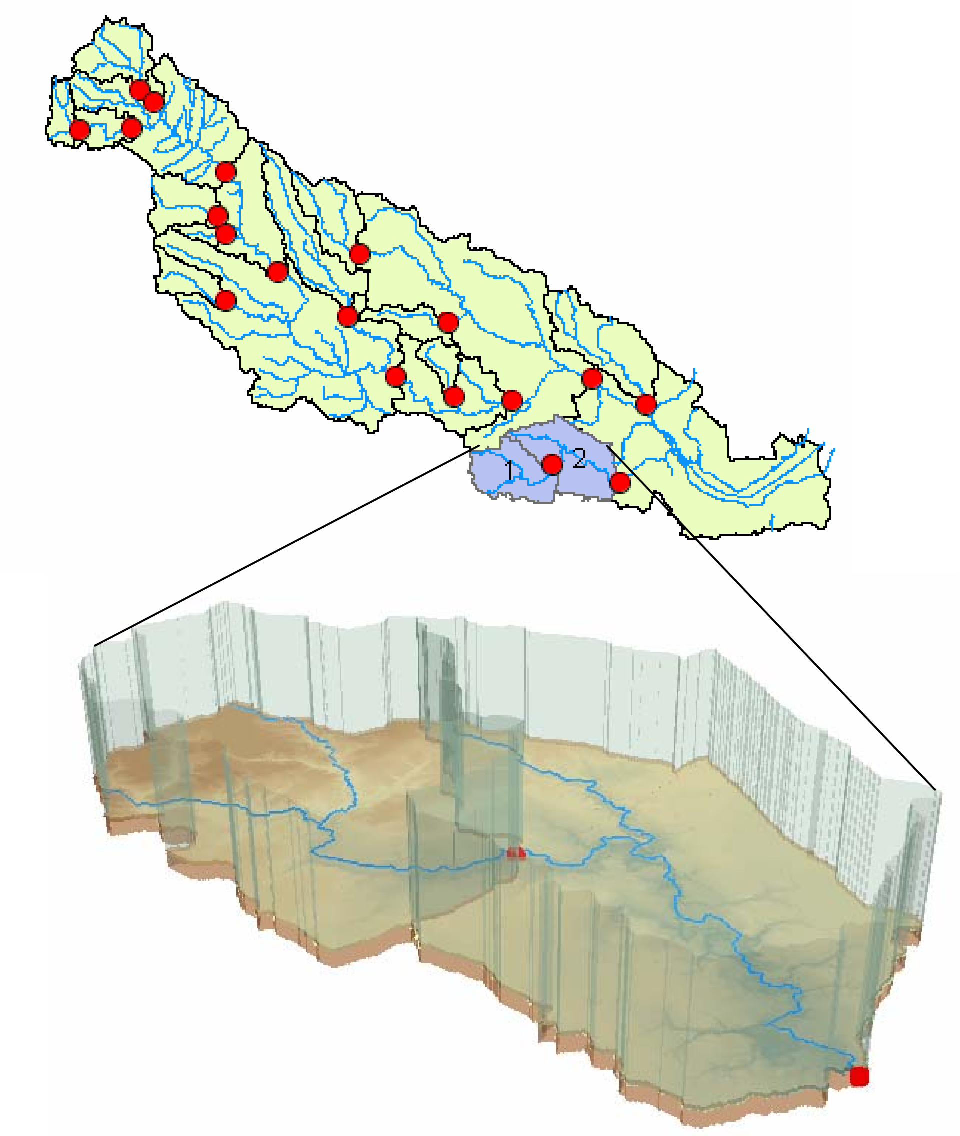
Hydrologic Information System
Status Report

CUAHSI
universities allied for water research



Version 1: September 15, 2005
Edited by David R. Maidment

Table of Contents

Chapters: Page

This material is based upon work supported by the National Science Foundation under Grant Nos. EAR 0412975 and 0413265.

Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

Developing a Hydrologic Information System

Executive Summary

A Hydrologic Information System (HIS) is a combination of hydrologic data, tools and simulation models that supports hydrologic science, education and practice. The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) is conducting an NSF-supported project to examine how such a system should be defined and to establish feasible pathways by which hydrologic information systems can be built. There are four goals: (1) to provide hydrologic scientists with better access to a large volume of high quality hydrologic data; (2) to develop a digital hydrologic observatory that presents to the viewer a seamless, comprehensive digital description of hydrologic region such as a river basin or aquifer; (3) to advance hydrologic science by enabling deeper insights into the functioning of hydrologic processes and environments; (4) to enhance hydrologic education by bringing the digital hydrologic observatory into the classroom. This report summarizes the findings to date of the CUAHSI HIS project, at a point when nearly 18 months of the 24 month project duration have occurred.

NSF has recently suggested that CUAHSI and the related program in environmental engineering, CLEANER (Collaborative Large-Scale Engineering Analysis Network for Environmental Research), examine how related functions in the two programs can be made interoperable.

An *environmental system* is a set of interrelated entities in the natural environment, such as rivers, lakes and aquifers, and the human infrastructure that interacts with those entities. Environmental systems have multiple scales, with subsystems of larger systems interacting through complex physical, chemical, and biological processes that govern the flow of water, sediment, nutrients and contaminants through these systems. Studying an environmental system requires definition of appropriate boundaries to be considered, which leads to a defined geographic region of interest and key inputs, outputs and governing processes within that region.

A *hydrologic system* is that part of an environmental system which contains the flow of the earth's natural waters, and the transport and transformation of sediment, nutrients, and contaminants carried by those waters.

CLEANER and CUAHSI are working together to better understand the processes governing these large-scale environmental systems by using advances in measurement, analysis and information technologies.

Key insights and accomplishments of the CUAHSI HIS project to date can be summarized under several headings: hydrology from a computer science perspective, the digital hydrologic observatory, HIS and web portals, and hydrologic data and functions.

Hydrology from a Computer Science Perspective

There are distinctive aspects of the hydrologic science community that make it slightly different from other data-oriented science communities:

- There is a great emphasis on “third party” data, i.e. data collected by another agency, typically a federal or state agency. Much of the data-oriented work is on acquiring that data and analyzing it, or using it in simulation models. This is true for point source time-series data from data sources such as NWIS (National Water Information System), NAWQA (National Water Quality Assessment), EPA Storet and Climate Data Online (with USGS, EPA, NCDC as the agencies) and also for remotely sensed data (with NASA, NOAA, USGS as the agencies). Thus, the HIS project is focused on providing web services to access remote data, digital libraries for storing large datasets acquired from third parties, and services-based mechanisms to easily include the different types of data in hydrologic analysis
- There seems to be sub-groups within the community, one that deals more with the point-source type of time series and another that deals with remotely sensed data series. These groups are trying to reconcile scaling issues between point-source and remotely sensed data. Thus, there is a need to understand the information technology needs and requirements of each subgroup and to serve all those needs.
- For those dealing with vector-oriented time series data, there is a widely adopted data model, Arc Hydro. This helps by providing a common basis for data structuring, at least for this subgroup of the community.
- At the science level, there are “natural” concepts, or objects, that form the basis for data and tool integration. These are concepts such as “digital watershed”, “digital aquifer”, etc. The “digital watershed” is the most important concept, and provides the conceptual basis for integrating data and tools. One can literally create a programming language construct (e.g. a complex Java object) to represent this — with a mapping from this object to real data, tools, and workflows, all of which together define a given digital watershed (or, at any rate, a particular scientist’s model of a given watershed).
- The primary focus is on the “local” or “regional” scale, e.g. a watershed, river basin or aquifer, rather than on a continental or tectonic scale. This means there is a natural network of nodes across the nation. Each node represents the data for a local region and there are natural “curators” for each region. These are the scientists who do work in the given area. Therefore, there is a need to provide information technology support to create a network of such nodes to enable data sharing. Digital Hydrologic Observatories may be developed at these nodes.

Digital Hydrologic Observatory

A *Digital Hydrologic Observatory* of a hydrologic region such as a river basin or aquifer is a comprehensive digital description using observations and simulation models of the functioning of this hydrologic system. Using data viewing tools, hydrologic scientists will be able to examine how water flows through the system and how sediment, nutrients and contaminants are transported and transformed as the flow occurs. A Digital Hydrologic Observatory has several components:

- A *Hydrologic Digital Library* indexes disparate sources of data, models and information using standardized metadata descriptions of each source, integrated using a metadata catalog, analogous to the card catalog in a traditional library. A digital library can also serve as a repository of large datasets describing the water environment of the observatory region. Technologies such as the San Diego Supercomputer Center Storage Resource Broker are being used to implement the digital library. A prototype hydrologic digital library has been constructed for the Illinois River basin.
- A *Digital Watershed* (or Digital Aquifer) is a fusion of point hydrologic observation data, GIS data, remote sensing images, and weather and climate grid information, linked to hydrologic simulation models. *Scientific workflow* tools such as ModelBuilder, D2K and Kepler are used to structure the flow of information among the data sources and models. A prototype Digital Watershed has been developed for the Neuse River basin.
- A *Hydrologic Flux Coupler* is a means of tracing hydrologic fluxes, flows and stores within and between components of a hydrologic system. The flux coupler serves as the interface between atmospheric, surface and subsurface water systems. A prototype hydrologic flux coupler has been created to link atmospheric and surface water in catchments of the Neuse River basin.

HIS and Web Portals

A *web portal* is a structured computer interface environment that integrates many kinds of information products and services from disparate sources on the internet. Web portals and services will serve several functions for CUAHSI HIS:

- Individual HIS components will reside on a linked network of computers at many geographic locations that perform as a connected system. Scientists will access HIS components through an HIS web portal that will provide them with data, tools and means of scientific collaboration. Selected HIS components may also be presented in other web portals, such as those for the CUAHSI National Center for Hydrologic Synthesis (NCHS), and the National Ecological Observatory Network (NEON). In this manner, data and functions developed within CUAHSI HIS can be made accessible to a variety of scientific communities.
- CUAHSI HIS is being developed using a *service-oriented architecture* so that CUAHSI HIS can function as a component of a collaborative, large-scale environmental observatory. This process leverages technologies and services developed by the San Diego Supercomputer Center in partnership with other cyberinfrastructure projects such as GEON (GEOinformatics Network) and SEEK (Science Environment for Ecological Knowledge).
- CUAHSI *web services* have been constructed to automatically access the USGS National Water Information System (NWIS) so that this national data archive is as accessible to the hydrologic scientist as if all the NWIS data resided on his or her

own local disk. Within the HIS web portal, there will exist a *common data window* in which the hydrologic scientist can select data of a particular type, have it searched out within NWIS and across a range of federal and other databases, and have the data served out in a consistent format regardless of the format in the original source database.

- New HIS tools, datasets and models developed and supported at CUAHSI institutions will be incorporated into the national HIS by being made accessible through web portals. A Time Series Analyst application developed at Utah State University, now made operational nationally by utilizing the CUAHSI NWIS web services, is demonstrated as a working example of this principle.
- A *cybercollaboratory* is a web portal that facilitates the activity of a community of scientists working jointly. CUAHSI HIS has adopted the cybercollaboratory technology of the CLEANER program. The CLEANER/CUAHSI cybercollaboratory will be used as an information portal to present sample prototypes of the data, tools and portal modules developed in this project for evaluation by the CUAHSI and CLEANER communities. Access to the CUAHSI common data window and to the Utah State University Time Series Analyst is already available from the CLEANER/CUAHSI cybercollaboratory.

Hydrologic Data and Functions

The manner in which hydrologic scientists use data has particular characteristics and functional requirements:

- A survey of CUAHSI hydrologic scientists shows that 96% of them use the Windows operating system and 36% also use one or more of the MacIntosh, Linux or Unix operating systems. The most widely used applications are Excel, ArcGIS and Matlab, followed by the programming languages Fortran, C/C++ and Visual Basic. The most widely used hydrologic simulation model is Modflow. Hydrologic scientists strongly desire better access to streamflow, water quality, remote sensing, precipitation, and groundwater data.
- Hydrologic observational data measured at points, such as gages and sampling sites, need a specially designed observations database in which the data are linked to metadata which describe their origin and character. A prototype *hydrologic observations database* design has been prepared. Case studies and benchmarking of implementation of this database are being conducted using high performance database technologies, such as IBM's parallel DB2 database, in order to support large-scale data sets.
- Hydrologic metadata use a hierarchy of concepts, called an *ontology*, to describe hydrologic data. A standardized CUAHSI *metadata profile* has been prepared and compared to the hydrologic metadata profiles used by various federal agencies. A minimal set of six basic metadata elements has been identified to facilitate quick description of data. Ontology-based services for dataset registration, search, and data integration developed in GEON are now being integrated into the CUAHSI HIS portal.

- A *CUAHSILink* ArcGIS extension has been created which provides access to CUAHSI digital library services from ArcGIS desktop, which has been identified as one of the most widely used applications by the hydrology community. This extension allows users to search digital library holdings and retrieve spatial data directly into ArcGIS.
- Tracing the movement of water and its constituents as a continuum through components of a hydrologic system such as watersheds, stream channels, and aquifers requires a *hydrologic data model* integrating space, time, and an array of hydrologic variables. The data will physically reside in a structured connection between a relational database and a set of binary data files. A *geotemporal reference frame* will define a common space and time coordinate system that the data share.
- To process large grid datasets from numerical simulations and remote sensors, and to meaningfully relate that data to other objects in a GIS framework, a *Modelshed* geodata model has been developed for diverse environmental science and hydrologic applications. It is capable of representing four-dimensional (space-time) model domains, vertical layering, environmental fluxes, dynamic spatial features, statistical time series data, and relationships among heterogeneous model domains.
- *Data driven discovery* is a new discipline that provides tools for accessing and handling a variety of very large data sets to illuminate patterns of relationships in information using data mining and space-time exploratory data analysis techniques. A prototype system of this kind has been developed using the D2K/I2K scientific workflow method and applied to analyze changes in remotely sensed characteristics of a hydrologic landscape.

During the final six months of this CUAHSI HIS project, attention will be focused on community engagement, feedback, and guidance for preparation of further plans for HIS development. A key goal will be refinement of the design for the Digital Hydrologic Observatory. A subsequent version of this status report will be accompanied by sample prototypes so that the community can directly test and examine the databases, tools, and portals referred to in this report.

Preface

The information contained in this report is being compiled during a research project sponsored by the National Science Foundation to investigate how a Hydrologic Information System (HIS) can be designed to meet the needs of faculty, students, and researchers, in US universities.   The project is being undertaken within the organizational structure of the Consortium of Universities for the Advancement of Hydrologic Science, Inc, (CUAHSI), an NSF-sponsored consortium of which more than 100 universities are members, which seeks to improve the infrastructure and services for the advancement of hydrologic science and education in the United States.   The CUAHSI HIS project started in April 2004 and will terminate in March 2006.

The intent of preparing and revising a status report on this CUAHSI HIS project during the last six months of the project's life is to provide a means for informing NSF, project partners, and the CUAHSI community of what has been learned in the HIS project, and of soliciting feedback and refinement of the concepts presented. In a later version of this report, links will be provided to sample prototypes of the tools, databases and web portals described in this report.

CUAHSI is presently engaged in a process of making its cyberinfrastructure development program interoperable with that of the related NSF program in environmental engineering called CLEANER (Collaborative Large-Scale Engineering Analysis Network for Environmental Research).   The CLEANER program is having its first meeting in Washington DC on September 20-22, 2005, following the establishment of the CLEANER program office.   The purpose of this draft of the HIS status report is to inform our colleagues in CLEANER as to the nature of our work so that we can explore with them how best we can proceed together towards the development of an interoperable program for cyberinfrastructure development for CUAHSI and CLEANER.

Many faculty, researchers and graduate students have contributed to the collective insights that are described in this report. In particular, I would like to acknowledge our colleagues at the San Diego Supercomputer Center: Chaitan Baru, Ilya Zaslavsky, Reza Wahadj, John Helly, Don Sutton and Tiffany Houghton; from the University of Illinois at Urbana-Champaign: Praveen Kumar, Ben Ruddell, Pratyush Sinha, Vikas Mehra, Barbara Minsker and Luigi Marini; from Drexel University: Michael Piasecki, Luis Bermudez, Bora Boran, Saiful Islam and Yoo-Ri Choi; from Duke University: Ken Reckhow, Jon Goodall and Peter Harrell; from the University of North Carolina: Larry Band and David Tenenbaum; from the University of South Carolina: Venkat Lakshmi and Ujjwal Narayan; from Utah State University: David Tarboton, Jeff Horsburgh and Christina Bandaragoda; from the University of California at Berkeley: Xu Liang and Seongeun Jeong; from the Lawrence Berkeley Laboratory: Norman Miller, Susan Hubbard and Deborah Agarwal; from Unidata: Ben Domenico, Russ Rew, Jeff Weber and Mohan Ramamurthy; Yao Liang from Virginia Tech, Chunmaio Zheng from the University of Alabama, Leroy Poff from Colorado State University, Upmanu Lall from Columbia University, Wendy Graham from the University of Florida, Anton Kruger from the University of Iowa, Dennis Lettenmaier from the University of Washington, Bill

Michener from the University of New Mexico, Kelly Redmond from the University of Nevada, Paul Morin from the University of Minnesota, Norman Jones from Brigham Young University, Randy Keller from the University of Texas at El Paso; Venkatesh Merwade, Gil Strassberg and Tim Whiteaker from the University of Texas at Austin; and last, but not least, Richard Hooper and Jon Duncan, our long serving guides and inspirations from the CUAHSI Program Office in Washington, and Douglas James our project manager at the National Science Foundation. Clearly, the collective wisdom expressed in this report is the product of many minds and efforts, and I wish to thank everyone who is participating in this project for their insights and assistance.

Please feel free to contact me directly if you have comments or suggestions concerning the content of this report. I served as Chairman of the CUAHSI Hydrologic Information System Committee from January 2002 to April 2004, and have served from April 2004 as the Principal Investigator of the CUAHSI HIS project, along with Chaitan Baru, Praveen Kumar, Michael Piasecki and Richard Hooper who are the co-Principal Investigators of this project. We welcome your comments and suggestions.

David R. Maidment
University of Texas at Austin
maidment@mail.utexas.edu

1. Introduction

By David R. Maidment
Center for Research in Water Resources
University of Texas at Austin

The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) is an organization representing more than a hundred US universities, sponsored by the National Science Foundation to develop infrastructure and services for the advancement of hydrologic science and education in the United States (http://www.cuahsi.org). The CUAHSI Hydrologic Information System (HIS) project is a component of CUAHSI's mission that is intended to improve infrastructure and services for hydrologic data acquisition and analysis. The CUAHSI Hydrologic Information System project is supported by the National Science Foundation and was initiated in April 2004 for a period of two years to investigate how best to construct a hydrologic information system and to define the pathway forward for building such a system in the years ahead. The purpose of this report is to present the conclusions that the HIS team has reached so far, and to invite feedback from the CUAHSI and CLEANER communities as to the future directions they see as having greatest merit.

The CUAHSI Hydrologic Information System has four goals:

(1) **Data Access:** to provide hydrologic scientists with better access to a large volume of high quality hydrologic data;
(2) **Digital Hydrologic Observatory:** to develop a digital hydrologic observatory that presents to the viewer a seamless, comprehensive digital description of hydrologic region such as a river basin or aquifer;
(3) **Hydrologic Science:** to advance hydrologic science by enabling deeper insights into the functioning of hydrologic processes and environments;
(4) **Hydrologic Education:** to enhance hydrologic education by bringing the digital hydrologic observatory into the classroom.

The HIS project has been undertaken by a network of investigators from CUAHSI institutions collaborating with researchers from the San Diego Supercomputer Center as our technology partner. The collaboration among the CUAHSI investigators actually began during the Spring of 2002 as a CUAHSI Hydrologic Information Systems committee which planned the HIS effort in parallel with other CUAHSI committees planning hydrologic observatories, a hydrologic synthesis center, and a hydrologic measurement program, with the hydrologic observatories being considered the centerpiece of this effort. Specific proposals have been made by CUAHSI to NSF for development of the synthesis center and the hydrologic measurement program, and initiation of those activities is expected to occur shortly.

In parallel with these developments, the National Science Foundation has been reorganizing the manner in which it supports computational infrastructure in science and engineering to focus on *cyberinfrastructure*, which includes the combination of high-speed telecommunications, distributed computing services, and advances in visualization, data analysis, and remote collaboration and knowledge sharing capabilities to produce an infrastructure which supports scientific and engineering research and education broadly across the nation. As described by the NSF Advisory Panel for Cyberinfrastructure (2003,): “The emerging vision is to use cyberinfrastructure to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments, and that operate at unprecedented levels of computational, storage, and data transfer capacity.” The NSF Advisory Committee for Environmental Research and Education (2003) describes an outlook for the first decade of the 21st century that focuses in part on advancing interdisciplinary study of the environment through better measurement and information management: “These new instrumentation, data-handling, and methodological capabilities have expanded the horizons of what we can study and understand about the terrestrial, freshwater, marine, and sedimentary environments, the atmosphere, and near-Earth environments in space”. The emerging field of *environmental cyberinfrastructure* is the component of this overall effort which will support environmental observatories.

CUAHSI and CLEANER[1]

The CUAHSI hydrologic observatories program has undergone a significant transformation during 2005 from the form in which it was earlier envisaged. The original concept was that NSF would select a set of hydrologic regions around the nation and make a significant investment in hydrologic investigation infrastructure at those locations. At a CUAHSI meeting held in Logan, Utah in August 2004, plans for 24 proposed observatory regions were presented by teams of CUAHSI institutions. The National Science Foundation is also contemplating parallel investments in other environmental observatory programs, for environmental engineering (CLEANER – Collaborative Large-Scale Engineering Analysis Network for Environmental Research http://cleaner.ncsa.uiuc.edu/), for ecology (NEON – National Ecological Observatory Network http://www.neoninc.org/) and for the ocean sciences (ORION – Ocean Research Interactive Observatory Networks http://www.orionprogram.org/). The combination of the financial requirements of all these observing programs and the budgetary constraints that have recently been placed on NSF, has caused the NSF leadership to ask for a more rationalized plan for environmental observatory development.

The immediate consequence of these changes is that their respective NSF directorates (Geosciences for CUAHSI and Engineering for CLEANER) have suggested that CUAHSI and CLEANER define a plan for joint use of the information and measurement systems and observatory infrastructure that they contemplate creating. The confluence of the aspirations of CUAHSI and CLEANER has been assisted by the fact that the CLEANER program has recently designated an organizational structure for its program development, led by investigators from the University of Illinois at Urbana-Champaign

[1] The material in this section has been prepared in collaboration with Barbara Minsker, Principal Investigator for the CLEANER Program Office and Co-Chairperson of the CLEANER Cyberinfrastructure Committee

and other institutions (see http://cleaner.ncsa.uiuc.edu). The joint planning of information systems and cyberinfrastructure development for the CUAHSI and CLEANER programs began during the summer of 2005 and fortunately, it appears that the proposed approaches for CUAHSI and CLEANER are quite compatible. In particular, the NSF cyberinfrastructure program has two main centers – the San Diego Supercomputer Center, and the National Center for Supercomputer Applications (NCSA) at the University of Illinois. With the CUAHSI effort collaborating with the San Diego center and the CLEANER effort focused at Illinois, the combined CUAHSI/CLEANER effort has the opportunity to draw on the strengths of both of these key cyberinfrastructure institutions. The CLEANER project office plans to define its cyberinfrastructure user requirements by December 2006. To facilitate the requirements gathering process, the project office plans to leverage environmental cyberinfrastructure demonstrations underway at NCSA, the HIS project demonstrations, and other available environmental cyberinfrastructure demonstrations to illustrate potential capabilities of cyberinfrastructure for supporting environmental researchers and educators.

