaslavsky (2008), Using an Observations Data Model in Hydrologic Information Systems, Presented at the AWRA Spring Specialty Conference on GIS and Water Resources V, San Mateo, CA, March 17-19.

Horsburgh, J. S., A. Spackman, D. K. Stevens, D. G. Tarboton, and N. O. Mesner (2008), An end-to-end system for publishing environmental observations data, Presented at the Utah State University Water Initiative Spring Runoff Conference, Logan, UT, March 31-April 1.

Tarboton, D. G., J. S. Horsburgh, D. R. Maidment, I. Zaslavsky, M. Piasecki, and J. Goodall (2007), Developing a Community Hydrologic Information System, Presented at the Environmental Sensing Symposium, Boise State University, Boise, ID, October 25-26.

Horsburgh, J. S., D. G. Tarboton, I. Zaslavsky, D. R. Maidment, and D. Valentine (2007), Deployment and Evaluation of an Observations Data Model, *Eos Trans. AGU*, 88(52), Fall Meet. Suppl., Abstract H11K-05.

Horsburgh, J. S. and D. G. Tarboton (2007), Development of a Community Hydrologic Information System, Presented at the Utah State University Water Initiative Spring Runoff Conference, Utah State University, Logan, UT, April 5-6.

Tarboton, D. G., D. K. Stevens, J. S. Horsburgh, and N. O. Mesner (2007), Building towards a Hydrologic Observatory in the Great Salt Lake Basin, Presented at the Utah State University Water Initiative Spring Runoff Conference, Utah State University, Logan, UT, April 5-6.

Mesner, N. O., J. S. Horsburgh, D. K. Stevens, D. L. Sorenson, R. Ryel, and D. Jackson-Smith (2007), Comparison of Water Quality Monitoring Techniques: Detecting Change in a Variable Environment, Presented at the Bear River Symposium, Utah State University, Logan, UT, September 5-7.

Spackman, A., D. K. Stevens, D. G. Tarboton, N. O. Mesner, and J. S. Horsburgh, (2007), Surrogate measures for providing high frequency estimates of total suspended solids and phosphorus concentrations in the Little Bear River, Presented at the Bear River Symposium, Utah State University, Logan, UT, September 5-7.

Stevens, D. K., D. G. Tarboton, J. S. Horsburgh, and N. O. Mesner (2006), Little Bear River Test-Bed: Tools for Environmental Observatory Design and Implementation, *Eos Trans. AGU*, 87(52), Fall Meet. Suppl., Abstract H21F-1437.

Tarboton, D. G., J. S. Horsburgh, I. Zaslavsky, D. R. Maidment, D. Valentine, and B. Jennings (2006), A Community Data Model for Hydrologic Observations, *Eos Trans. AGU*, 87(52), Fall Meet. Suppl., Abstract H21F-1431.

Stevens, D. K., J. S. Horsburgh, N. O. Mesner, T. Glover, A. Caplan, and A. Spackman (2006), Integrating Historical and Realtime Monitoring Data into an Internet Based Watershed Information System, Presented at the 2006 National Water Quality Monitoring Council National Monitoring Conference, San Jose, CA, May 7-11.

Stevens, D. K., J. S. Horsburgh, N. O. Mesner, D. Jackson-Smith, D. L. Sorenson, R. Ryel, (2006), A real-time water quality monitoring network for investigating the strengths and weaknesses of existing monitoring techniques, Presented at the 2006 National Water Quality Monitoring Council National Monitoring Conference, San Jose, CA, May 7-11.

Horsburgh, J. S., D. K. Stevens, and J. Goodall (2006), Time Series Analyst – An Internet Based Application for Analyzing Environmental Time Series, Presented at the AWRA Spring Specialty Conference on GIS and Water Resources IV, Houston, TX, May 8-10.

Tarboton, D. G., J. S. Horsburgh, D. R. Maidment, I. Zaslavsky, D. Valentine, and B. Jennings (2006), Testing a Community Data Model for Hydrologic Observations, Presented at the AWRA Spring Specialty Conference on GIS and Water Resources IV, Houston, TX, May 8-10.

Horsburgh, J. S., D. K. Stevens, D. L. Sorenson, N. O. Mesner, D. Jackson-Smith, and R. Ryel (2006), Real time water quality monitoring for investigating the strengths and weaknesses of existing monitoring techniques, Presented at the Water Environment Association of Utah (WEAU) Conference, St. George, UT, March 2006.

Ames, D. P. and J. S. Horsburgh (2005), A Slinky Space Streamflow Estimator for TMDLs: Time Series Analysis Meets Geostatistics, AWRA Annual Conference. Seattle, Washington, November 2005.

Horsburgh, J. S. (2005), An Internet Based Watershed Information System for the Bear River Basin, Presented at the Utah Nonpoint Source Conference, Salt Lake City, UT, September 27-29, 2005.

Horsburgh, J. S. (2004), Development of an Internet Based Laboratory Watershed Information System for the Bear River Basin, Presented at the Utah State University Water Initiative Spring Runoff Conference, March 25-26, 2004.

Ames, D. P. and J. S. Horsburgh (2003), MapWindow GIS, a Data Distribution Solution, Presented at the Utah Geographic Information Council Annual Conference, September 8-10.

Professional Activities

Instructor, CUAHSI Hydrologic Information System – Pre-conference seminar. 2008 AWRA Spring Specialty Conference, GIS and Water Resources V, San Mateo, CA, March 16, 2008.

Instructor, Better Assessment Science Incorporating Point and Nonpoint Sources (BASINS) Training, 2000 – 2002.

Reviewer for the following Journals:

Institution of Civil Engineers (ICE) - Water Management
Advances in Water Resources

Awards

American Water Resources Association, Award for Outstanding Student Presentation at AWRA's Spring Specialty Conference, GIS and Water Resources V, March 17-19, 2008.