Environmental and Hydrologic Systems

In order for the scientific missions of CUAHSI and CLEANER to be understood in an integrated way, it is useful to develop abstract definitions of the fundamental concepts involved in each program. An *environmental system* is a set of interrelated entities in the natural environment, such as rivers, lakes and aquifers, and the human infrastructure that interacts with those entities. Environmental systems have multiple scales, with subsystems of larger systems interacting through complex physical, chemical, and biological processes that govern the flow of water, sediment, nutrients and contaminants through these systems. Studying an environmental system requires definition of appropriate boundaries to be considered, which leads to a defined geographic region of interest and key inputs, outputs and governing processes within that region.

A *hydrologic system* is that part of an environmental system which contains the flow of the earth’s natural waters, and the transport and transformation of sediment, nutrients, and contaminants carried by those waters.

CLEANER and CUAHSI are working together to better understand the processes governing these large-scale environmental systems by using advances in measurement, analysis and information technologies.

The present draft of this report reflects the investigation effort and perspective of the CUAHSI program. Subsequent drafts of this report before its final publication in March 2006 will show more fully how a combined information infrastructure can be developed for CUAHSI and CLEANER together.

Cybercollaboratory

A particular theme of the CLEANER Cyberinfrastructure project is the idea of a *cybercollaboratory*, which is a web portal combining tools for facilitating social networking among groups of researchers working on common themes, including data access and analysis. Interaction among the investigators on the HIS project, which has been accomplished mainly by email and conference calls, has resulted in a kind of collaboratory where the investigators have worked towards a common purpose, and conducted a continuous dialog to facilitate their collective understanding of the subjects that they are addressing. Email and conference calls work effectively for a group of a dozen or so investigators, but this method of interaction will not scale up readily to interaction with the whole CUAHSI community. The cybercollaboratory will provide a more scaleable means of involving a larger array of input to the HIS project from the whole CUAHSI community to be achieved. CUAHSI HIS has adopted the CLEANER cybercollaboratory technology so that there will be a single source for both the CUAHSI and CLEANER communities to view the results of the HIS effort (see Figure 1).

This report refers to many kinds of software tools, databases, and web portals and services. It is useful for the reader to be able to try out these tools, investigate the databases, to work with the portal components themselves and not to rely simply on what is described about them in the report. The CUAHSI HIS team is preparing a set of "sample prototype" systems that will be connected to a later revision of this report and will be available for review and discussion, probably through the cybercollaboratory. This will facilitate feedback and guidance as to appropriate user requirements definitions for the future CUAHSI and CLEANER cyberinfrastructure programs.

CLEANER
Collaborative Large-scale Engineering Analysis Network for Environmental Research
CUAHSI
The Consortium of Universities for Advancement of Hydrologic Science, Incorporated
Home | Collaboration | Data | Analysis | Library | Search

Figure 1. The CLEANER/CUAHSI cybercollaboratory (http://cleaner.ncsa.uiuc.edu)

Digital Observatory

Central to the CUAHSI/CLEANER vision is the idea of a *digital observatory*, which is a comprehensive characterization of an environmental system using integrated data and simulation models. For example, if the system contains river basins, aquifers, and bay and estuary systems, the digital observatory has components of a *digital watershed* to describe the river basins, a *digital aquifer* to describe the groundwater resources, and a *digital estuary* to describe the bay and estuary systems. The digital observatory may cover many thousands or even tens of thousands of square kilometers in area, sufficient to study the large-scale, multimedia dynamics of the system. Its digital description covers both the natural environment and also constructed infrastructure, such as dams, water diversion, treatment, piping and discharge systems. It contains means for tracking the movement of water, sediments, contaminants and nutrients through the environment. New information on environmental functioning is continuously acquired through *sensor networks*.

If a digital observatory is to be comprehensive, it must embrace the best available information produced from all sources about the environmental system, which may include data and simulation models produced by federal, state and local agencies, water authorities and districts, cities, counties, and in some cases by the consultants who work with these organizations. For example, the Federal Emergency Management Agency (FEMA) is presently reconstructing in a digital form the flood plain maps of the entire nation, county by county, with flood plain maps being created for streams whose drainage area is one square mile or greater. The projected budget of this effort is $1.5 billion over a five year period beginning in 2005. Hundreds of engineering contractors, cities, counties and other agencies are presently involved in this effort, and river morphology terrain data and hydraulic modeling being created for this program are being stored in a single large centralized database being maintained by FEMA. The hydrologic data and models arising from a national investment of this magnitude (which is many times what NSF will be able to afford to spend on hydrology) should be incorporated into digital observatories in order to make their representation of the hydrologic environment as complete as possible. The task of building a digital observatory is extensive and complex, and requires a significant effort to create appropriate environmental cyberinfrastructure tools and data structure designs.

Report Outline

The scope of this report is very wide, from an abstract conceptualization of how various information spaces can be contained within a hydrologic information model to the arcane details of how observational data should be stored in relational database tables, from consideration of the representation of hydrologic processes in atmospheric, surface and subsurface water to mechanisms for rapidly acquiring water resources information through automated web data services using the Simple Object Access Protocol (SOAP). The authors of this report do not claim to be clairvoyant and to understand everything about this complex and extensive subject, whose many aspects are themselves evolving continually. What is presented here is simply the best assessment that we can make of the facts as they are apparent now. No doubt, further investigation and a wider assessment and feedback from the CUAHSI and CLEANER communities, will lead to improvement in the insights and directions suggested here. This is a vision of how a hydrologic information system could function and feasible directions by which its components could be constructed.

This report is divided into eleven chapters. After this introduction, Chapter 2 lays out the conceptual framework by which the CUAHSI HIS will be constructed, Chapter 3 describes the proposed system architecture, Chapter 4 assesses the user needs for such a system, Chapter 5 outlines a recommended approach to hydrologic metadata development, Chapter 6 defines a relational database model for storing hydrologic observations, Chapter 7 discusses the acquisition and processing of remotely sensed data, Chapter 8 describes how a digital watershed has been constructed for the Neuse basin, Chapter 9 shows how hydrologic fluxes, flows and storage can be characterized for the Neuse basin considered as a hydrologic system, Chapter 10 shows some examples of how the HIS can be used to support data driven discovery in hydrology, and Chapter 11 outlines how HIS and CUAHSI National Center for Hydrologic Synthesis are connected.

References

NSF Advisory Panel on Cyberinfrastructure, 2003, “Revolutionizing science and engineering through cyberinfrastructure”, Report of a Blue Ribbon Advisory Panel, February 3, 2003, National Science Foundation, Arlington VA 22003, p. ES-2.

NSF Advisory Committee on Environmental Research and Education, 2003, “Complex environmental systems: synthesis for earth, life and society in the 21st century”, National Science Foundation, http://www.nsf.gov/ere , p.5.

Chapter 2

Conceptual Framework

By David R. Maidment
Center for Research in Water Resources
University of Texas at Austin

Richard P. Hooper
Consortium of Universities for the Advancement of Hydrologic Science, Inc.
Washington, DC

Introduction

The conceptual framework of the CUAHSI Hydrologic Information System rests upon four kinds of information "spaces" as shown in Figure 1: the environment space, measurement space, concept space, and simulation space. Each of these spaces contains a particular kind of information representation appropriate for its contents, and the character of these information representations may differ substantially from space to space. The CUAHSI Hydrologic Information Model combines the information representations of these four spaces with mechanisms for information transformations among them so that they function as an integrated system. Although there could be a single measurement space (there is some total number of measurements), it is more likely that people will create different measurement spaces by selecting different subsets of the available measurements for study. CUAHSI explicitly wants to enable multiple conceptual and simulation model spaces. The Hydrologic Information System will help make their construction easier and help scientists communicate what they are doing and why they are doing it.

Environment Space

The environment space describes the natural environment within which water flows and its phenomena occur. It is the "real world," or at least our perception of it through our senses (e.g., the visible spectrum) and the measurement tools that extend our senses (e.g., infrared images and concentrations of chemical elements in various media). Some aspects of the environment is sensed with quantitative measurements that include the land surface terrain, soils, vegetation, land cover, watersheds, stream networks, aquifers, hydrogeological units, and water infrastructure such as dams, bridges, and conveyance systems. This type of information is well described by geographic information systems that contain representations for continuous surfaces and discrete space objects (points, lines, areas and volumes) whose location can be considered fixed in space at any given point in time. Modern geographic information systems, such as ArcGIS, are implemented on top of relational database systems, such as Microsoft Access, SQL/Server, Oracle or DB2.

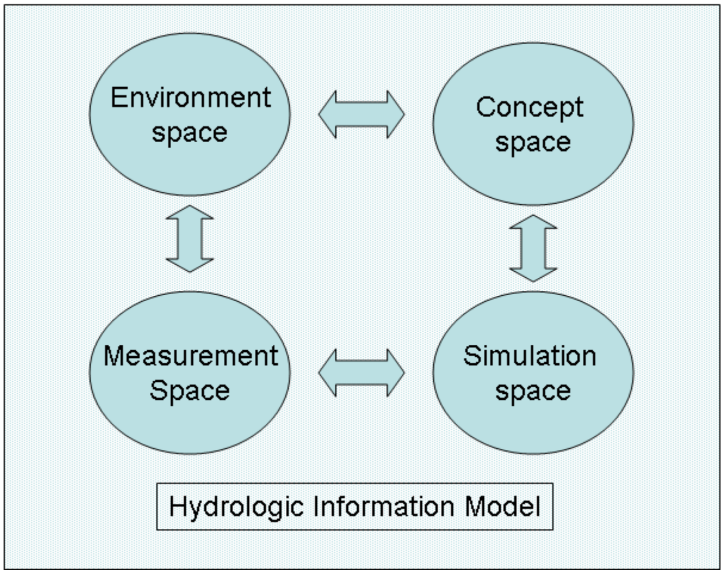


Figure 1. The hydrologic information model is an interoperable set of information representations of the natural environment, water measurements, simulation models and hydrologic concepts.

Measurement Space

The measurement space contains the sets of measurements of hydrologic phenomena made by observing systems and networks. These include *point observations* made using gages or water sampling at collections of fixed geographic locations such as networks of streamflow gages, water quality sampling sites, groundwater wells and climate stations. They also include *distributed observations* where sensors are moved through space such as measurements of water properties in a lake using a moving boat, or seismic measurements in geophysics where waves produced by impacts at some locations are sensed by receivers at other locations. Another category of measurements is *remote sensing* from satellites, aircraft or ground-based units where images of phenomena are produced by sensors mounted at a distance from the phenomena they record.

Point observations are spatially simple, requiring just a latitude and longitude pair and elevation to specify the earth location of the measurement site, but the information recorded there may be very complex, especially if real-time streaming of data is occurring. However, this is the most traditional area of hydrologic observations, carried on for more than a century, and practice in this field has converged on the use of relational databases for storing tables of processed data from observation sites. This is a function of hydrologic services such as the USGS in the United States and corresponding agencies in other countries.

Distributed observations are spatially a little more complicated, requiring sets of fixed measurement locations or tracking of a moving location and the character of what is measured may be both tabular data and images, such as geophysical images of subsurface strata. Remote sensing observations are spatial images, snapshots in time, but extensive in space, with a gridded set of image values often in the range 0-255 in each image band.

There is no single information system available that can conveniently store and process all kinds of measurement information – what is needed is a combination of a relational database for tabular information and a structured file system for image and real-time streaming information. Several forms of relational databases are available – Microsoft Access and SQL/Server, Oracle and DB2 in the commercial domain, and MySQL and PostGresSQL in the public domain. The National Center for SuperComputer Applications at the University of Illinois has developed a structured file format called HDF or Hierarchical Data Format, and NASA has adopted a variant of this format called HDF-EOS as a standard for storing remotely sensed images.

The measurement space is also closely associated with the idea of metadata. *Metadata* is generally understood to be information that describes data. Thus, the Federal Geographic Data Committee has defined *geospatial metadata* standards for spatial coverages of the natural environment used in geographic information systems. Hydrologic observation data are supplied along with *hydrologic observation metadata* that specify what instrument or procedure was used make the measurements and process them. Remote sensing data have associated *remote sensing metadata* files that describe the character and structure of the information contained in them. Climate models have *model output metadata* files that specify what variables were computed and the structure in space and time of values of those variables in the data file. Thus, if the hydrologic information model is to be able to support interactions among its four spaces, there must be some process by which the metadata describing the components of those spaces can be connected.

These connections can be accomplished by a process called semantic mapping in which sets of structured concepts, or *ontologies*, are connected by saying this concept in this ontology A is equivalent to that concept in ontology B. For example, both the USGS and EPA provide access to stored archives of water quality data. But they describe these data in quite different ways so that although the underlying character of the scientific measurement may be identical, the USGS metadata description and the EPA metadata description are expressed in different languages of coded values and text fields. *Ontology web language* is a mechanism by which each of the USGS and EPA data descriptions can be reduced to a comparable form, and then equivalences established between them. *Semantics* is the process of assigning meaning to things, and *semantic mapping,* the process of connecting things that have related meanings. Semantic mapping is a complex process that is as yet not well understood but some significant progress had been made by the GEON project at the San Diego SuperComputer Center which maintains a library of ontologies of geologic concepts which is used to interrelate the many existing descriptions of the structure and properties of rocks.

Concept Space

The concept space is the abstraction (and usually simplification) of measurement space to focus on the hypotheses being tested. Typically, a complex landscape may be reduced to a small number of units or compartments that are considered homogeneous in some property, such as permeability, soil horizon, or slope. For some spatially distributed models, such as SHE (Systeme Hydrologique European), there may be a one-to-one mapping of the environment space to the concept space as each cubic meter of soil becomes an element in the model. Although this abstraction is generally subjective, the advent of GIS enables objectively mapped areas to be determined based on rules, such as topographic indices based upon digital elevation models.

In the past, the concept space has not always been described explicitly, but is implicit in the application of a certain model to a watershed. However, it is important to distinguish the simplification of the landscape into homogeneous units (the concept space) from the specific set of algorithms used to describe hydrologic phenomena in these units (the simulation space). The concept space and the simulation space (described next) form two distinct sets of hypotheses that must be independently evaluated. An important contribution of a hydrologic information system is to permit multiple conceptualizations of an environment space to be constructed and to be evaluated..

Simulation Space

The action of constructing and operating hydrologic *simulation model*s involves describing hydrologic phenomena with sets of mathematical equations operating on the conceptual units described in the concept space, transforming the equations into computer models, and calculating results through time and across space. A plethora of well-known simulation models exist, some so widely accepted such as Modflow (a model of groundwater flow) as to be considered standards in the field, and others invented as part of their research by individual investigators to explain phenomena of interest to them.

Simulation modeling is so widely practiced in water resources that the phrase “the model” is taken to be synonymous with “simulation model”. It should be understood here, however, that simulation is just one component of a hydrologic information model and that other components, such as point observation data and remote sensing, each have their own information representations, or *data models*. Thus, the interaction between simulation models and data models is critical to the successful functioning of a hydrologic information system.

Hydrologic simulation models are computer codes that have associated input and output files. The structure of these codes and files is particular to each simulation model and is difficult to generalize. A structured file system is needed to archive and document this information. A new version of HDF called HDF5 has more flexibility than its predecessors, and may be suitable for storing simulation models and their input and output files. The Environmental Simulation and Modeling Laboratory at Brigham Young University has designed a special version of HDF5 called XMDF for storing numerical simulation model meshes and their output results. The output of climate models is stored in an array structure of which the most widely used variant is netCDF, supported by Unidata, part of UCAR in Boulder, Colorado. NetCDF stores the sampled values of an n-dimensional function space in a binary array structure. NetCDF is also used to store the results of hydrodynamic simulations of water bodies, and spatially distributed observations of properties of the ocean, such as sea surface temperature.

Digital hydrologic observatory

A *digital hydrologic observatory* is a comprehensive information depiction of a river basin that describes its natural environment, its hydrologic measurements, simulation models of its processes and phenomena, and conceptual frameworks for thinking about its hydrologic functioning. A digital hydrologic observatory is produced by the application of the Hydrologic Information Model to a hydrologic region defined by river basin or aquifer boundaries. *Cyberinfrastructure* is the combination of computer tools, telecommunications, database structures, and distributed computer networks which collectively support advancements in science and engineering through integrated information access and processing. Cyberinfrastructure has many modes of application, each of which has to be focused on the needs of the particular science or engineering community, in this case, hydrology.

The cyberinfrastructure for a digital hydrologic observatory is shown in Figure 2. Users of the digital hydrologic observatory enter through a *digital hydrologic observatory portal*, which is an internet-based computer interface which provides a local user with access to information resources scattered across a distributed domain of many remote computers, data sources, formats and software tools. Dotnetnuke is a portal technology built on Microsoft's dotnet technology which has a large body of open source implementation additions, and which may be suitable for application in the CUAHSI Hydrologic Information System. LiveRay (used in the CLEANER cybercollaboratory) and Sakai are other candidate portal technologies. Users will go to the portal input, display, query and output information from the Digital hydrologic observatory. The portal will be linked to a metadata catalog so that users can appropriately interpret the various kinds of information they are dealing with.

Underneath the portal will reside the *digital hydrologic observatory information repository* which will be a combination of a relational database containing point observation data and GIS data, and a structured file system storing remote sensing, netCDF and hydrologic simulation model files. The suggestions in Figure 2 for using MS SQL/Server and HDF5 for the relational database and structured file components of the repository are not intended to be definitive – other relational databases could be substituted for MS SQL/Server, for example. The detailed structure of the digital hydrologic observatory cyberinfrastructure is described separately in chapter 3 of this report. It is intrinsic in concept of the distributed nature of cyberinfrastructure that not all components of a digital hydrologic observatory will physically reside at the same location – in other words a digital hydrologic observatory is an assembly of components that may reside on several different computers and geographic locations. By interacting with this infrastructure through the portal, the user is screened off from the complications that result from distributed computer networks and just interacts with an integrated hydrologic information system.

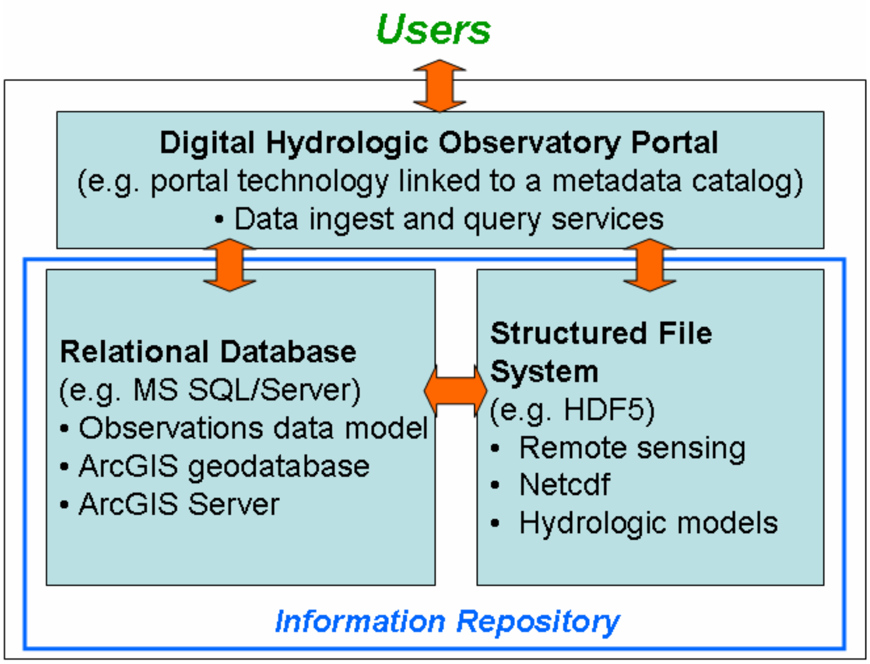


Figure 2. Cyberinfrastructure for a digital hydrologic observatory

Information Portal

Central to the functioning of the digital hydrologic observatory is the information portal which provides access to it. At an NSF Sensors for Environmental Observatory workshop held at UCLA in June 2005, representatives of CUAHSI, CLEANER and NEON jointly sketched out a design for an information portal as shown in Figure 3. At the head of this diagram is a *cyberdashboard* that comprises a set of windows through which the user interacts with the information system. These windows contain maps, graphics, sound feeds (such as the sounds of song birds), web access, wave forms (such as frequency domain representations of time series information), the topology or layout and current status of the network of observational sensors, and tools for interactive collaboration, such as chat rooms, electronic white boards and the like. There are many potential windows (or *modules* or *portlets*) of an information portal and at any given moment a user may be viewing only a few of them, rather like the set of windows to various software programs that a user maintains open on a desktop computer system when normally operating on it.

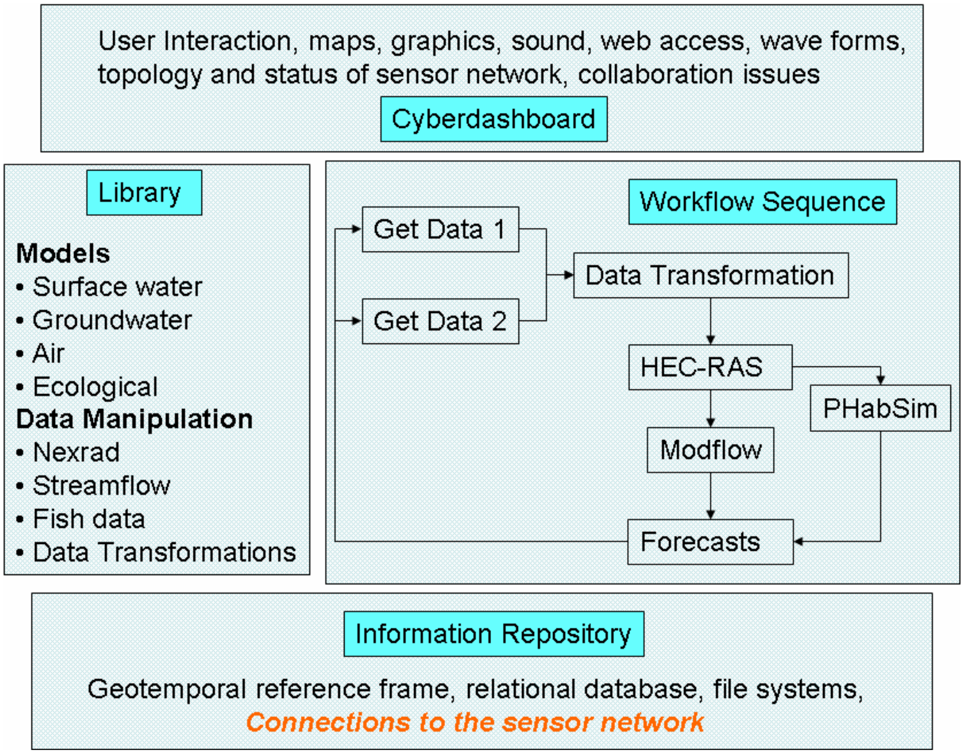


Figure 3. Components of the information portal

Beneath the cyberdashboard there will exist a library of simulation models and tools for data acquisition and manipulation. These might include simulations of surface water (HEC-RAS) and groundwater (Modflow), air, or ecological phenomena (PHabSim). Data acquisition tools will provide access (Get Data tools) to Nexrad, streamflow, fish and other kinds of data, and mechanisms for data transformations so that the data can be viewed in a consistent format and framework.

Implied by this is a *geotemporal reference frame* which means a single spatial and time coordinate system which the user chooses to operate within. For example, Nexrad data are a National Weather Service product and because NWS operates in all time zones of the United States the default time coordinate system for weather information is Universal Time Coordinates (UTC) or Greenwich Mean Time (GMT). If one wishes to plot USGS streamflow data, measured in Eastern Standard Time, and compare it to the corresponding Nexrad data, the time offset between universal time and local time has to be accounted for to allow the data to be displayed correctly and consistently. Likewise, data come in many spatial coordinate systems and GIS transformations are needed to achieve comparable data in a single spatial coordinate system.

Workflow Sequence

The accomplishment of all these data ingestion, conversion and transformation tasks is a tedious process involving a large number of small tasks of various kinds that often take a great deal of time to carry out. A *workflow sequence* is a defined set of operations that can be executed in order with data passing automatically from one operation in the sequence to the next. Several workflow sequencing environments are available. The astronomy community has created Kepler, which is supported by the San Diego SuperComputer Center; the NCSA has developed I2K (Information to Knowledge) and D2K (Data to Knowledge), and the ArcGIS system has ModelBuilder. Other software systems such as Erdas for image processing and SAS for statistical manipulation have their own workflow environments. Regardless of the type of workflow sequence, the principles are the same – take a network of operations and execute them in a defined way that may include branching and looping as in normal programming languages.

These operations may be preprogrammed tools built as standard for the workflow environment, they may be custom tools created by the investigator within the environment, and custom tools may execute simulation models. Thus, a simulation model can be thought of simply as a tool in an information system, which takes in information from other tools, and produces information which goes on to other tools. For example, the Center for Research in Water Resources of the University of Texas at Austin has produced Map2Map which is a ModelBuilder workflow sequence that takes Nexrad radar rainfall as input, executes HEC-HMS to convert rainfall into stream discharge, executes HEC-RAS to convert discharge to water surface elevation, and uses ArcGIS tools to produce the resulting inundation map, as shown in Figure 4. This whole process operates automatically and executes in minutes what would otherwise take hours or even days of work to accomplish. In Figure 4, the oval shapes represent data inputs and outputs, the square boxes represent operators, and the arrows between them show the workflow sequence.

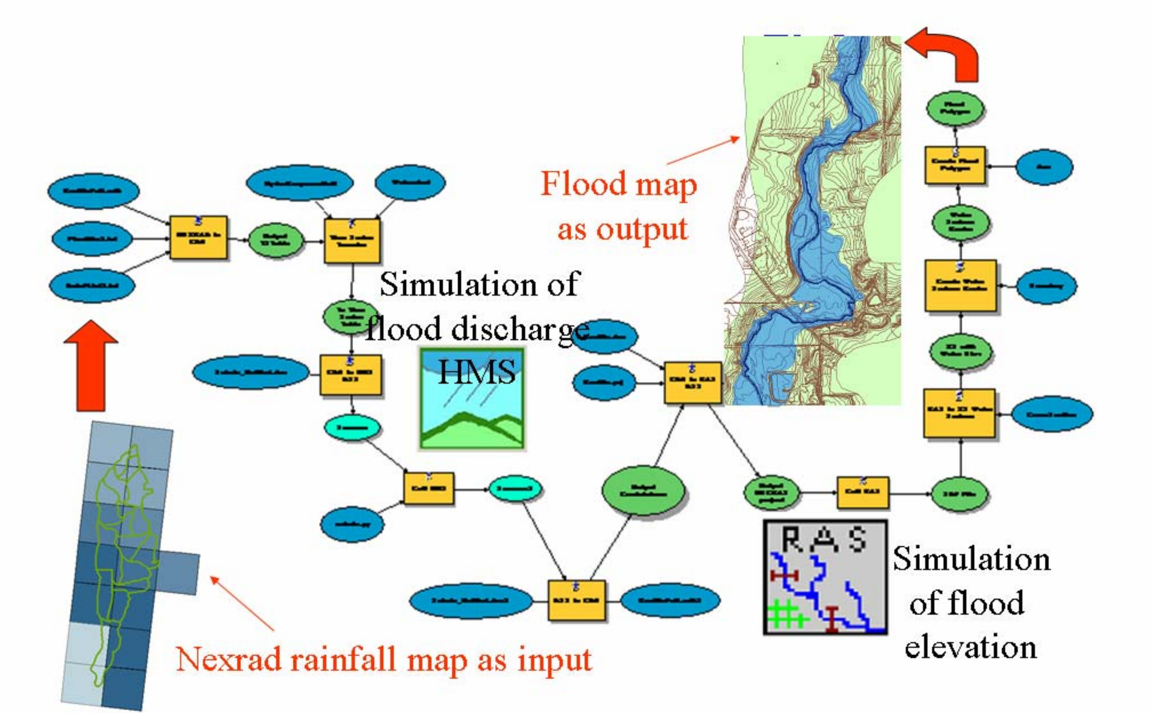


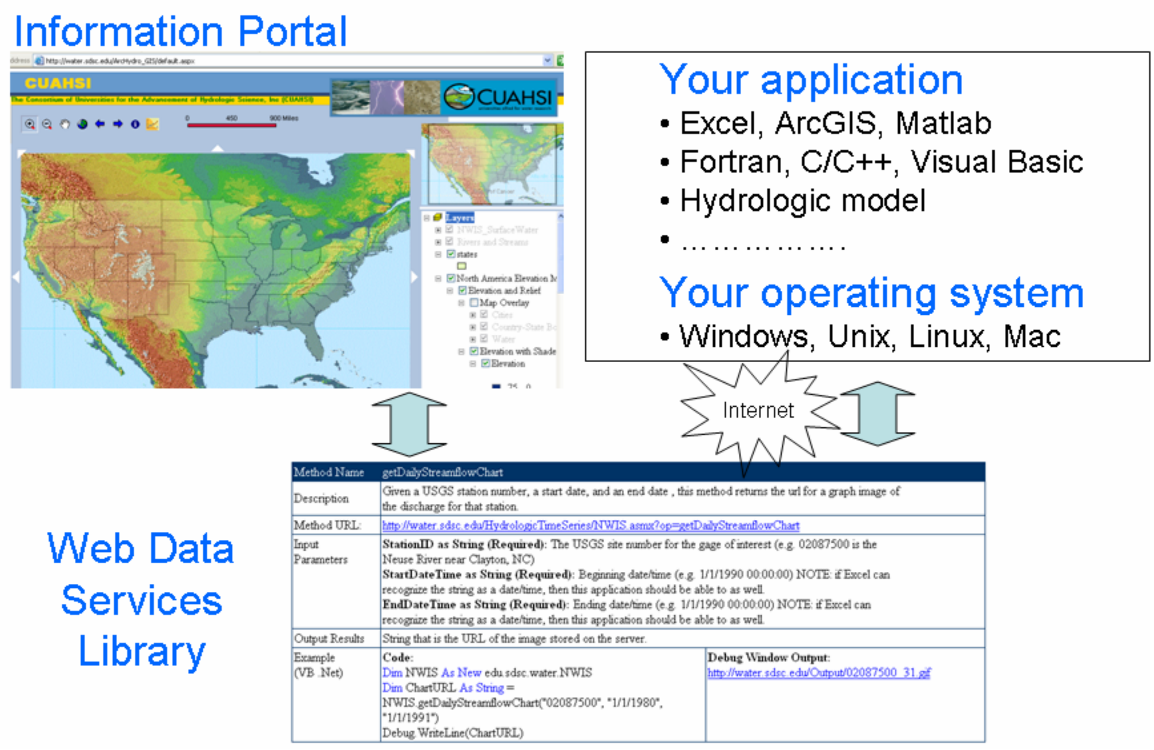
Figure 4. Workflow sequence for transforming Nexrad radar rainfall to flood inundation maps using ModelBuilder and the HEC-HMS and HEC-RAS simulation models

Web Data Services

One of most critical functions of the cyberdashboard of an information portal is to provide rapid access to information sources. In hydrology, a very important class of information sources are the federal agencies and other entities that routinely measure hydrologic variables such as rainfall, streamflow, water quality, groundwater levels, climate at point locations such as gages and sampling sites. These agencies provide access to their data through the internet but the number and variety of the websites, all having their own individual structure and method, means that it practically requires a PhD in websitology to know how to go and work all of them, even if a user knows that they exist. And all the operations on each web site are manual, tedious and time consuming, so that it takes hours to accomplish what should be a task of minutes or seconds. What is needed is some “common data window” to which a hydrologist can go and acquire data through a single interface, which itself is interacting with the various internet data sources with their many formats and methods but the user does not have to know anything about that. This is like the service that Travelocity provides – American Airlines and Hilton Hotels may have completely different reservation services, but a user of Travelocity chooses to fly American and stay in a Hilton though a single information system.

Figure 5 shows the CUAHSI Information Portal which provides access to point observation sites with measured hydrologic information across the United States using a map background to provide the spatial context of where the measurement sites are located. In responding to information requests, the portal is accessing an underlying *web data services library* which is a set of elementary functions like “getStreamflowdata”, “getWaterQualityParameters”, “getStationInfo”, which are executed directly on the USGS National Water Information System, without any manual interaction through the NWIS web site. The results returned can be graphed in the portal to enable data inspections before downloading any information.

The beauty of web data services is that through standard information protocols called SOAP and REST the same services can be accessed by any operating program on any internet connected computer. Thus, the same CUAHSI web data services for NWIS can be called from an Excel program on your local computer and the result is that data is downloaded straight from NWIS into your Excel spreadsheet. And it does not matter whether Excel is running in Windows or on a MacIntosh because the computer-computer interactions are handled by SOAP or REST. The same services can be programmed into a Perl script on a Linux machine and NWIS data will automatically be ingested there also.



Information Portal

Your application

- Excel, ArcGIS, Matlab
- Fortran, C/C++, Visual Basic
- Hydrologic model
- ……………….

Your operating system

- Windows, Unix, Linux, Mac

Web Data Services Library

Method Name	getDailyStreamflowChart	
Description	Given a USGS station number, a start date, and an end date , this method returns the url for a graph image of the discharge for that station.	
Method URL:	http://water.sdsc.edu/HydrologicTimeSeries/NWIS.asmx?op=getDailyStreamflowChart	
Input Parameters	**StationID as String (Required)**: The USGS site number for the gage of interest (e.g. 02087500 is the Neuse River near Clayton, NC) **StartDateTime as String (Required)**: Beginning date/time (e.g. 1/1/1990 00:00:00) NOTE: if Excel can recognize the string as a date/time, then this application should be able to as well. **EndDateTime as String (Required)**: Ending date/time (e.g. 1/1/1990 00:00:00) NOTE: if Excel can recognize the string as a date/time, then this application should be able to as well.	
Output Results	String that is the URL of the image stored on the server.	
Example (VB .Net)	**Code:** Dim NWIS As New edu.sdsc.water.NWIS Dim ChartURL As String = NWIS.getDailyStreamflowChart("02087500", "1/1/1980", "1/1/1991") Debug.WriteLine(ChartURL)	**Debug Window Output:** http://water.sdsc.edu/Output/02087500_31.gif

Figure 5. Web data services underlying the CUAHSI Information Portal.

As shown in Figure 6, what will be constructed are CUAHSI web data services for the National Water Information System (NWIS), the National Climate Data Center (NCDC), the EPA Storet system for water quality, the USGS National Water Quality Assessment (NAWQA), the Ameriflux network of atmospheric flux towers, the Long Term Ecological Research (LTER) network, and the National Center for Atmospheric Research. In one form or other, connections have been made by the CUAHSI Hydrologic Information System team with each of these data systems and with the people who operate them, and sufficient research has been done to verify that the proposed technical approach is feasible in each case, though the degree of difficulty varies from one information source to another. A successful set of CUAHSI web data services has been built to the NWIS data system and is presently being reviewed for functional integrity by the USGS NWIS team.

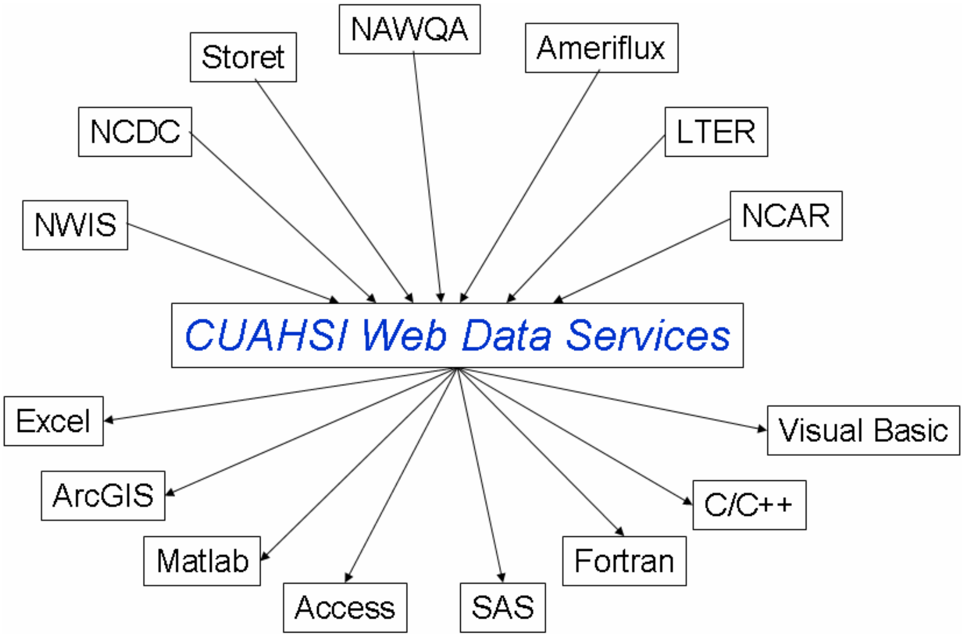
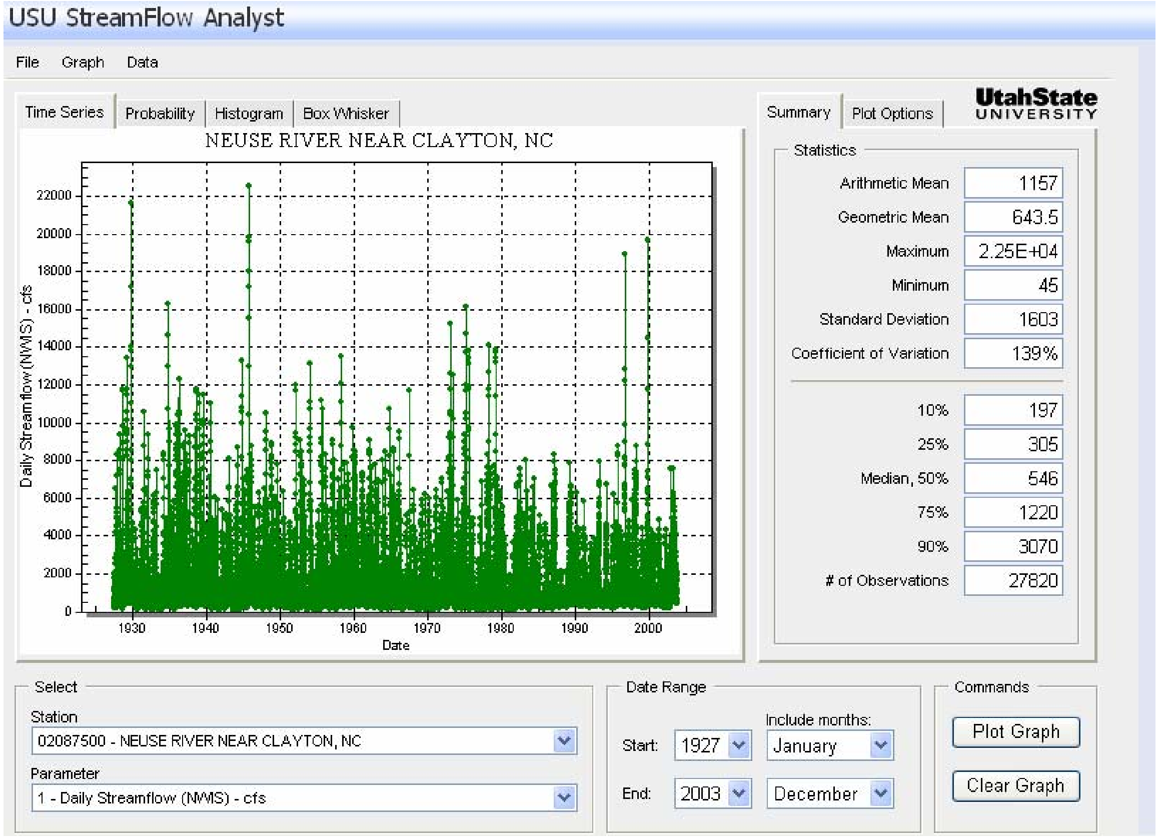


Figure 6. CUAHSI web data services directly connect hydrologic data sources with application programs and programming languages.

The Hydrologic Information System user needs survey identified the highest priority applications and programming languages favored by significant numbers of the CUAHSI community. These include Excel, ArcGIS, and MatLab as application systems; Fortran, C/C++ and Visual Basic as programming languages, and MS Access and SAS for data archiving and statistical analysis, respectively. The mechanisms by which CUAHSI web data services can be imbedded in these systems are being investigated. Of course, the CUAHSI community uses many more tools and languages than those shown in Figure 6, and the CUAHSI web data services will be open to all CUAHSI institutions so that they can be imbedded into whatever applications or programming languages that an investigator wishes to use.

Utah State University Streamflow Analyst

A remarkable example of the utility of the CUAHSI information portal and web data services is provided by their utilization by the Utah State University Streamflow Analyst (also called Time Series Analyst). This software tool, programmed by Jeff Horsburgh at the Utah State University Water Research Laboratory, is designed to allow uses to access time series of water observations, plot them as a time series graph, as a cumulative frequency curve, as a histogram, and as a monthly varying box and whisker plot, and to summarize their statistical characteristics. When originally programmed, this system accessed data from its own local database with its own special data structure, into which NWIS and other data from the Bear River watershed in Utah had been loaded. When reprogrammed so that it could also draw data from the CUAHSI web data services for NWIS, the USU Streamflow Analyst was immediately able to access and plot NWIS data from anywhere in the United States. Then the USU Streamflow Analyst was included as a module in the CUAHSI information portal so it is accessible to CUAHSI members anywhere in the nation. In this manner, a tool written by one person in one CUAHSI institution accessing only its own local information is transformed into a tool can be accessed by CUAHSI members anywhere in the nation and can be applied to observational data anywhere in the nation! This is a completely remarkable transformation, increasing by many thousands of times the value of the original tool. And this USU Streamflow Analyst was developed independently of the CUAHSI HIS project. This example shows how CUAHSI information services can magnify the value of services contributed by CUAHSI and provide access to them throughout the CUAHSI community. There is no question that the impact of having a Consortium like CUAHSI representing 100 universities means that the federal data agencies treat the process of building web services to their data systems seriously when CUAHSI does it, as distinct from what reaction they may have if individual hydrologic investigators tried to do this on their own.



Statistics

Arithmetic Mean	1157
Geometric Mean	643.5
Maximum	2.25E+04
Minimum	45
Standard Deviation	1603
Coefficient of Variation	139%
10%	197
25%	305
Median, 50%	546
75%	1220
90%	3070
# of Observations	27820

Select

Station: 02087500 - NEUSE RIVER NEAR CLAYTON, NC

Parameter: 1 - Daily Streamflow (NWIS) - cfs

Date Range

Start: 1927 Include months: January

End: 2003 December

Commands

Figure 7. Utah State University Streamflow Analyst

Data Cube

The description of any information system is replete with terms like cyber this and web that but, in the end, a hydrologic information system has to describe hydrologic phenomena. In hydrology, this is a complex task, because water flows through the atmosphere, land surface and subsurface, and these are three very different kinds of flow environments (Figure 8). Atmospheric water is transported by air in a continuous fluid domain extending to perhaps 15 km up into the atmosphere. The land surface is the only flow boundary and the exchange of water and energy between the land surface and the atmosphere is an important boundary condition for atmospheric circulation. Surface water flows in a concentrated fashion along flow paths through rills and hollows, streams and rivers. Driven by gravity, these flow paths are closely related to the gradient and curvature of the earth's surface, itself a product of erosion processes occurring over eons of time. Subsurface flow occurs through soil and rock strata, some permeable, others not. The flow is often visualized as being spatially continuous through a homogeneous subsurface medium as in Figure 8, but in fact there are subtle variations and gradients in subsurface properties, and sometimes sharp changes in them at the boundaries of subsurface strata.

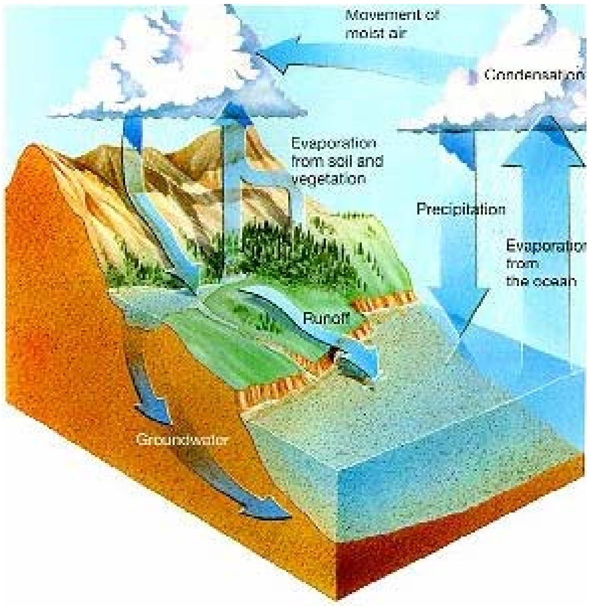


Figure 8. Water flows through the atmosphere, land surface and subsurface.

In general terms, if there is a variable V whose values are a function of spatial location L and time T, a data value D can be visualized in a *data cube* as a function D(V, L, T), as shown in Figure 9. There may be many variables on the variable axis; space may be represented in one, two or three dimensions, or may refer to the properties of discrete spatial objects represented by points, lines, areas or volumes; the data values may occur regularly or irregularly in time.

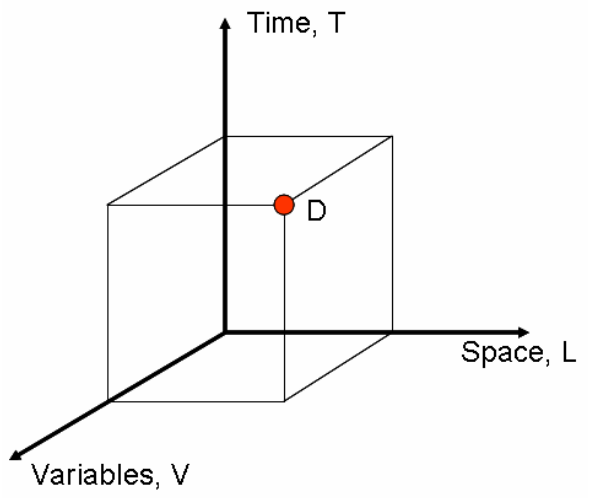


Figure 9. The data cube

The *netCDF* data model for continuous variation in space and time is illustrated in Figure 10. All variables in a netCDF file are called *dimensions* – those that serve to index the location in space and time (the set {X} in Figure 10) are called *coordinate dimensions*, while those whose values are defined at points in the coordinate space, are called *variable dimensions* (the set {Y} in Figure 10). The analogy is with regression analysis where dependant variables, y, are estimated as a function of the independent variables, x. NetCDF files describe the sampled values of an n-dimensional function space. For example, suppose that two variables, temperature and relative humidity, are defined at a set of latitude and longitude pairs and at specified points in time. There would then be five dimensions to the function space of which three (latitude, longitude and time) are the coordinate dimensions and two (temperature and relative humidity are the variable dimensions). Tools to manipulate netCDF files are provided by Unidata, an organization under UCAR in Boulder, Colorado, which plays the same role for academic data support for the atmospheric sciences as CUAHSI Hydrologic Information Systems plays for hydrology. NetCDF is the most widely employed data representation format for spatially distributed information in the atmospheric and ocean sciences.

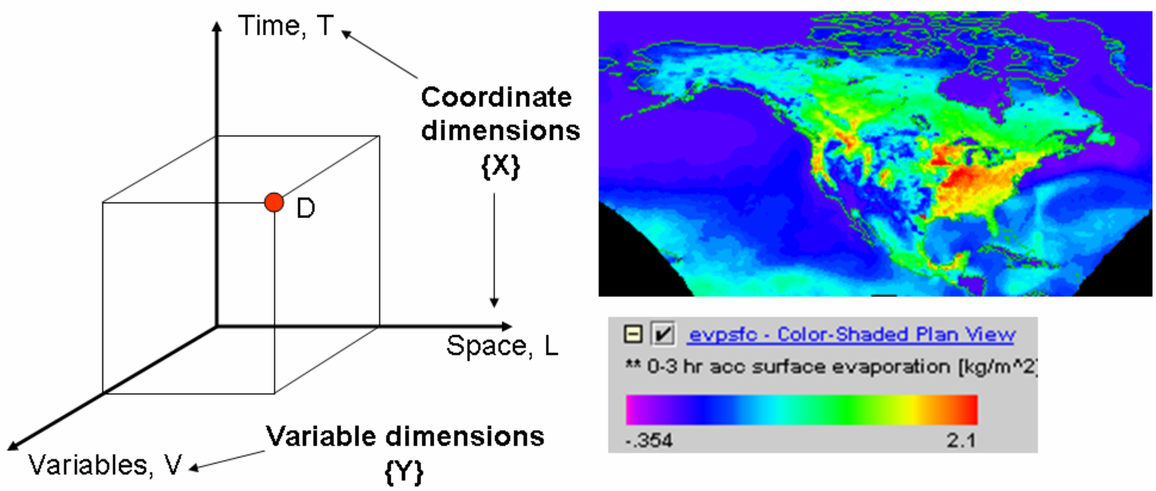
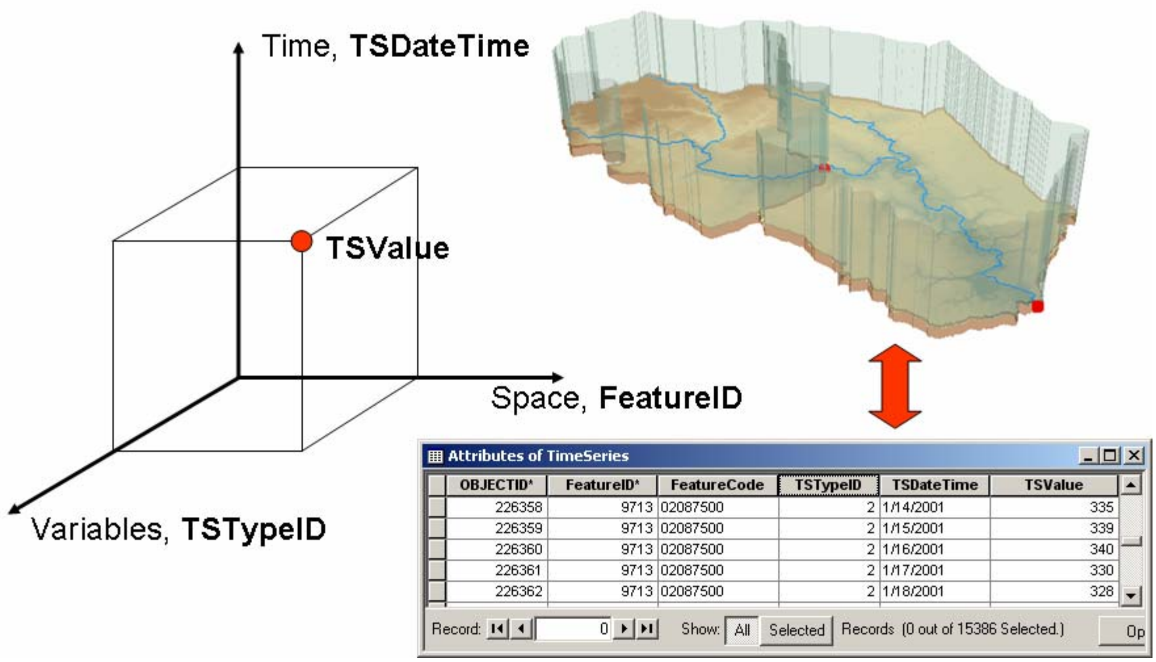

Time, T
Coordinate dimensions {X}
D
Space, L
Variables, V
Variable dimensions {Y}
⊟ ☑ evpsfc - Color-Shaded Plan View
** 0-3 hr acc surface evaporation [kg/m^2]
-.354
2.1

Figure 10. The netCDF data model for continuous variation in space and time illustrated by the distribution of forecast evaporation over North America.

The Arc Hydro data model for describing time varying variables in a discrete space domain is shown in Figure 11. Points may represent gaging stations, river segments may be lines, watersheds may be areal polygons, and hydrogeological units may be volumes – all these are discrete space objects, or *hydrofeatures*, because they have definite spatial location and boundaries. They are usually fixed in location and shape but may occasionally have time-varying geometries, such as flood inundation boundaries as a flood passes along a river valley. In Arc Hydro, all points, lines, areas and volumes are individually and uniquely identified by their *HydroID*, an integer identifier that is applied carefully using a toolset that prevents a HydroID once assigned to a hydro feature, from ever being reassigned to another feature in the same dataset.

Arc Hydro is a customization of ArcGIS for water resources information so the values of its variables are stored as *attributes* in tables, such as that shown in Figure 9c. The variable type is indexed by the TSTypeID field, the value of the HydroID of the hydrofeature described by the variable is contained in the FeatureID field of its time series table, time is indexed by the TDDateTime field, and the actual value of the variable is stored in the TSValue field. Some additional fields are shown in Figure 11 – FeatureCode serves to store the permanent public identifier of this feature when outside Arc Hydro – the USGS site number in this instance, and the ObjectID is a row index used by ArcGIS. Tabular data models that index information in space and time are appropriate for storing hydrologic observations data measured at point locations and for describing time varying spatial properties of hydrofeatures, such as the average moisture content of the soils of a watershed.



Attributes of TimeSeries

OBJECTID*	FeatureID*	FeatureCode	TSTypeID	TSDateTime	TSValue
226358	9713	02087500	2	1/14/2001	335
226359	9713	02087500	2	1/15/2001	339
226360	9713	02087500	2	1/16/2001	340
226361	9713	02087500	2	1/17/2001	330
226362	9713	02087500	2	1/18/2001	328

Figure 11. The Arc Hydro data model for describing time varying values on discrete spatial features.

Up to this point, it has been assumed that all variables are deterministic – that is, at any point in space and time, there is a single value of the variable. Suppose instead that the variable is random, and that the given spatial location and time point, it is described by a probability distribution as shown in Figure 12. If one considers spatial variation alone, *geostatistics* is well-developed to describe random functions; similarly, *time series analysis* is well suited to describing random variables whose values change in the time dimension; *multivariate analysis* describes the mutual interaction of sets of random variables. Putting together space, time, and probability is the ultimate challenge in representation of hydrologic variables, and sound theories and methods for dealing with nonhomogeneous, nonstationary random fields need further development.

The netCDF, Arc Hydro and random variable data representations all have advantages and limitations for application in hydrology. NetCDF is fine for describing water properties and motion in continuous fluid domains, such as the air, lakes, estuaries and the oceans, but it breaks down in discrete space domains such as watersheds and stream networks. The Arc Hydro tabular method for storing information works well for point observation data and static spatial attributes of maps, but breaks down when discrete spatial hydrofeatures have arrays of dynamic variables defined upon them.

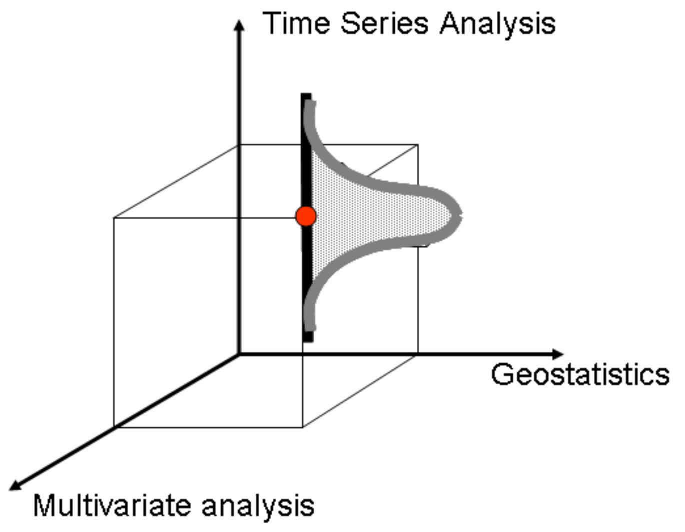


Figure 12. A random variable described by a probability distribution whose properties vary in time and space.

Suppose that the netCDF array model were applied to discrete spatial objects – in this instance, spatial location would be indexed by the HydroID of the hydrofeatures which would be treated as a coordinate dimension in the netCDF dataset. In this manner the freedom of the array model for describing mathematical processes would be combined with the geographic precision of the GIS method for describing discrete space objects. This combination would enable simulation of hydrologic processes in surface and subsurface flow, where the spatial location and interaction among the hydrofeatures would be indexed using spatial analysis in the GIS. This data model, here termed Geographic NetCDF may be a useful approach for describing the fluxes, flows and mass balances of interest in hydrology.

Summary

The preceding discussion has covered a very wide scope from the conceptualization of a hydrologic information model into four information spaces, the translation of this model into a cyberinfrastructure design for a digital hydrologic observatory using a hydrology data portal to access a digital hydrologic observatory information repository that includes web data services for ingestion of hydrologic observation data from government agencies, and culminating in variations on the data cube to capture various approaches to representing hydrologic variables that vary in space and time, and may be random as well. This conceptual framework for the CUAHSI Hydrologic Information System probably omits some important factors, but it serves at least as a point of departure for discussion and refinement of the definition of key elements of this system.

Chapter 3

System Architecture

By Chaitan Baru, Ilya Zaslavsky and Reza Wahadj
San Diego Supercomputer Center

1. Open services-oriented hydrologic observatory architecture

1.1 The vision

The CUAHSI HIS system architecture is envisioned as a component of a large scale environmental observatory effort, which emerges as a network of seamlessly integrated data collection, information management, analysis, modeling and engineering endeavors implemented across disciplinary boundaries. The CUAHSI HIS system architecture design is presented here as one of the building blocks of a larger cyberinfrastructure supporting digital observatories (Figure 1). This supporting cyberinfrastructure represents a network of data repositories, digital libraries, and analysis and modeling services annotated with metadata and knowledge bases and organized into analytical workflows, which are accessible from a variety of data portals. CUAHSI HIS develops components of the cyberinfrastructure specific for hydrologic research, focusing on representing, manipulating and sharing data objects that reflect fundamental hydrology concepts, such as the digital watershed. At the same time, CUAHSI HIS cyberinfrastructure components are being developed in a standards-compliant way, compatible with other large NSF-supported research and infrastructure earth sciences projects, including GEON, SEEK and NEON, and in particular with the environmental engineering agenda within the emerging CLEANER effort.

The CUAHSI community has already developed a plethora of databases, data analysis and visualization models and tools, including various watershed and flow models and mapping and time series visualization systems. Other data resources of interest to CUAHSI community, as confirmed by the survey in Chapter 3, are provided by federal agencies and include large repositories such as the USGS's NWIS, the EPA's STORET, etc. The goal of CUAHSI HIS architecture development is to create an environment where these different elements work in concert to support advanced data intensive hydrology research. This includes providing easy analytical access to the distributed data resources, ability to publish and manage local observational and model data, and interface the data with a variety of community models and analysis and visualization codes. The system architecture outlined in this chapter addresses these goals, specifically focusing on integrating platforms, research and analysis tools, and data sources uncovered in the course of CUAHSI user needs assessment.

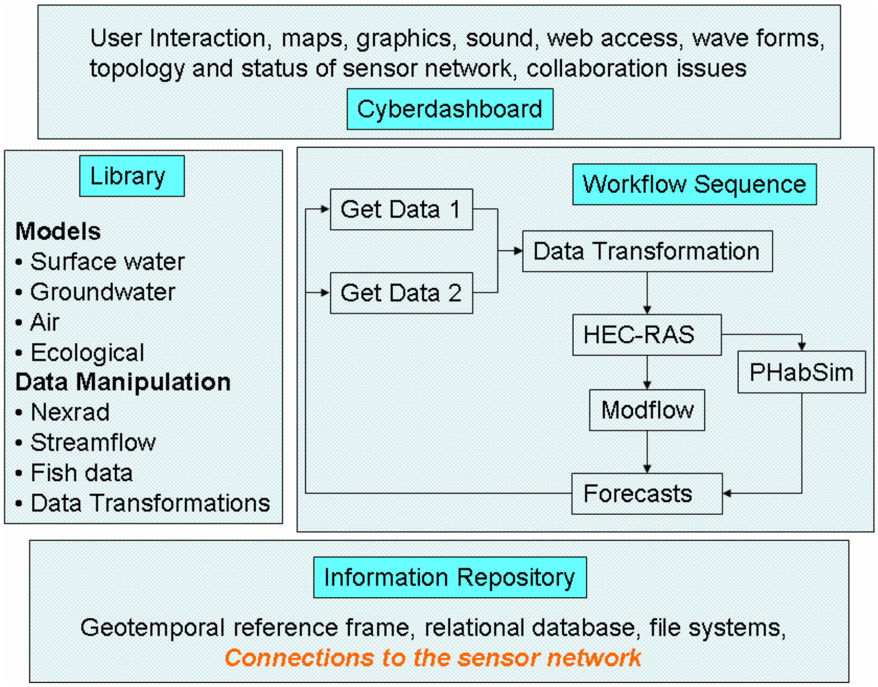


Figure 1. Supporting cyberinfrastructure for the Digital Observatory: a general vision

1.2 The services model for a digital hydrologic observatory

The CUAHSI Hydrology Information System design follows the open services-oriented architecture model that has been explored and developed in several large-scale federally funded cyberinfrastructure projects. The *services-oriented architecture* (SOA) relies on a collection of loosely coupled self-contained services that communicate with each other and can be called from multiple clients in a standard fashion. Services provide a useful abstraction for functionality accessible over the web, by establishing a standard protocol (e.g. SOAP – Simple Object Access Protocol, or REST – REpresentational State Transfer) for invoking services irrespective of their underlying language, and by establishing a standard "contract" between a service provider and service client that can be used to formulate correct requests against a service (e.g. WSDL – Web Services Description Language). Common benefits associated with services-oriented architecture include: scalability, security, easier monitoring and auditing; standards-reliance; interoperability across a range of resources; plug-and-play interfaces. Internal service complexity is hidden from service clients, and backend processing is decoupled from client applications. In other words, different types of clients, including Web browsers and such desktop applications as Matlab, ArcGIS and Excel, exposed as the primary desktop client environments by the CUAHSI user needs assessment, will be able to access the same service functionality, leading to a more transparent and easier managed system.

The services model has substantial backing from the industry, with SOAP/WSDL web services being the first large-scale interoperability standard jointly supported by the

Microsoft and Java development communities. Web services-aware development tools and systems are now becoming ubiquitous in both commercial and academic applications. The model emphasizes external service interfaces rather than the internals of the service; the latter are hidden from the user thus making it possible to develop services in multiple programming languages, using open source or commercial components as needed. With the size and diversity of CUAHSI, and different software development environments and constraints of different CUAHSI software development teams, such openness and flexibility are important components of a successful infrastructure implementation.

1.3. Web services

Under the services-oriented architecture model, all processing and database access functions are “wrapped” in web service wrappers. Examples of such web service wrappers include interfaces to CUAHSI digital library, grid management infrastructure for CUAHSI nodes, NWIS data access functions, etc. They are reviewed in more detail below.

The reality of data intensive disciplines like hydrology is that most data extraction and manipulation services have to be near databases, to minimize unnecessary data transmission in the systems. The CUAHSI HIS team has been communicating with database experts at USGS and EPA (managers of NWIS, NAWQA and STORET online databases) to establish data access via web services installed at the agencies. In the meantime, all data access web services are running on a SDSC computer with ESRI’s ArcGIS Server. This organization allows us to query and retrieve agency data into an ArcHydro-like structure (described in detail in the chapter on hydrologic data models) and analyze and visualize the data using ArcObjects and other components available through ArcGIS Server, deposit the data into a digital library, etc.

Web services are alluringly easy to develop and consume, especially in the .NET environment where Visual Studio IDE hides a lot of infrastructure details from the coder. The CUAHSI effort would certainly benefit if many partners develop web services on their own. However, this increases the potential for unsystematic development, duplication, and possibly incompatibilities between related services. Therefore, one of the goals of this chapter, in addition to reporting the development status, is to outline a reference collection of CUAHSI web services to be developed within the CUAHSI HIS project, as a guide for external developers. Additionally, we outline web services leveraged from neighbor projects. These projects include, most importantly, the NSF-funded GEON, NEON, SEEK and CLEANER efforts, NIH-funded BIRN, as well as several NASA-sponsored undertakings such as Geobrain and DataFed.

1.4 On applicability of web services, and best practices within CUAHSI

Web Services is not a panacea for all architectural issues. Each technology has limitations. Here are some relevant limitations:

- Web services are not suitable for fine grained interactions, or in tightly coupled, high-volume internal applications. Determining the right level of granularity is the most difficult portion in developing a SOA. Basically, having as few methods as possible to represent functions of a subsystem, such that each method does one complete logical function, is the preferred model. The NWIS driver services developed within CUAHSI so far are being revised from this perspective, as described below.
- If a web service-based processing pipeline involves cycles of serializing objects into XML, sending them over network then importing the data back into similar objects, this results in unnecessary overhead. Instead, it would be better to persist objects where the processing occurs, thus avoiding serialization overhead. With respect to this limitation, treatment of hydrologic time series objects, once they are populated from a data source, deserves careful consideration. In particular, the common CUAHSI web service design scenario suggests that a data retrieval service has at least two return options: returning either the serialized data or a reference to data object ID residing on a server. This consideration becomes especially important in the design of a workflow system based on Web services. For efficiency, web services in such a system would generally exchange references to object IDs rather than XML-ized data.
- Web services should align with data objects developed within CUAHSI, such as the hydrologic time series object described in Chapter 6, though at a coarser level. For example, NWIS web services are currently being organized into a data retrieval service and a digital watershed service each having multiple methods that mirror methods in the object model, and share much of the code base.
- Web services are most suitable when we need to interoperate across different computing platforms. CUAHSI services have been developed mostly in .NET so far while many services in related cyberinfrastructure projects are Java-based and run mostly on Linux servers. However, interoperation between Java and .NET services shouldn’t present a problem, also because Java services (as in GEON, especially) are mostly core infrastructure services, while in the CUAHSI project’s .NET services support distributed data access and retrieval, and charting and mapping applications. Therefore, chances that such services would exchange complex objects and potentially collide are low.
- Services should have *high performance* (which implies the use of lightweight parsers and simple data types, and sending compressed XML in messages; *reliable, fault-tolerant, and accessible* (i.e. supporting load balancing and graceful fallback), *scalable* (ability to run in a cluster), *secure* (support encryption, role-based security), have means to control *data integrity* (standard conformance, plus checksums where applicable). These principles are followed in CUAHSI web services development.

Building on this outline of the general principles of web service development, applicability limits, and dos and don’ts with respect to CUAHSI HIS effort, we continue by laying out the architecture of the CUAHSI HIS information system. The following section defines groups (layers) of web service interfaces enabling access to and communication between different system components. A list of services follows from the architecture discussion, noting services already developed within CUAHSI or other projects, services to be developed within CUAHSI, and services that we can reasonably expect to be developed by partner projects. This outline is primarily based on the intense discussions among CUAHSI HIS partners in late spring – early summer 2005, and several design documents generated at that time, including the design of CUAHSI data portal, principles of web services-based architecture, and the list of 85 proposed CUAHSI services developed by David Maidment. All the design documents, as well as individual web service descriptions, are available from CUAHSI HIS portal at http://gis.sdsc.edu/cuahsi.

2. CUAHSI Services Oriented Architecture development

2.1 CUAHSI web services architecture

In this section, we extend the architecture diagram in Figure 1, populating it with CUAHSI-specific objects. Figure 2 presents a vision of digital observatory as a higher-level object that is comprised of discipline specific objects such as digital watershed, digital estuary, digital aquifer, etc. Such objects are generated by applying various information models and data integration and workflow techniques to a range of lower-level objects. Later in this chapter, we propose a formal representation of digital watershed supporting: a) querying a watershed at the symbolic level, and assessing digital watershed consistency and completeness, b) publication of digital watershed data as web services (mapping services, in particular), and c) instantiating a digital watershed configuration in a data warehouse for further off-line analysis. The lower-level objects comprising a digital watershed would include time series objects, channel objects, spatial data layers, etc. These objects are populated from a variety of data sources, both hosted by the observatory and external data repositories (federal data sources at USGS, EPA, etc.), and from metadata and knowledge descriptions associated with the sources.

The focus of CUAHSI HIS effort to date has been on constructing the data foundation for the digital hydrologic observatory. Thus the bulk of this chapter describes the data-related services that are either already developed or being developed within the project. The next phase is management of higher-level objects such as digital watershed, which is discussed in the last part of the chapter.

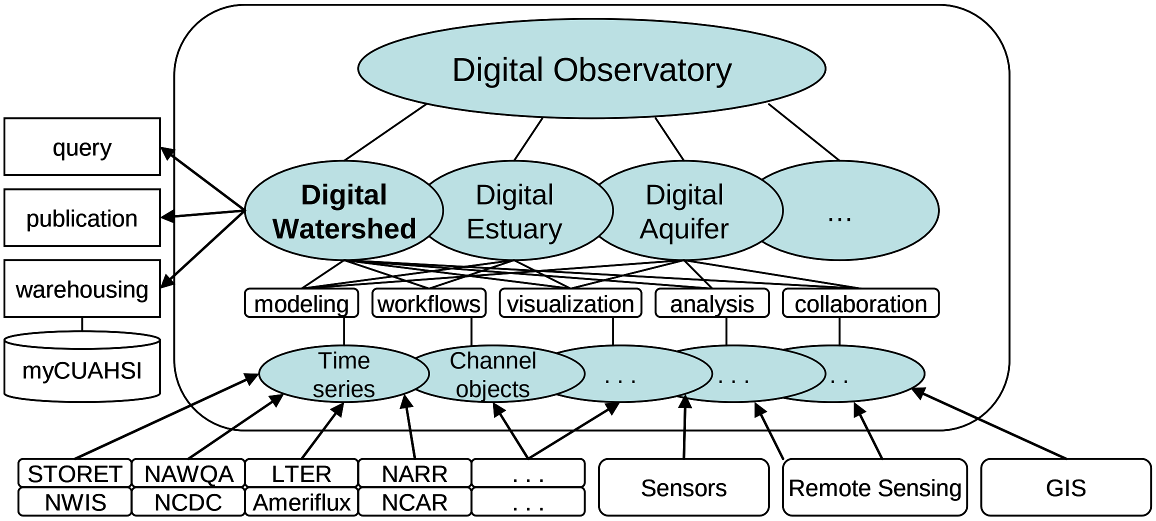


Figure 2. Higher-level object hierarchy in CUAHSI HIS.

The architecture diagram (Figure 3) further describes the main structural components of CUAHSI HIS, and groups of web services we are developing. The numbers below correspond with the numbers on the diagram.

1. Core services: monitoring and managing CUAHSI PoP (Point of Presence) nodes. Such services have been prototyped in the GEON project.

2. Resource driver services: access to various external resources, including federal hydrology-related databases. Services in this group support data search and retrieval against individual external databases. An example is services being developed for accessing the USGS NWIS repository, which are considered in detail below. In addition to data search and retrieval, services in this group will support periodic metadata updates (i.e. updating NWIS stations layer and other information that should not be requested from external sources at query time).

3. Sensor management services: monitoring and managing sensors and other continuous observation instruments. In a hydrologic observatory setting, such services may be developed in collaboration with CLEANER, LTER, NEON and other observatory projects.

4. Sensor data filtering services: managing aggregation/filtering of sensor data, depositing the data in a digital library and updating corresponding metadata records.

5. Digital library services: harvesting, uploading, searching and retrieving data hosted by CUAHSI data nodes. These services are described below, several of them are operational.

6. Services for global query and data retrieval: querying both across external data sources and CUAHSI-hosted metadata, and orchestrating data retrieval from multiple sources.

7. Ontology services: ontology generation, ontology query, translation and update services primarily used in conjunction with the global query and data retrieval services.

8. Application level ("digital watershed") services. This is a large group of services that supports manipulation of higher-level HIS objects, in particular the digital watershed object. The services include: compilation of a digital watershed from lower-level objects, analysis of digital watershed for completeness (data gaps) and consistency (incompatibilities in data resolution, temporal or spatial frameworks, access mechanisms, etc), transformations of digital watershed data; workflow orchestration services, etc.

9. Authentication/authorization services (usually tied to portal framework)

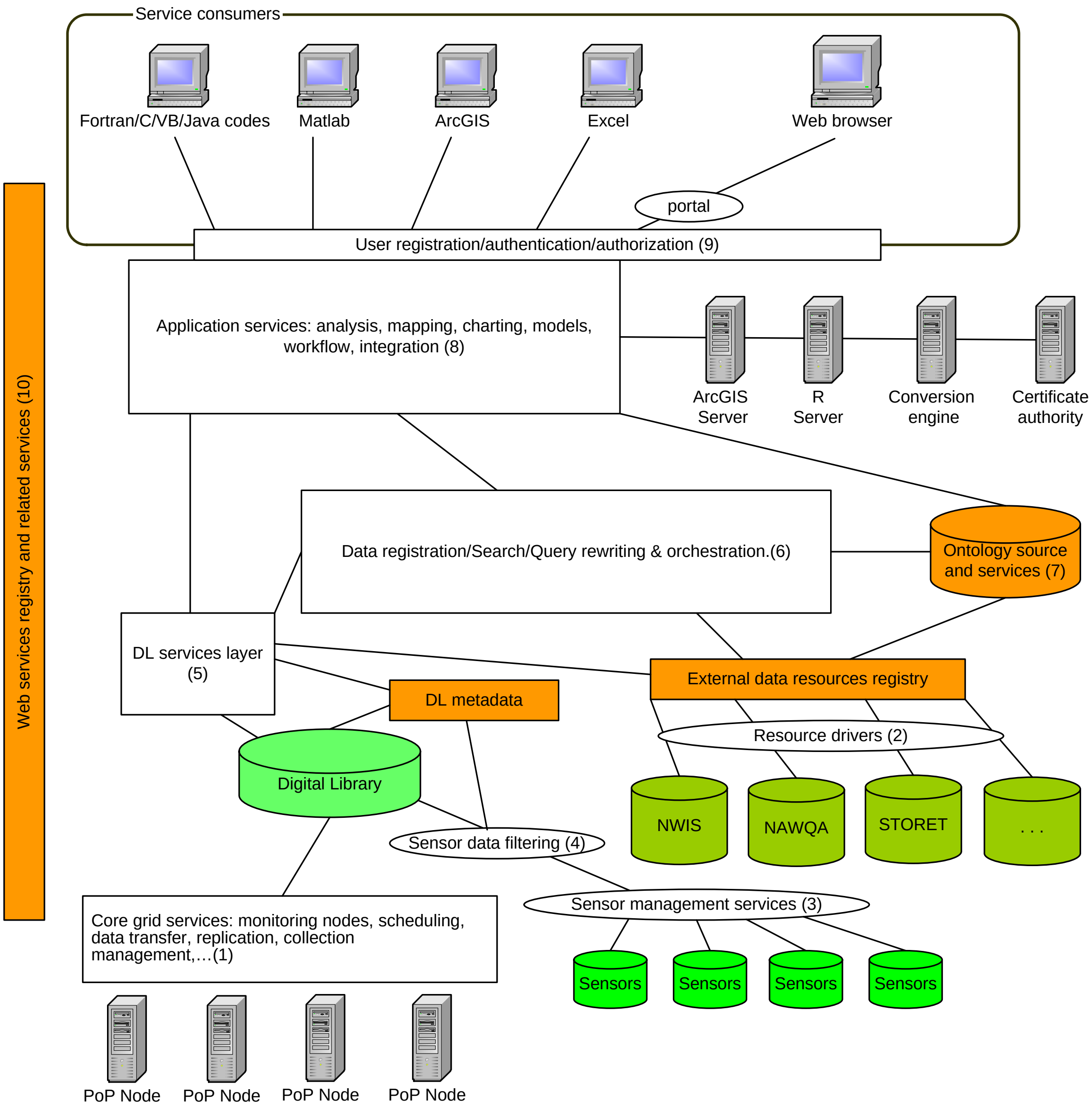
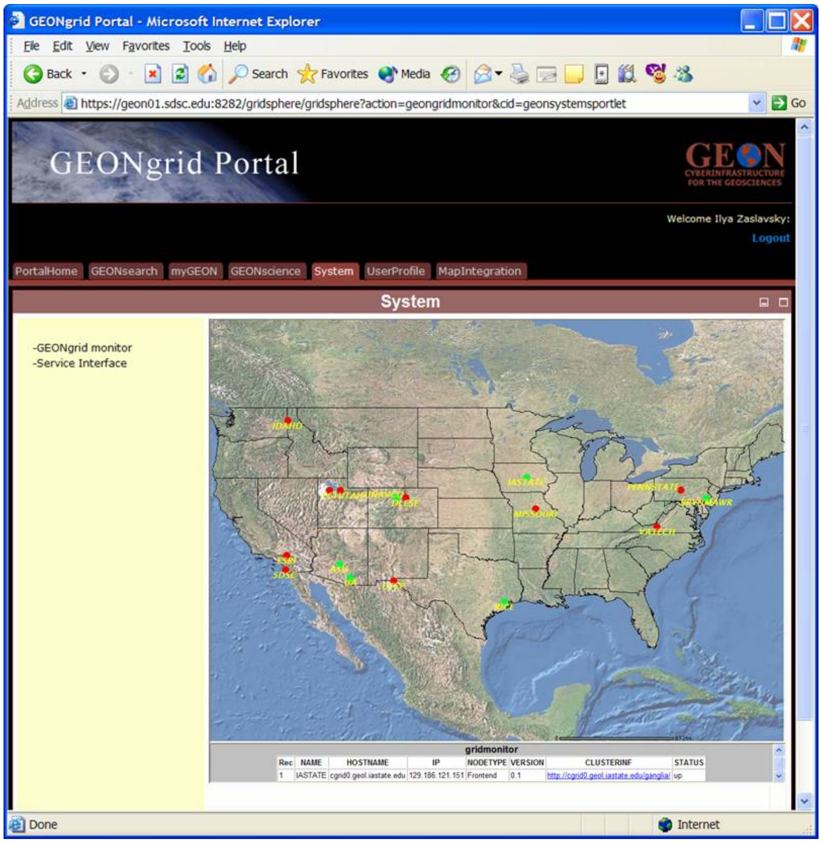

Service consumers
Fortran/C/VB/Java codes
Matlab
ArcGIS
Excel
Web browser
portal
User registration/authentication/authorization (9)
Web services registry and related services (10)
Application services: analysis, mapping, charting, models, workflow, integration (8)
ArcGIS Server
R Server
Conversion engine
Certificate authority
Data registration/Search/Query rewriting & orchestration.(6)
Ontology source and services (7)
DL services layer (5)
External data resources registry
DL metadata
Resource drivers (2)
Digital Library
NWIS
NAWQA
STORET
. . .
Sensor data filtering (4)
Sensor management services (3)
Core grid services: monitoring nodes, scheduling, data transfer, replication, collection management,…(1)
Sensors
Sensors
Sensors
Sensors
PoP Node
PoP Node
PoP Node
PoP Node

Figure 3. Architecture of CUAHSI web services

2.2 Individual services outline

This section adds service description details for each group outlined above.

1. Core services. These services for managing and monitoring CUAHSI PoP nodes are being developed primarily within the GEON project. GEON software stacks with the core node management services are installed on nodes at SDSC and at UIUC. The SDSC-UIUC collaboration on CUAHSI digital library for Illinois River Watershed is described below, in the section on digital library services.


GEONgrid Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://geon01.sdsc.edu:8282/gridsphere/gridsphere?action=geongridmonitor&cid=geonsystemsportlet Go
GEONgrid Portal
GEON CYBERINFRASTRUCTURE FOR THE GEOSCIENCES
Welcome Ilya Zaslavsky:
Logout
PortalHome GEONsearch myGEON GEONscience System UserProfile MapIntegration
System
-GEONgrid monitor
-Service Interface
IDAHO
UTAH
UNAVCO
DLESE
IASTATE
PENNSTATE
BRYNMAWR
MISSOURI
VATECH
ESRI
SDSC
ASU
UA
RICE
Done
Internet

gridmonitor

Rec	NAME	HOSTNAME	IP	NODETYPE	VERSION	CLUSTERINF	STATUS
1	IASTATE	cgrid0.geol.iastate.edu	129.186.121.151	Frontend	0.1	http://cgrid0.geol.iastate.edu/ganglia/	up

Figure 4. GEON grid monitor service

Of the core node management services, *GeonPingService* and *GeonResources* are the core services that can be reused in CUAHSI. These services, in particular, are used to monitor the status of grid nodes installed at partner institutions (Figure 4). Currently, a Point-of-Presence (PoP) and a Data node are installed at UIUC, and can be monitored and managed via a similar interface.

Additional scheduling, load balancing, replica management, synchronization, and data transfer with encryption/compression, will be developed within GEON or related projects.

2. Resource driver services. We consider two types of drivers: services invoked at query time, and services used to update station metadata. The first group includes drivers that populate TimeSeries objects from a given data resource and optionally serialize the data or perform various operations on the objects. The second group includes services that retrieve station lists and metadata from each source to update external data resource registry at regular intervals.

    a. The following drivers of the first group have been created for NWIS: *getDailyStreamFlowChart*, *getGWLevelValues*, *getWQValues*, *getDischargeValues*. Since all of them take the same inputs (with minor variations) and return either a TimeSeries document or a chart image, it may be useful to unify them in a single *NWISTimeSeries* web service with a *getValues* method that accepts stationID, parameterName, startdate and enddate as inputs, and returns a time series string or a URL (CUAHSI-ID) to a TimeSeries object on the server. Optionally, the output may include a URL to a chart, or URL to a thematic map – but these would rather be implemented as separate methods. Ideally, an *NWISTimeSeries* web service would closely mirror the TimeSeries object being developed by T. Whiteaker, UT-Austin. Web services developed within the NASA DataFed project, have similar design, though connect with different data sources.

b. Other services developed currently for NWIS, belong to the second group: *getStationsWithWQParameter*, *getDischargeInfo*, *getStationInfo*, *getWQParamName*, *getWQParamUnits*. The first of them (*getStationsWithWQParameter*) is actually a special case of the Search service outlined later in this section (6). The current *getStationsWithWQParameter* will be converted into a method in a more generic search web service, capable of handling different types of spatial selection (bounding box, by state_county, by hydrologic units) and multiple attribute filters.

   Of the remaining services, two operate on metadata for a given station: *getDischargeInfo(stationID)* is a prototype service for a *getMeasuredParameterInfo (parameterName, stationID)* method in a *MeasuredParameterInfo* service, while the *GetStationInfo* would be the main method in a *StationInfo* service. The remaining two services (*getWQParamName*, *getWQParamUnits*) really belong to the group of ontology services (7).

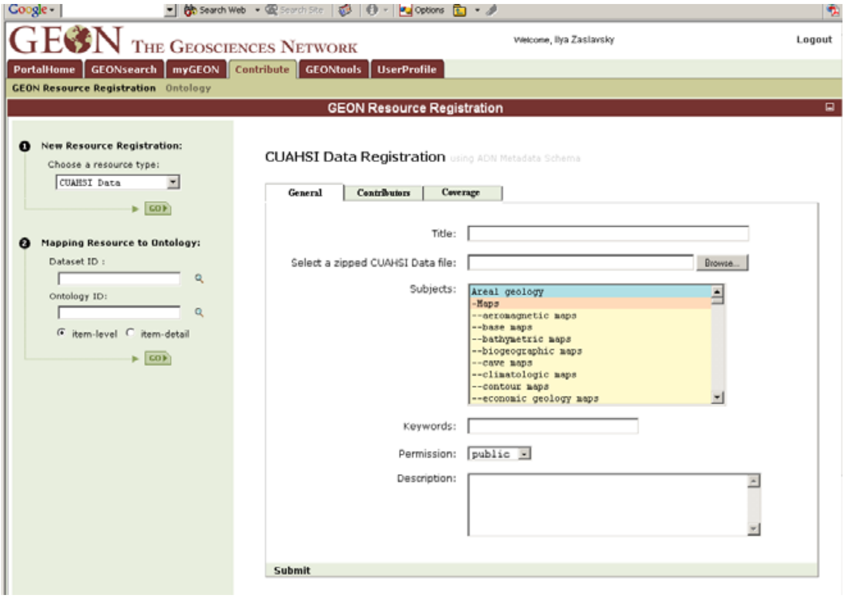
   This review of data access services is far from complete. There is an ongoing discussion between members of CUAHSI HIS team, and data managers at USGS and EPA, about development of the most robust set of data access web services, to be eventually installed at the agencies.

3. Sensor management services; monitoring and managing sensors, and implementing data transmission policies. We don't consider them in detail here, since such services have been in the center of other projects. For example, services implemented in Antelope real-time monitoring system (the ROADnet, LOOKING and related projects) include: *getAllProcesses; getAllOnProcesses; getStatus4Process; turnOnAProcess; turnOffAProcess; restartAProcess; addANewProcess; removeAProcess; getPfFile.*

4. Sensor data filtering: services for aggregating/filtering sensor data, depositing them in digital library and updating corresponding metadata records. Like the previous group, this set of services will become one of the foci of hydrology digital observatory development during its next phase.

5. Digital library services layer: services managing harvesting, searching, retrieving data under control of CUAHSI data nodes. The following web services are operational: *FindADO* (which searches CUAHSI metadata; this service must be updated to rely on ISO/CUAHSI metadata profile); *getADOurl*, and *transferADOs* (list at http://cuahsi001.sdsc.edu:8080/axis/servlet/AxisServlet). A recently developed service is *ADOtoImageService*, which converts spatial data found in an ADO (Arbitrary Digital Object, a compressed collection of data files and documents of different types), into an ArcIMS image service. The following services need to be additionally developed: *getMetadata* (ADOurl), *putMetadata*, and *putADO* – the latter two services to enable scenarios where ADOs are deposited from data retrieval applications (such as CUAHSI HIS data portal) or uploaded by users directly.

   Describing these services, we note that the concept of digital library in CUAHSI has evolved from being a central data gateway, to serving as a repository hosting data from CUAHSI partners, and ensuring data provenance. A typical workflow involving the digital library is as follows: CUAHSI researcher uses the CUAHSI data portal or a desktop application to find and retrieve hydrologic time series data from one of external sources, and uses it for analysis and visualization. When the time series data become available at the portal or at the application, the user has an option to save the retrieved data in the digital library to ensure that a repeated analysis can be performed on the same data set. In this process, the portal or desktop application calls registration/upload web services to upload the data to the hosted repository and create a respective record in CUAHSI metadata catalog. Another digital library scenario involves a digital library curator or power user uploading and registering CUAHSI-hosted datasets manually or in bulk. This scenario is now being explored at the CUAHSI Illinois site (P. Kumar, B. Ruddell) where significant local hydrology data resources have been assembled and registered to the digital library, described using CUAHSI/ISO metadata profile developed at Drexel University (M. Piasecki, B.Boran) and arranged in GEON-like data management infrastructure developed at SDSC (C. Baru, K. Lin, V. Nandigam, G. Memon). A snapshot of CUAHSI data registration portal (implemented currently as part of GEON portal) is in Figure 5. The data search and registration modules are now being migrated to CUAHSI HIS Data Portal described in section 2.3. Also, the web services developed initially for the ADO-based digital library have been merged with GEON data search and registration services, while HDF5 format is being explored as a potential standard container for heterogeneous set of documents. This completes CUAHSI digital library transition toward an architecture that reflects the digital hydrologic observatory vision.



GEON Resource Registration

1 New Resource Registration:
Choose a resource type: CUAHSI Data

2 Mapping Resource to Ontology:
Dataset ID :
Ontology ID:
● item-level ○ item-detail

CUAHSI Data Registration using ADN Metadata Schema

Title:
Select a zipped CUAHSI Data file:
Subjects: Areal geology
-Maps
--aeromagnetic maps
--base maps
--bathymetric maps
--biogeographic maps
--cave maps
--climatologic maps
--contour maps
--economic geology maps
Keywords:
Permission: public
Description:

Figure 5. The current system for registering CUAHSI data sets.

6. Services for querying both external data registry and digital library metadata; searching and retrieving data from them, with or without serialization (abstracted from individual source schemas). The *search service*, is

```
search (
        sources={Ameriflux|NWIS|Reanalysis|Storet|DL, or a group of those},
        attributeFilter={list of categories from a controlled vocabulary},
        type_of_geo={state_county|bbox|hydro_unit},
        geoFilter={state_county_FIPS, or 4 numbers of bbox in lat/lon,
        or hydro_unit_code},
        type_of_time={local|GMT},
        timeFilter={startDate, endDate}
        )
```

The service spawns search requests against external data sources (data resource drivers described in (2), the *getStationsWithWQParameter* service in particular) and against the digital library (the *FindADO* service mentioned in (5)), and returns stationIDs for each source. These stationIDs become the input for the *DataRetrieve* service, which populates a series of TimeSeries objects and returns either a data string, or an associated graphic object (chart, map), or a reference to an object.

7. Ontology services are used for registering heterogeneous data sources, rewriting queries from global to individual ontologies, and data integration. Several of these services have been explored in the course of GEON project, which now allows users to annotate datasets with an existing or user-defined ontology, create and edit ontologies, and formulate selection requests in terms of a selected ontology. In particular, ontology handling methods are implemented in the *DataRegistration*

and *ImageQuery* services in GEON, which are accessible from the GEON portal (www.geongrid.org).

8. Application level ("digital watershed") services. This is a large group of services that support managing *digital watershed* objects, including querying, publication as services, and instantiation of as warehouses. A variety of services support manipulation of digital watershed objects including:
    a. Data analysis services, including statistical analysis (with R server). Some work on incorporating R server in a workflow system was done within the Kepler project at SDSC.
    b. Workflow orchestration services (e.g. Model Builder, Kepler, D2K, etc.)
    c. Model wrappers (e.g. for Modflow and other models).
    d. Spatial analysis, e.g. polygon overlay, buffers, map algebra services (with ArcGIS Server). A web service wrapper for Map Algebra has been developed at SDSC. Wrapping GRASS modules for inclusion in the Kepler workflow system is available as a result of collaboration between UCSB and SDSC.
    e. Map integration services. These services, critical for digital watershed construction, have been partly developed within GEON. In particular, these services manage query-based integration of ArcIMS, shapefile, WMS/WFS services and data as long as they are registered the GEON registration system. Services in this group manage conversion of different types of data into a set of compatible formats, development of an integrated map legend, data shipment across nodes, generation of a composite map configuration file, and its instantiation as a map service.
    f. Presentation services: map generation; chart generation, 3D watershed visualization, animation, etc. Some of this functionality is now found within the resource drivers services (e.g. *getDailyStreamFlowChart).* While more efficient, it would be preferable to logically separate data retrieval from presentation in this context.
    g. Format conversion; projection; unit conversion services. Services of this group have been developed within multiple projects. Within GEON, there are services for converting XML and ASCII files to shapefiles and then to mapping services. At UCSB, there are web service wrappers for GDAL (developed within SEEK). Implementing conversion of this type, CUAHSI HIS may rely on ArcGIS server or FME.

9. Authentication/authorization services. These services are managed through the CUAHSI HIS portal described below, and support personalization of user experience, including the ability to create and manage personal research space within CUAHSI (myCUAHSI). This would allow users to manipulate custom versions of higher-level objects, register datasets for personal or group use, share their resources with authorized users, etc.

10. Service registry, with functionality allowing registration, description, search, and retrieval of services. SDSC is working on developing a searchable registry of web services, with the prototype available at water.sdsc.edu/uddi/

2.3 CUAHSI Hydrology Data Portal

CUAHSI HIS Web data portal is one of user interfaces for accessing CUAHSI resources. Its goal is to provide a uniform view over multiple concurrent project efforts, facilitate access to federal and CUAHSI-hosted hydrologic data, provide preliminary data analysis and visualization tools and organize the multiple resources into executable workflows. Other data management and analysis environments from which access to hydrologic resources is desirable (according to the CUAHSI user survey) include MS Excel, ArcGIS and Matlab desktop applications, as well as programming languages including Fortran, VB, C and Java.

Data sources currently under consideration that provide tabular hydrologic information at point locations include:

- **USGS National Water Information System (NWIS) –** streamflow, water quality and groundwater levels; http://waterdata.usgs.gov/nwis
- **National Climate Data Center (NCDC) Climate Data Online (CDO) –** precipitation, temperature, and other climate variables; http://cdo.ncdc.noaa.gov/CDO/cdo
- **EPA Storet –** water quality; http://www.epa.gov/STORET/
- **USGS National Water Quality Assessment (NAWQA) –** hydrology, water quality and biology in NAWQA study units; see “Data” at http://water.usgs.gov/nawqa/
- **Ameriflux –** land-atmosphere flux data from flux towers; http://public.ornl.gov/ameriflux/
- **Long Term Ecological Research (LTER) --** climate, hydrology and ecology data; http://www.lternet.edu/data/

Other data sources that provide gridded weather and climate information include:

- **North American Regional Reanalysis of climate (NARR) –** land-atmosphere fluxes and atmospheric conditions from weather model reanalysis; http://www.emc.ncep.noaa.gov/mmb/rreanl/
- **NCAR Community Data Portal –** weather and climate datasets from NCAR research (especially VEMAP); https://cdp.ucar.edu/

Based on CUAHSI Hydrologic Information System user assessment, the order of importance of these data sources is:

1. USGS Streamflow
2. NCDC Precipitation
3. Other NCDC weather and climate information

4. EPA Storet water quality
5. USGS NAWQA data
6. USGS groundwater data

EPA STORET is, in particular, regarded as a valuable data source that is difficult to access at present.

At present, mechanisms are developed to access USGS NWIS and NAWQA repositories, EPA STORET, and Ameriflux. The NWIS access is already facilitated through the data portal as described below. Methods of accessing NAWQA and Ameriflux are similar (the difference is in controlled vocabularies used by each system). In order to access the EPA STORET repository, SDSC developed an EPA STORET web wrapper. However, the common challenge of web site wrapping is that changes in site layout or underlying database (often unpredictable) would require modifications in web wrapper code. A better solution is collaboration with the EPA STORET development team to establish web service-based access to the repository. This direction is being pursued now.

A conceptual outline of the data portal, or *cyberdashboard* as it is being referred to in the first architecture diagram, is shown in Figure 6. The portal is being implemented as a group of three main portal applications: metadata search, data analysis/visualization, and control center (where authorized users can curate CUAHSI data holdings, manage grid resources, upload and register datasets, eventually control sensor network, etc.).

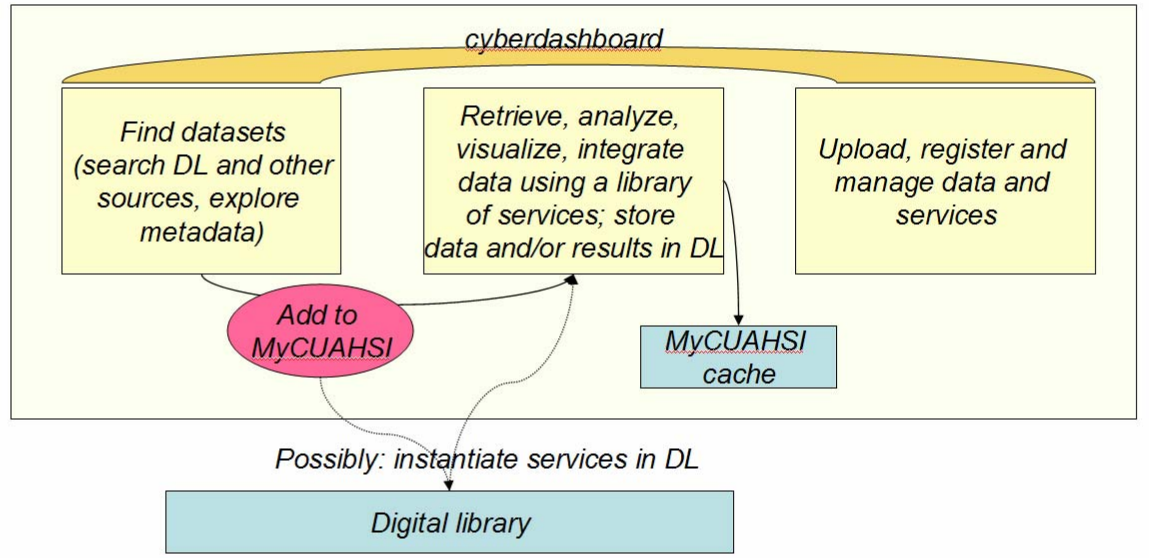
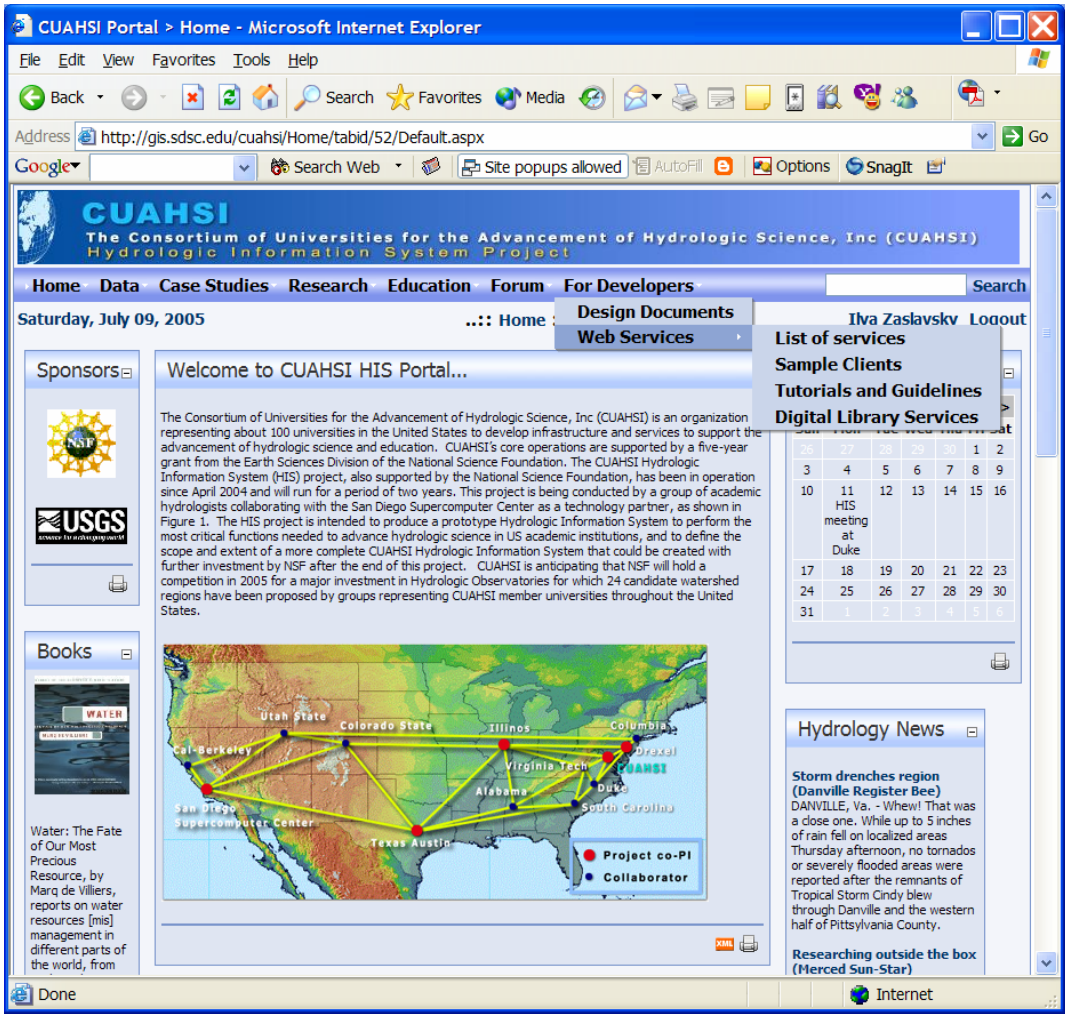

cyberdashboard
Find datasets (search DL and other sources, explore metadata)
Retrieve, analyze, visualize, integrate data using a library of services; store data and/or results in DL
Upload, register and manage data and services
Add to MyCUAHSI
MyCUAHSI cache
Possibly: instantiate services in DL
Digital library

Figure 6. Conceptual organization of CUAHSI portal.

The current prototype implementation relies on a common open-source portal framework called DotNetNuke at www.dotnetnuke.com (Figure 7). This is a popular framework, with nearly 200,000 registered participants, active development community, and a large number of various modules available free of charge. The CUAHSI HIS portal is built from a large number of modules managing user access to hydrology resources and linking to various CUAHSI applications. In particular, the main part of the portal is the gateway to federal and CUAHSI-hosted data resources available under Data – Data Portal part of the menu. (Figure 7). This application, however, is not tied to the particular DotNetNuke development framework; recently it has been also included in the CLEANER Collaborative portal based on a different technology (Liveray, Java portlets).



Welcome to CUAHSI HIS Portal...

The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) is an organization representing about 100 universities in the United States to develop infrastructure and services to support the advancement of hydrologic science and education. CUAHSI's core operations are supported by a five-year grant from the Earth Sciences Division of the National Science Foundation. The CUAHSI Hydrologic Information System (HIS) project, also supported by the National Science Foundation, has been in operation since April 2004 and will run for a period of two years. This project is being conducted by a group of academic hydrologists collaborating with the San Diego Supercomputer Center as a technology partner, as shown in Figure 1. The HIS project is intended to produce a prototype Hydrologic Information System to perform the most critical functions needed to advance hydrologic science in US academic institutions, and to define the scope and extent of a more complete CUAHSI Hydrologic Information System that could be created with further investment by NSF after the end of this project. CUAHSI is anticipating that NSF will hold a competition in 2005 for a major investment in Hydrologic Observatories for which 24 candidate watershed regions have been proposed by groups representing CUAHSI member universities throughout the United States.

Water: The Fate of Our Most Precious Resource, by Marq de Villiers, reports on water resources [mis] management in different parts of the world, from

26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11 HIS meeting at Duke	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Hydrology News

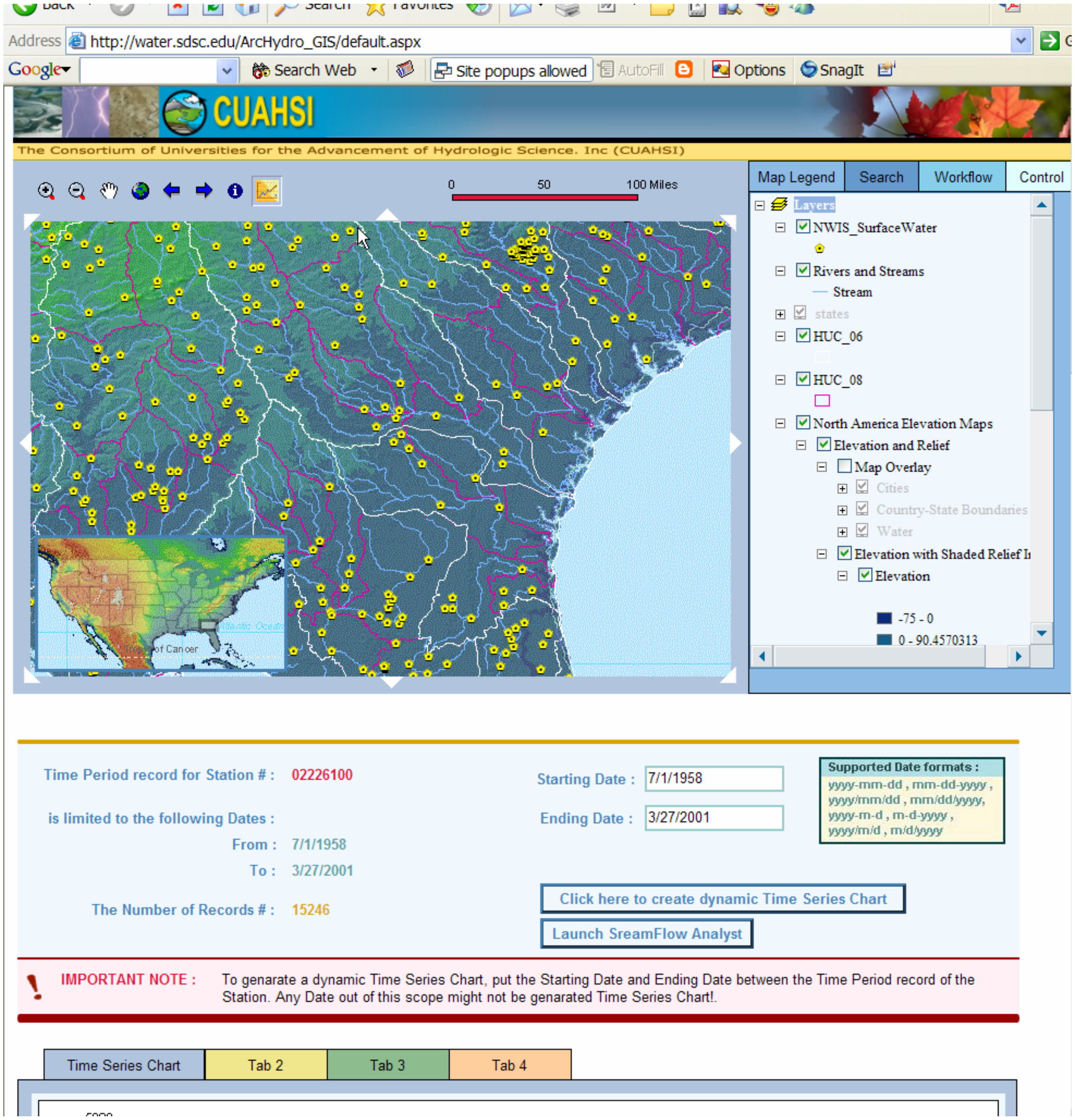
Storm drenches region (Danville Register Bee)
DANVILLE, Va. - Whew! That was a close one. While up to 5 inches of rain fell on localized areas Thursday afternoon, no tornados or severely flooded areas were reported after the remnants of Tropical Storm Cindy blew through Danville and the western half of Pittsylvania County.

Researching outside the box (Merced Sun-Star)

Figure 7: The current prototype implementation of the CUAHSI HIS portal

Opening the data portal, users can see a map of the US, zoom in to area of interest, and select one or several stations to explore. Currently the list of stations contains NWIS stations only, but stations for other federal repositories will be added soon. Additional tabs in the right top corner of the interface let users switch between search, workflow, and "control/contribute" components of the portal, as in Figure 8. Once a station of interest is identified, the portal calls data access web services to retrieve station metadata and create a time series chart, either by launching portal's charting services or calling the USU StreamFlow Analyst charting tool.

The data portal is created using ESRI's ArcGIS Server which appears the most flexible and powerful web mapping application development platform, as it streamlines development of web services based on the extensive ArcObjects library.


Address http://water.sdsc.edu/ArcHydro_GIS/default.aspx
Google Search Web Site popups allowed AutoFill Options SnagIt
CUAHSI
The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI)
0 50 100 Miles
Map Legend
Search
Workflow
Control
Layers
NWIS_SurfaceWater
Rivers and Streams
Stream
states
HUC_06
HUC_08
North America Elevation Maps
Elevation and Relief
Map Overlay
Cities
Country-State Boundaries
Water
Elevation with Shaded Relief I
Elevation
-75 - 0
0 - 90.4570313
Click here to create dynamic Time Series Chart
Launch SreamFlow Analyst
Time Series Chart
Tab 2
Tab 3
Tab 4

Time Period record for Station # : 02226100

is limited to the following Dates :

From : 7/1/1958

To : 3/27/2001

The Number of Records # : 15246

Starting Date : 7/1/1958

Ending Date : 3/27/2001

Supported Date formats : yyyy-mm-dd , mm-dd-yyyy , yyyy/mm/dd , mm/dd/yyyy, yyyy-m-d , m-d-yyyy , yyyy/m/d , m/d/yyyy

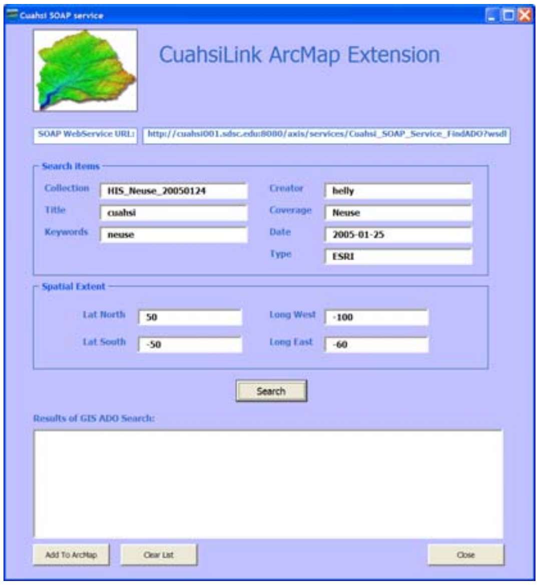
IMPORTANT NOTE : To genarate a dynamic Time Series Chart, put the Starting Date and Ending Date between the Time Period record of the Station. Any Date out of this scope might not be genarated Time Series Chart!.

Figure 8. CUAHSI HIS Data Portal snapshot

2.4 CuahsiLink

CUAHSI user needs survey described in Chapter 3, identified ArcGIS, Excel and Matlab as the desktop applications most commonly used the community. The CuahsiLink ArcGIS extension application was developed to let users access data stored in digital library from ArcGIS desktop environment. A snapshot of the application is shown in Figure 9. Upon launching CuahsiLink the user can select an area of interest and enter additional selection filters. The system would invoke the *FindADO* web service to discover digital objects that satisfy the search criteria, and prompt the user to select objects for downloading. In the next step, the compressed objects would be downloaded

onto local machine’s staging area, with spatial data being extracted from the archives and loaded into ArcGIS. This small application significantly eased user access to digital library, while utilizing the same set of web services as the data portal.
Figure 9. CUAHSILink application



SOAP WebService URL: http://cuahsi001.sdsc.edu:8080/axis/services/Cuahsi_SOAP_Service_FindADO?wsdl

Search Items

Collection: HIS_Neuse_20050124
Title: cuahsi
Keywords: neuse
Creator: helly
Coverage: Neuse
Date: 2005-01-25
Type: ESRI

Spatial Extent

Lat North: 50
Lat South: -50
Long West: -100
Long East: -60

3. Formal representation of CUAHSI higher-level objects

3.1 Digital watersheds as central to CUAHSI HIS cyberinfrastructure

As we discussed before, *digital watershed* is one of the central concepts in the CUAHSI HIS project. Beyond being a research concept, it plays a critical *infrastructure role* as a bridge between the project's data retrieval and management modules, on one side, and analysis, modeling and visualization components, on the other side. This central role is manifested in several fundamental CUAHSI workflows. For example, modeling water quality, patterns of water supply and demand, analysis of local ecosystem, delineation of impaired water bodies require compiling a canonical set of data layers which may reside in different databases located at different servers. The data layers may reflect both natural and social environments, and cover area within or beyond the boundaries of a given watershed. Such a higher-level digital observatory object (figure 2) comprised of a canonical set of linked data layers that reflect comprehensive nature of a watershed and support a set of watershed modeling workflows, is being referred to as *digital watershed*.

3.2 Rationale for conceptually representing digital watersheds in CUAHSI HIS

Development of a robust digital representation of watersheds is one of CUAHSI research goals. Such a robust representation implies:

a) ability to present a watershed using a standard platform- and software-independent template, and following mainstream computing standards, to ensure that the watershed representation is both computer and human-readable, and can be easily parsed by a variety of existing codes,
b) ability to convert the digital watershed representation into various standard or vendor-specific documents or services including, ESRI geodatabases, ArcIMS services, OGC map services, WSDL web services, SVG (W3C's Scalable Vector Graphics spec), XAML, etc.,
c) ability to express how digital watershed integrates different types of data objects, using various spatio-temporal or attribute join models,
d) ability to instantiate a particular digital watershed from available remote data, where data may come from federal or state data repositories, include various project datasets, etc.,
e) ability to query digital watershed representations, e.g. to identify data gaps when compared to a canonical watershed description, or check the referenced data for consistency (projections, formats, temporal reference, etc.)
f) ease of integration with other emerging digital representations being developed in related projects, including digital estuary, etc.,
g) compatibility with emerging data management cyberinfrastructure, including ontology handling, reliance on web/grid services, XML representations of source schemas and capabilities via a collection of XML wrappers, etc.,
h) ease of update as new knowledge or data sources become available

An interesting characteristic of digital watershed is the multitude of integration models that it can express. An obvious model that formally holds digital watershed elements together is co-location in space, where boundaries of most data layers are defined by watershed boundaries, which are derived from analysis of DEM and hydrologic networks. In other words, the watershed boundary, or some function of it (e.g. a distance buffer), is used as a "cookie-cutter" when populating watershed data layers from external sources.

However, other important components of digital watershed are not bounded by its surface boundaries. These include atmospheric parameters, groundwater flows, underlying geology, as well as demographic and economic variables and processes that not necessarily coincide with natural boundaries. For them other types of integration models are needed based on functional relationships between watershed parameters. As a further example of the type of integration, watershed characterization shall point to upstream and downstream watersheds or water bodies. Ideally, a watershed representation shall explicitly outline the types of joins between different watershed elements, to make automatic instantiation and update of digital watersheds possible.

Following these desired features, we propose the following formal representation of a digital watershed.

3.3 Digital watershed as a system of declarative integrated views, and its XQuery formalization

Declarative integrated views is a common mechanism for expressing patterns of information integration across multiple distributed resources. XQuery is a W3C standard XML query language for querying XML documents and specifying integrated reusable system-independent views over data distributed across the enterprise, and is widely used in this role in industry. Here, we outline how XQuery-based formalisms can be used to describe data integration within a digital watershed, and outline services for processing such views.

Following the discussion of different integration models to be expressed by digital watershed template, we propose to specify digital watershed as a collection of *spatio-temporal integrated views*, and *mediator integrated views* (S-views and M-views, as they are referred to below).

A sample XQuery specification of a digital watershed as an S-View is shown in Figure 10. The focus here is not on a complete digital watershed specification, but rather on general outline and examples of different integrated views.

An S-view abstraction represents a collection of distributed mapping services and views, which together form a spatial database (or a map) with a given initial spatial extent, projection, units, etc. S-views reference one or more M-views that describe map layers as well as valid queries against services declared within the S-view, in addition to outlining cartographically-meaningful layer ordering and including abstractions for various standard map components. The output of an XQuery is an XML fragment which represents a composite map configuration document specifying a sequence of spatial layers to be included in digital watershed, and a collection of queries/functions to be exposed to the user as part of map interface.

Figures 11-12 provide examples of M-views referenced from the S-view in Figure 8.

Note in Figure 10, that XQuery has a query header part and an output specification part. The header part lists watershed parameters (name, bounding box, temporal bounds, and various external parameters for watershed functions referenced in the output specification. The output part is a skeleton of an XML document where different layers are represented as views over one or several remote resources.

```
Declare Function digital_watershed
($name as xs:string, $coordsys as xs:int,
$minx as xs:double, $miny as xs:double,
$maxx as xs:double, $maxy as xs:double
$startdate as xs:date, $enddate as xs:date
$discharge-parameter as xs:double) As element() {

let $env := envelope($minx,$miny,$maxx,$maxy) cast as ogc:polygon
let $period := period($startdate, $enddate) cast as time:period
return
<Sview>
  <name> {$name }</name>
  <projection>{ $coordsys }</projection>
  <envelope>$minx, $miny, $maxx, $maxy </envelope>
  <period>$startdate, $enddate</period>
  <layers>
  <group id=”baselayers” status=”core”>
    <mview>
    {
       for $DEM in source("basemap")//DEM
       where overlap(projection($DEM/Shape,$coordsys), $env ) = 1 AND
              inside ($DEM/Timestamp, $period) = 1
       return
       <source>{ $ocean }</source>
    }
    </mview>
    <mview>
    {
       for $hydrology in source(“usgs_hydro”)//river
       where overlap (projection($hydrology/Shape,$coordsys), $env) = 1 AND
              inside ($hydrology/Timestamp, $period) = 1
       return
       <source>{ $hydrology }</source>
    }
    </mview>
  </group>
  <group id=”streamflow” status=”core”>
  <mview>
  {
     for $NWIS_Stations in NWIS_Stations_with_data($discharge-parameter,$startdate, $enddate)
     where overlap (projection($NWIS_Stations/Shape,$coordsys), $env) = 1
     return
     <source>{ $NWIS_Stations }</source>
  }
  </mview>
  </group>
  <group id=”atmospheric” status=”core”>
. . .
  </group>
</layers>
 <mviews>
  <mview>
  {
     for $watersheds in  watersheds_upstream($name)
     where “”
     return
     <source>{ $watersheds }</source>
   }
   </mview>
</m_views>
</Sview>
}
```

Figure 10. A sample XQuery specification of a digital watershed

Further, notice that the watershed output specification has two groups of views. Views that compose a "map" are grouped into the "layers" group. Other datasets and integrated views that are not critical to have on a map proper (or a geodatabase underneath) but essential for certain watershed-related analysis/modeling workflows, are included in a separate <mviews> group. In turn, each layer, be it a map layer or just an essential data component, is included into a named <group>. The groups reflect geology, soils, streamflow, water quality, atmospheric, socio-economic, etc. thematic categories, both core and auxiliary (precise list and content of these categories to be determined).

The flexibility in defining the shape of query output that XQuery offers, is very important for describing spatial data collections generated from distributed sources. I feel that this advantage justifies the use of XQuery for defining watersheds as compound documents, despite XQuery's relative novelty compared with more traditional ways of managing integrated views.

The M-views can simply reference a single data service, or integrate over several of them, based on spatial or attribute joins. For example, Figure 9 shows a parameterized query "find NWIS stations that have data on *$parameter* between *$startdate* and *$enddate,* which is included as a thematic layer in the map. We assume here that time series data obtained from USGS or EPA on such queries (implemented as Web services in CUAHSI) are an integral part of digital watershed description. Note the "*ontology part-of*" operation that wraps the "*$parameter*" requested from the NWIS system. This construct handles an ontology of water characteristics measured at USGS stations stored as an OWL (Web Ontology Language) file, and expands a given "parameter" to include its child concepts: for example, querying "nutrient" data would expand to include phosphorus, carbon, sulfur, iron, etc.; phosphorus, in turn, would expand to dissolved and particulate, etc.

```
Declare Function NWIS_Stations_with_data
($parameter as xs:string, $startdate as xs:date, $enddate as xs:date) as element() {
        for
$station IN source("NWIS")//station
        where
Ontology:part-of($station/parameter, $parameter) AND
$station/startdate > $startdate AND
$station/enddate < $enddate

return $station
}
```

Figure 11. An M-view over NWIS stations delivering a particular time-series extract.

Alternately, an M-view may be translated not into a map layer but into a parameterized query exposed on the client interface (as, for example, a pointer to upstream watersheds - Figure 12).

```
Declare Function watersheds_upstream
($thisWatershed as xs:string) AS element() {
            for
$watershed IN source("watersheds")//watershed,
            where
upstream($watershed/Shape, $thisWatershed) = 1
return $watershed
}
```

Figure 12. A sample M-view declaring a parameterized query “find watersheds upstream of the current watershed”. Note that this M-view is not a “map layer” here.

Once M-views are built, they can be queried as all other data sources, and selection queries against them are no more difficult to formulate than queries against primary XML sources.

Given a digital watershed integrated view expressed as above, CUAHSI services shall be able to a) analyze if/how the watershed can be instantiated with available data, and identify data gaps; b) convert this representation into a geodatabase (materializing all or selected views), or to a software-specific configuration file. Such possible outputs would include ArcIMS image service configuration file, SVG presentation, or Microsoft XAML document. For different types of output, users shall be able to control the degree to which digital watershed is materialized: from symbolic representations (such as ArcIMS service configuration file accompanied by placing referenced data on a local staging area) to digital watersheds more fully materialized as geodatabases. As discussed below, some services developed in GEON and related SDSC projects, with little modification would support generation of ArcIMS or SVG documents from such XQuery descriptions.

Various implementations of integrated views in mediators have been discussed in the literature; while handling spatial views in XQuery and converting integrated views to compound map documents seems to be new. However, several previous efforts shall make the implementation task easier.

One potential building block is provided by map assembly services in GEON. The purpose of these services is to generate composite maps on the fly from multiple distributed resources. The services include a range of conversion routines that handle retrieving spatial data fragments from remote databases to a local staging area, transforming spatial data from various formats into formats supported by ArcIMS, generating an ArcIMS configuration file, and instantiating an ArcIMS image map service. This code can be re-used when converting digital watershed representation to an ArcIMS service.

Another component is provided by our previous work on automatic generation of SVG maps from remote sources, where SVG documents were generated by retrieving and re-styling feature geometry from ArcIMS sources. Since many database vendors support

SVG output, generating SVG fragments from other databases shall not be difficult. In a similar way, we shall be able to generate XAML documents (the Microsoft's presentation format in Longhorn), for rendering on a range of smart clients.

Many components of the system outlined in this chapter are already constructed, within CUAHSI or related projects. In particular, CUAHSI HIS has developed functional prototypes of services for retrieving time series data from federal sources, populating and searching digital library, visualizing and analyzing time series data and maps. In our opinion, most promising areas of further research and development include construction of CUAHSI object hierarchies and developing formal models for representing and manipulating core CUAHSI HIS objects, most importantly the digital watershed with associated workflow and data integration services. Additional research directions of critical importance for CUAHSI HIS system architecture include formal metadata and knowledge-based annotation and querying of digital watershed objects and spatial data layers, and a detailed investigation of efficient data structures and services supporting hydrologic observatory and providing advanced spatio-temporal analysis, visualization and modeling capabilities. This effort would contribute to a better understanding of digital observatory and its structural components, and create a working prototype of a digital hydrologic observatory.

Chapter 4

User Needs Assessment

Christina J. Bandaragoda and David G. Tarboton,
Dept of Civil Engineering, Utah State University, Logan UT

David R. Maidment
Center for Research in Water Resources
University of Texas at Austin

Introduction

This chapter reports on a data collection effort targeting the Hydrologic Information System (HIS) User Community: who they are, what they do, and how they do it. Here we present the results of a web-based survey of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) members and their affiliates which has clarified important HIS development issues, informed HIS project decision-making and will help create an effective, efficient, and functional HIS. CUAHSI is an organization representing more than 100 universities, sponsored by the National Science Foundation to develop infrastructure and services for the advancement of hydrologic science and education in the United States. The CUAHSI Hydrologic Information System (HIS) project is a component of CUAHSI's mission that is intended to improve infrastructure and services for hydrologic information acquisition and analysis. You can learn more about CUAHSI from the website www.cuahsi.org.

We have learned from the survey that there is a definitive, quantifiable need for a Hydrologic Information System. Most researchers spend a significant amount of time preprocessing data for their research and believe an information system such as the CUAHSI HIS would be helpful and relevant to their work. Data services are the most important services for the HIS to provide while addressing critical data use difficulties such as inconsistent data formats, the existence and consistency of metadata, and irregular timesteps.

The overall research objective of the HIS User Needs Assessment process was to assess how hydrologic information is used in research and to assess what functions are of greatest importance among the services that CUAHSI HIS may provide. The information collection focused on three main goals.

- Define the hydrologic data user community.
- Collect raw data on data use patterns, preferences, issues relevant to key decision points in HIS development.
- Prioritize future HIS developments.

The HIS User Assessment survey process included three main steps: preliminary information gathering, a pilot survey, and a web survey. The process began with collaborators to the HIS project team gathering preliminary information from their institutions. This preliminary information was presented at the HIS Symposium at

Austin, Texas, March 2005. At the symposium, a pilot paper survey was conducted using feedback from the preliminary information gathering efforts. The results of both the information gathering and pilot survey were then used to develop the web survey that was conducted in May 2005. This chapter primarily reports results from the web survey. Results from the information gathering and preliminary surveys are included as appendix 4.

Sampling Method

The web-based survey was developed to guide development of the CUAHSI HIS and to assess how hydrologic information is used in research and what functions are of greatest importance among the services that the HIS may provide. The hyperlink was emailed to approximately 100 CUAHSI contacts including representatives of CUAHSI member institutions or participants in CUAHSI sponsored projects or activities with the request that they take the survey as well as forward the request to others at their institution. The questionnaire used is given in Appendix 1. We received 76 responses from researchers at 39 different universities.

Results

Respondents

The HIS User Assessment respondents were approximately 40% hydrologists (surface and groundwater) with a notable majority of respondents identifying themselves as being from other disciplines including water resources, water quality/chemistry, environmental, ecology, atmosphere, GIS/spatial analysis, geomorphology, geology, statistics/mathematics, biology, social science and economics. (Figure 1). The current position of respondents was primarily University faculty (72%), but also included graduate students (20%), University professional/post-doc, working professionals, and others.

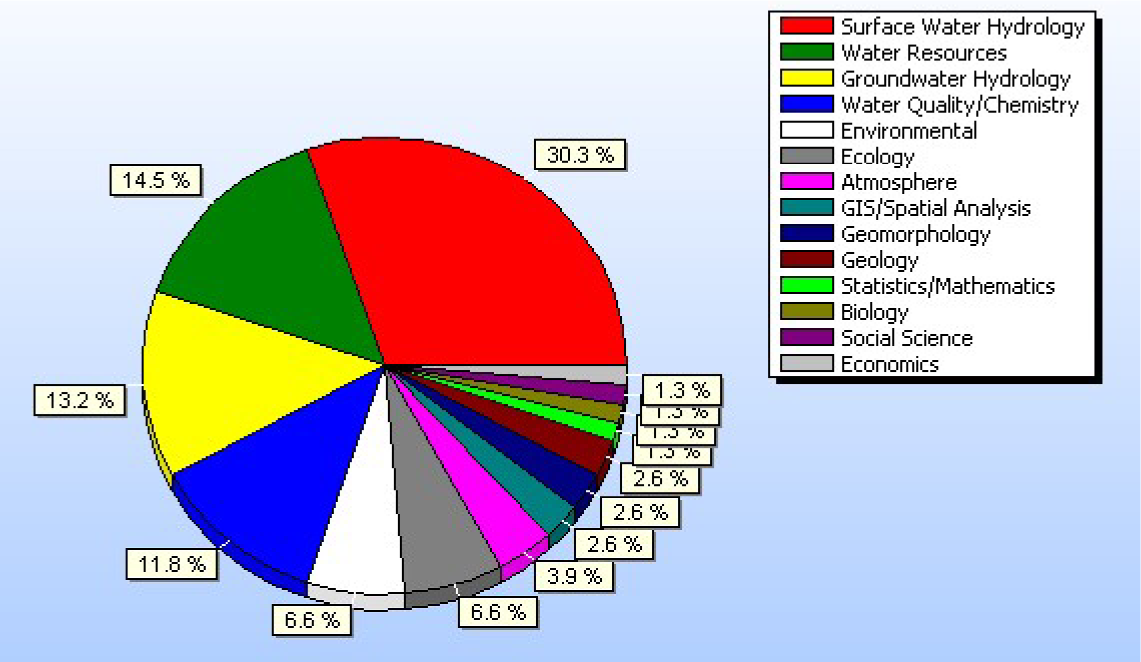


Figure 1. Distribution of respondent fields of research

Software Used for Hydrologic Research

The web survey began with questions related to software used in hydrologic research. Compatibility, inter operability and reliance on open source or professionally supported commercial systems are factors in the design of the HIS. Prior to this survey the distribution of operating system use in the HIS user community was unknown. The results show how many respondents use multiple operating systems (Figure 2). Findings show that 96% of respondents use the Microsoft Windows Operating systems for research, and 36% of respondents also use another operating system in addition to Windows. This shows that although nearly everyone uses Windows, a significant number of researchers also rely on other operating systems for their research.

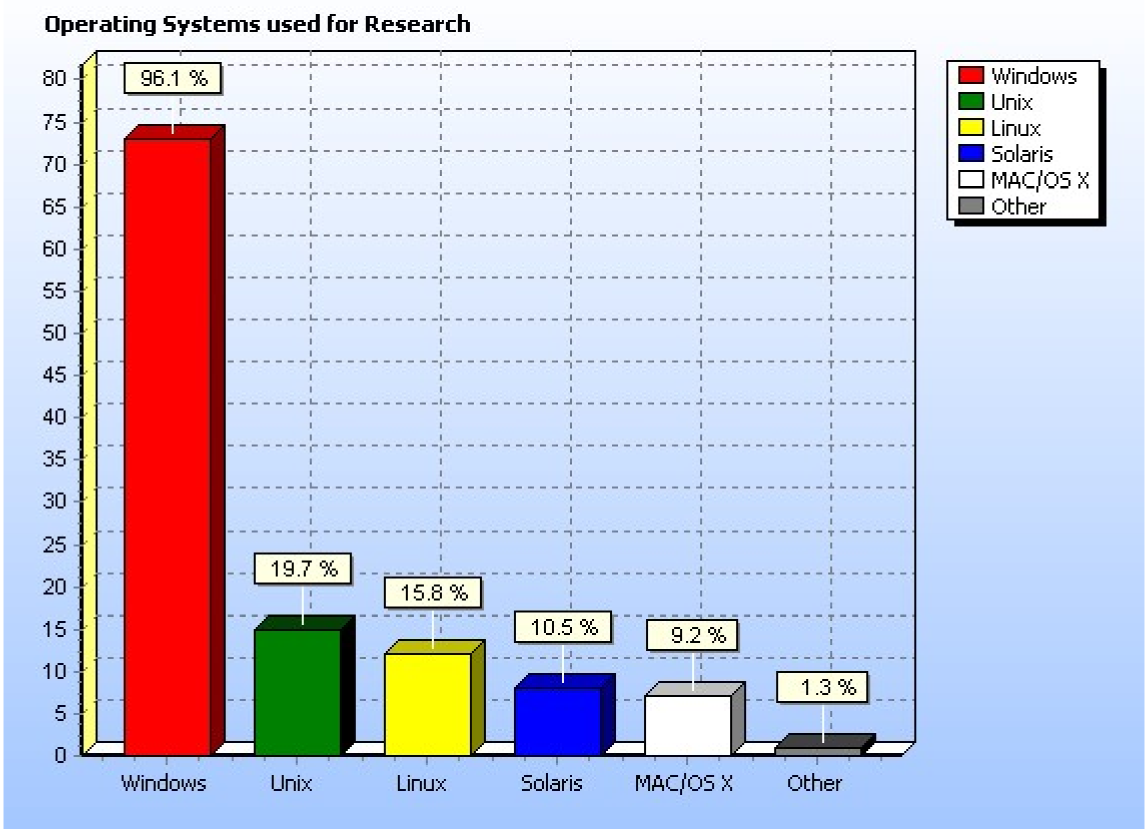


Figure 2. Distribution of operating systems used for hydrologic research. Respondents could indicate more than one operating system, resulting in percentages totaling more than 100%. 36% of respondents indicated one or more non Windows operating systems.

Software Used by category

The HIS may include the capability to interact with other software. Respondents were asked which software packages they use for hydrologic analysis in areas including programming, data management, programming for database client software, GIS (Geographic Information Systems), Mathematics/Statistics, and Hydrologic Models. Our question allowed respondents to select a first and second choice from lists of software programs specific to each category in order to determine which software is the most important among the numerous software programs available. The complete results and weighted average ranking of software programs is provided in Tables 1 through Table 6. Respondents were offered two drop down lists from which to select their first and second choices. Weighted averages were calculated simply with the first choice having twice the weight of the second choice.

Programming

FORTRAN is the most popular programming language used for research, followed by C/C++ and Visual Basic (VB). 85% of respondents indicated a programming language used in their research with 15% indicating programming is 'not applicable to my research'. Considering the first choice in programming languages, FORTRAN is twice as popular as C/C++ or VB. However, for the weighted average of first and second choices, the percentage of users selecting FORTRAN, C/C++ and VB is much closer. Java, Python, AWK and PERL are used by relatively few of the respondents.

Table 1. Programming languages for hydrologic research

Rank	Software	1st Choice	2nd Choice	Weighted Average
1	FORTRAN	42.1%	18.6%	34.3%
2	C/C++	19.7%	27.1%	22.2%
3	Visual Basic	18.4%	23.7%	20.2%
4	not applicable to my research	15.8%	15.3%	15.6%
5	Java	2.6%	5.1%	3.4%
6	Python	1.3%	5.1%	2.6%
7	AWK	0.0%	3.4%	1.1%
8	PERL	0.0%	1.7%	0.6%

Data Management

Microsoft Excel is the most popular software for managing data, followed by Microsoft Access. At least 93% of respondents indicated data management software that they use for their research with between 5% and 7% indicating ‘not applicable to my research’. Almost 70% of respondents use Microsoft Excel as their first choice for managing data. This could be due to the simplicity of using Excel for the relatively small datasets common in hydrology. Less than half of the respondents program database client software to access data, but when they do, Visual Basic is primarily used (Table 3).

Table 2. Data Management software for hydrologic research

Rank	Software	1st Choice	2nd Choice	Weighted Average
1	Excel	69.3%	18.2%	52.3%
2	MS Access	10.7%	58.2%	26.5%
3	SQL/Server	12.0%	16.4%	13.5%
4	not applicable to my research	5.3%	7.3%	6.0%
5	PostgreSQL	2.7%	0.0%	1.8%

Table 3. Programming languages used to access database client software. Only a single selection was permitted for this question.

Rank	Software	1st Choice
1	Not applicable to my research	50.8%
2	Visual Basic	23.1%
3	FORTRAN	9.2%
4	C/C++	9.2%
5	Other	3.1%
6	Java	1.5%
7	Perl	1.5%
8	Python	1.5%
9	Awk	0.0%

GIS

ArcGIS (ESRI ArcMap, ArcInfo, ArcView) dominates GIS software use (Table 4). 92% of respondents selected ArcGIS as their first choice with the highest ranking second choice receiving only 30% of second choice selections apart from the 43% who indicated that a second choice was not applicable to their research. Apparently, most respondents rely only on ArcGIS and reliance on other GIS software is rare.

Table 4. GIS software for hydrologic research

Rank	Software	1st Choice	2nd Choice
1	ArcGIS (ESRI ArcInfo, ArcView, etc)	92.1%	0.0%
2	not applicable to my research	6.6%	43.3%
3	IDRISI (Clark Labs)	1.3%	13.3%
4	MapInfo	0.0%	30.0%
5	GRASS	0.0%	13.3%
6	TAS	0.0%	0.0%

Mathematics/Statistics

Matlab is the most popular software for mathematics and statistics, followed by Microsoft Excel and SAS, but there is a wide variability in software used (Table 5). Mathematics/Statistics software programs are used by at least 97% of respondents (less than 3% reported that use of a software program in this category was not applicable to their research). Matlab is the first choice of 42% of respondents, Excel is the first choice of only 24% of respondents. However, the difference in the weighted average between Matlab and Excel is only 9%. If use of mathematics/statistics software were to be incorporated into the HIS, both Matlab and Excel would need to be accommodated.

Table 5. Mathematics/Statistics software for hydrologic research

Rank	Software	1st Choice	2nd Choice	Weighted Average
1	Matlab	41.3%	19.0%	33.9%
2	Excel	24.0%	25.4%	24.5%
3	SAS	10.7%	11.1%	10.8%
4	SPSS	5.3%	11.1%	7.2%
5	R (Open Source Splus)	2.7%	14.3%	6.6%
6	Mathematica	5.3%	6.3%	5.6%
7	Minitab	2.7%	4.8%	3.4%
8	IDL	2.7%	3.2%	2.9%
9	Splus	1.3%	3.2%	1.9%
10	not applicable to my research	2.7%	1.6%	2.3%
11	Scilab (Open Source Matlab)	1.3%	0.0%	0.9%

Hydrologic Models

80% of respondents indicated that they use hydrologic models in their research, however the models used vary widely. The most important result reported in Table 6 may be that 'not applicable to my research' was the highest ranking response for choice in hydrologic model. Modflow is the most popular groundwater model but there is no predominant surface water model. A general, simple, standard, and open interface that could connect with many systems would be the only way to accommodate all of the models used.

Table 6. Hydrologic Models used in hydrologic research

Rank	Software	1st Choice	2nd Choice	Weighted Average
1	not applicable to my research	21.9%	15.3%	19.7%
2	Modflow/Visual Modflow	19.2%	16.9%	18.4%
3	U.S. Army Corps HEC models	11.0%	10.2%	10.7%
4	GMS Groundwater Modeling System	8.2%	11.9%	9.4%
5	TOPMODEL	11.0%	8.5%	10.2%
6	Sacramento/NWS/HSPF	5.5%	8.5%	6.5%
7	SMS Surface Water Modeling System	2.7%	6.8%	4.1%
8	SHE System Hydrologique European/Mike-SHE	0.0%	8.5%	2.8%
9	Groundwater Vistas	4.1%	5.1%	4.4%
10	TIN-based real time Integrated Basin Simulator (tRIBS)	4.1%	1.7%	3.3%
11	EPA Basins	2.7%	5.1%	3.5%
12	WMS Watershed Modeling System	2.7%	0.0%	1.8%
13	SWAT	4.1%	0.0%	2.7%
14	MMS/PRMS	2.7%	1.7%	2.4%

The questionnaire included space for respondents to list other software packages that should be considered for interfacing with the CUAHSI HIS. The complete list of responses received is given in Appendix 2. These responses mention a total of 35 additional software packages.

An interesting result from our preliminary survey at the CUAHSI HIS Symposium at Austin, Texas, March 2005, came from the comparison of all software programs without restriction to categories. Respondents were asked to rate each software program between 1 and 5, where 1 is "never use or do not find useful" and 5 is "use frequently and find indispensable". Figure 3 presents these results which show that Excel and ArcGIS scored highest as the two most popular software programs for hydrologic research among the Symposium participants.

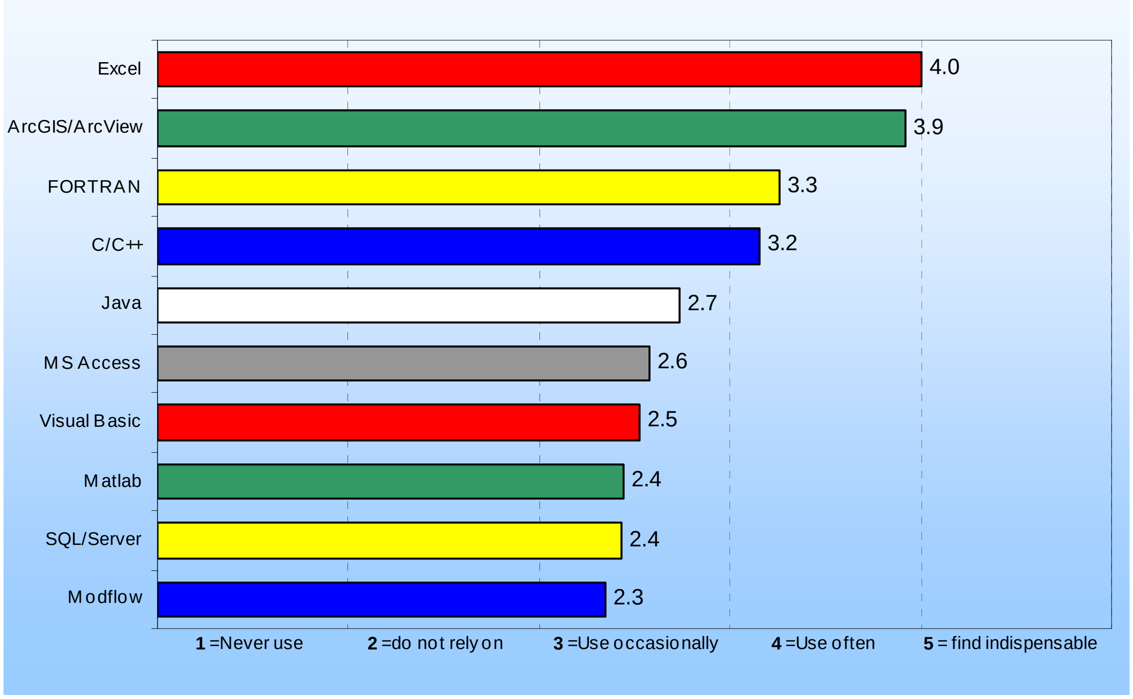


Figure 3. Rating of software packages and programming languages with respect to how important they are for hydrologic analysis. Results taken from preliminary survey at the CUAHSI HIS Symposium held in Austin, Texas, March 2005. Sample size n=39.

In building the CUAHSI HIS choices need to be made with respect to reliance on the capability of existing software, both proprietary and open source. Reliance on other software takes advantage of existing technology, avoids the need to repeat existing capability and may be more reliable and have professional support and maintenance. Respondents were asked opinions regarding the selection of open source or commercial software platforms for the CUAHSI HIS. Respondents are predominantly in favor of HIS client software being open source, but at the same time would like to leverage commercial software and have the capability to work on all operating systems (Table 7).

Table 7. Opinions on software development.

	Strongly Disagree	Disagree	Agree	Strongly Agree	No Opinion
HIS Client software should work on all computer operating systems	2.7%	10.7%	49.3%	30.7%	6.7%
HIS Software should leverage commercial software systems	2.7%	9.3%	40.0%	22.7%	25.3%
HIS Software should be open source	1.3%	8.0%	29.3%	42.7%	18.7%

In addition to opinions regarding the software platform issue, we were interested to know which issues researchers were concerned about when considering the use of open source or commercial software platforms. If the community has strong preferences for open source or proprietary software, it is useful to know why, or which concerns need to be addressed when decisions are made by the HIS development team. We developed the following list of common issues related to choice of operating platform and asked respondents to rank the three most important to them:

- Cost of commercial software required by the HIS user to exploit full HIS capability.
- Long term stability of commercial software and continuation of support by provider
- Existence of support and upgrade options for open source solutions
- Flexibility to scrutinize and modify source code
- The professional support provided by commercial software
- The functionality available in commercial software

The results show that the cost to the user of commercial software required to use the HIS is the greatest concern. This is closely followed by concern that the HIS have the stability, long-term support, and functionality available in commercial software (Figure 4).

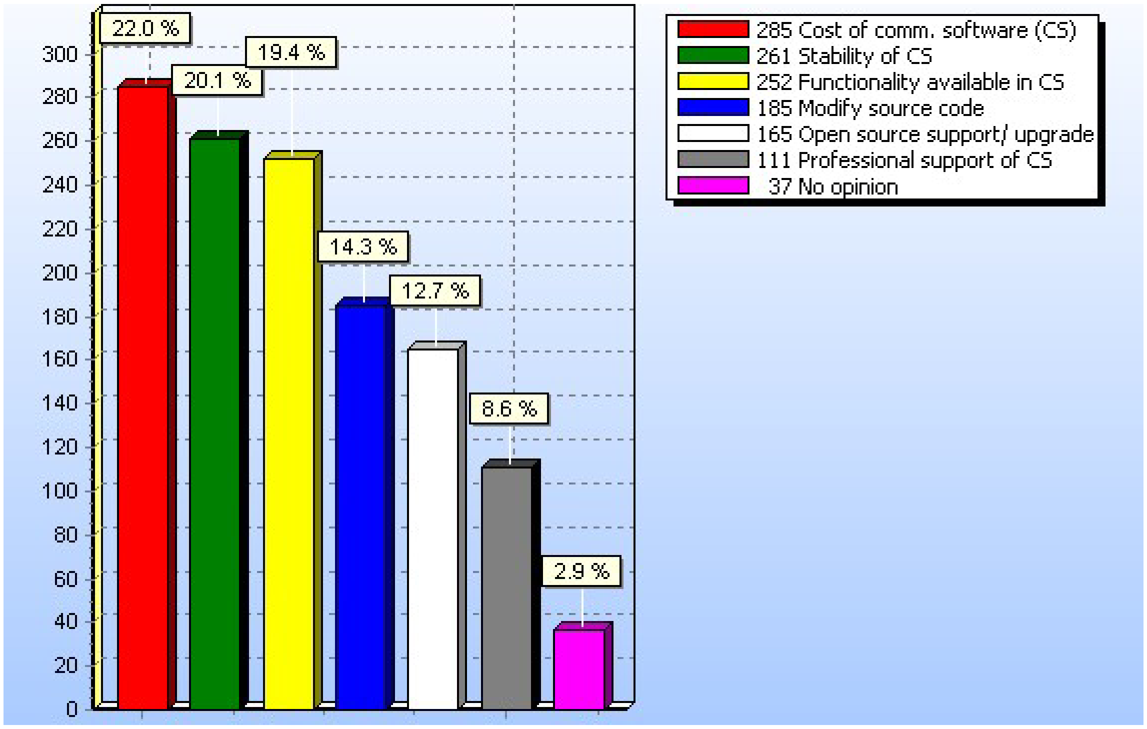


Figure 4. Importance of issues related to use of commercial and open source software (using a value score where first choice has a score of 3 points, second choice has a score of 2 points, and third choice has a score of 1 point).

Hydrologic Data Acquisition and Preparation

To understand the current patterns in hydrologic data acquisition and preparation, we asked what proportion of research time is spent preparing or preprocessing data into appropriate forms needed for research purposes. A significant fraction of research time is spent preparing and preprocessing data (Figure 5).

- More than 80% of respondents spend more than10% of research time preparing data.
- More than 35% of respondents spend more than 25% of research time preparing data.
- More than 12% of respondents spend more than 50% of research time preparing data.

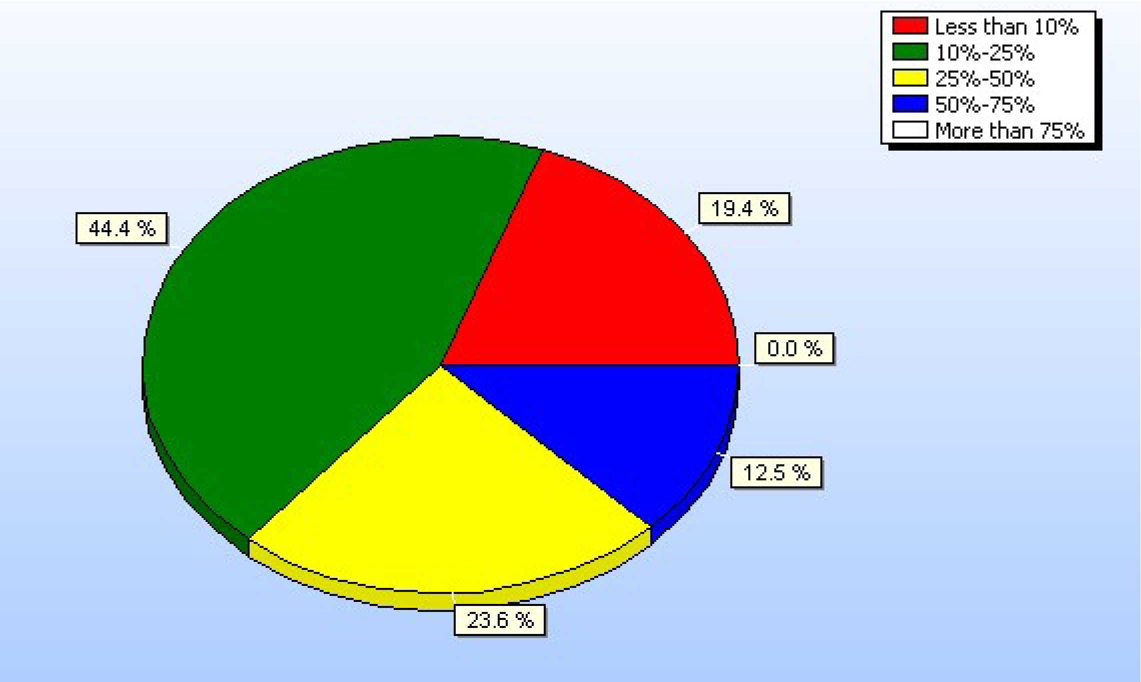


Figure 5. Proportion of research time spent preprocessing or preparing data.

A matrix of datasets that the CUAHSI HIS may incorporate was presented with four choices for rating the priority of each dataset for inclusion in the HIS. For each dataset, the respondent could choose 1) Essential to my research, 2) Am likely to use in my research, 3) I am aware of this, but not likely to use it, and 4) I have not heard of this dataset. The following datasets ranked highest for incorporating into the CUAHSI HIS[2].

1. USGS Streamflow
2. NCDC Precipitation
3. Remote Sensing data (e.g. LANDSAT, GOES, AVHRR)
4. National Elevation Dataset and derivatives (EDNA)
5. Other NCDC Weather and Climate Data
6. USGS Groundwater levels
7. National Land Cover dataset (NLCD)
8. Soils Data (STATSGO/SSURGO)
9. National Hydrography Dataset (NHD)
10. NCDC Pan Evaporation

The tabulated results (Table A3.1) and responses to the question about additional datasets to consider for inclusion in the HIS are listed in Appendix 3.[3] Isotope data,

[2] Rank was determined using a weighted average, ("Essential"*2 + "Likely to Use")/3

[3] Other datasets respondents were asked to rank, but which scored lower, include in the following order: USGS Water Chemistry (NASQAN, HBN, Cooperative data), SNOTEL, EPA STORET Water Quality, NEXRAD Radar Precipitation, National Water Quality Assessment (NAWQA), Biological Data, USGS National Geology data, USGS Hydrologic Landscape Regions, PRISM Precipitation data, Climate Model Reanalysis data (e.g. NARR), Aquatic Ecoregions (AQUAECO), and Acidic Surface Waters (A_WATER).

water use and management data, listed among the additional dataset responses, are important datasets not previously identified.

In addition to ranking the above datasets for inclusion in the HIS, we also asked respondents which of these datasets are the most difficult to access and use (Figure 6). The HIS can provide a service to researchers by facilitating the dissemination of important data that is currently challenging to utilize. The top four datasets respondents believed would most benefit from increased ease of access through a Hydrologic Information System are:

1. EPA STORET Water Quality
2. USGS Streamflow
3. Remote Sensing data (e.g. LANDSAT, GOES, AVHRR)
4. NEXRAD Radar Precipitation

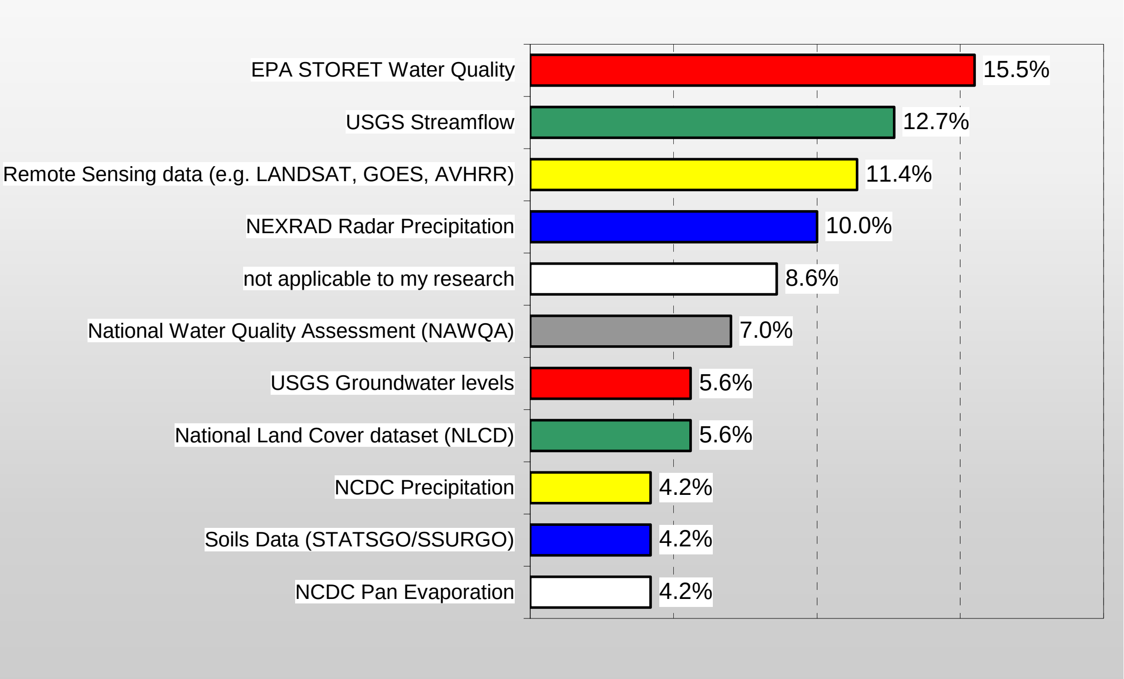


Figure 6. Datasets that are difficult to access and use which would most benefit from increased ease of access through a HIS. All datasets that appear in Table A3.1 were available for respondents to select.

Respondents were asked which spatial scales are most relevant to the data resolution used in their research with the option to select multiple responses. The watershed scale was indicated as most relevant, followed closely by the Field and Sub-watershed resolutions (Figure 7). The prevalence of researchers investigating questions at more than one spatial scale indicates the importance for the HIS to integrate datasets so they may be used at multiple spatial scales. Respondents are predominantly interested in studies at a watershed scale or smaller.

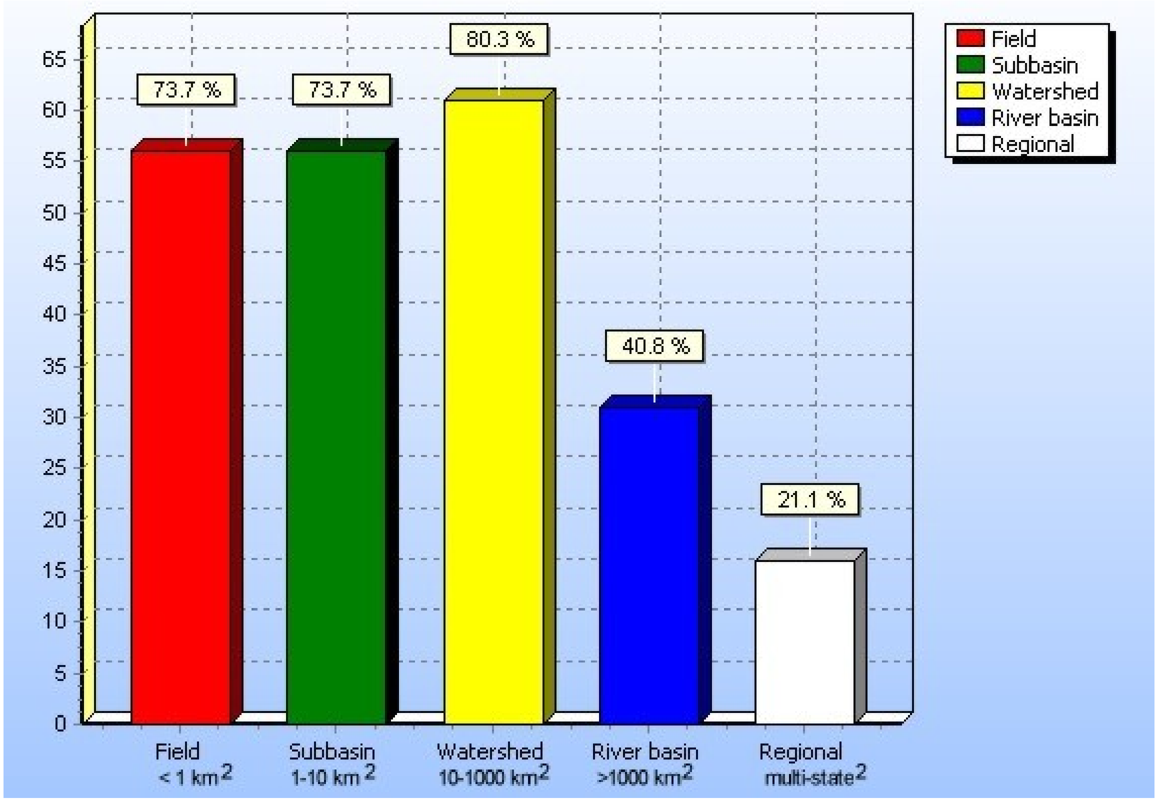


Figure 7. Spatial scales most relevant to data resolution used in research. Respondents could indicate more than one scale, resulting in percentages totaling more than 100%.

The HIS hopes to improve capability for integrating, analyzing, and synthesizing data from disparate sources. Respondents were asked to choose and rank three of the most critical difficulties for the HIS to address. Results show that the HIS under development must address the following common difficulties encountered when using hydrologic data for research (Figure 8):

1. Inconsistent data formats
2. Existence and consistency of metadata
3. Irregular and different timesteps

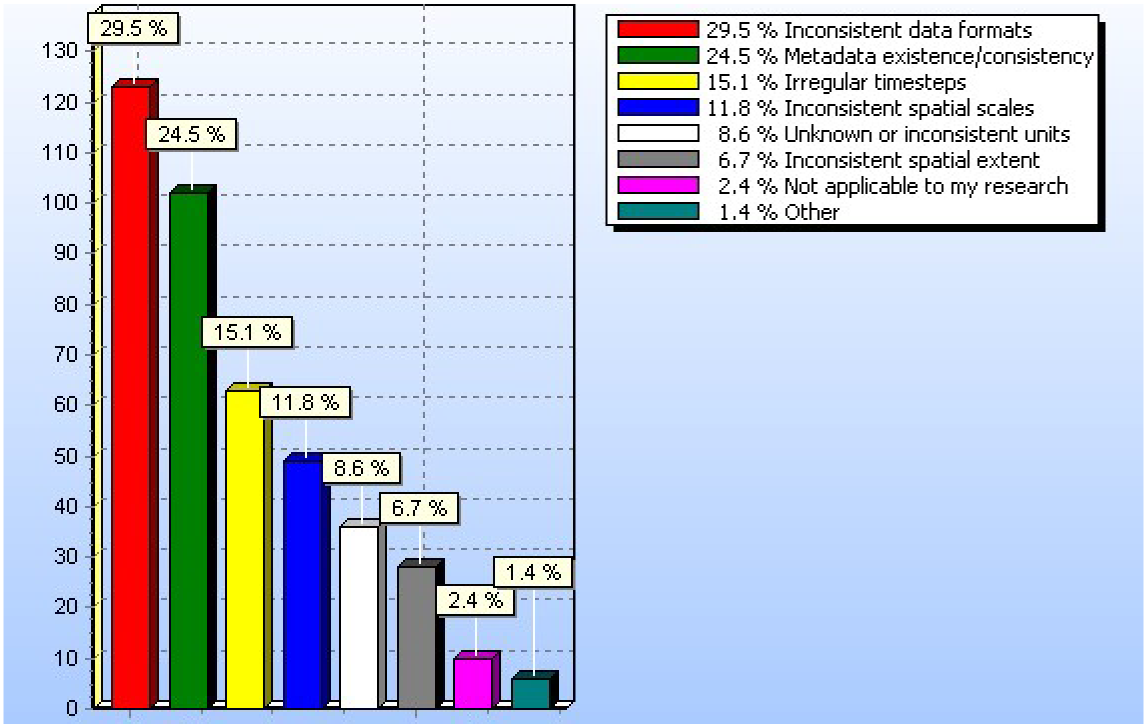


Figure 8. Difficulties in integrating, analyzing and synthesizing data that should be addressed by the HIS (using a value score where first choice has a score of 3 points, second choice has a score of 2 points, and third choice has a score of 1 point).

HIS Services

The CUAHSI HIS has under consideration four main categories of service to the hydrologic research community:

1. Hydrologic data services – these are services that can be used by a hydrologic researchers or students anywhere in the nation to obtain the hydrologic data they require quickly and easily, and in forms that they can readily use;

2. Hydrologic observatory services – these are information services that a CUAHSI Hydrologic Observatory will require to process, archive and display the data measured at the observatory;

3. Hydrologic science services– these are services needed to build the complex digital representations of hydrologic environments needed to support advanced hydrologic modeling, hypothesis testing, and constructing water, energy and mass balances of hydrologic systems;

4. Hydrologic education services – these are services needed to advance the use of hydrologic information in the classroom.

Respondents were asked to rank these four goals. Results in Figure 9 identify Data Services as the functionality that users perceive to be most important for the HIS to provide.

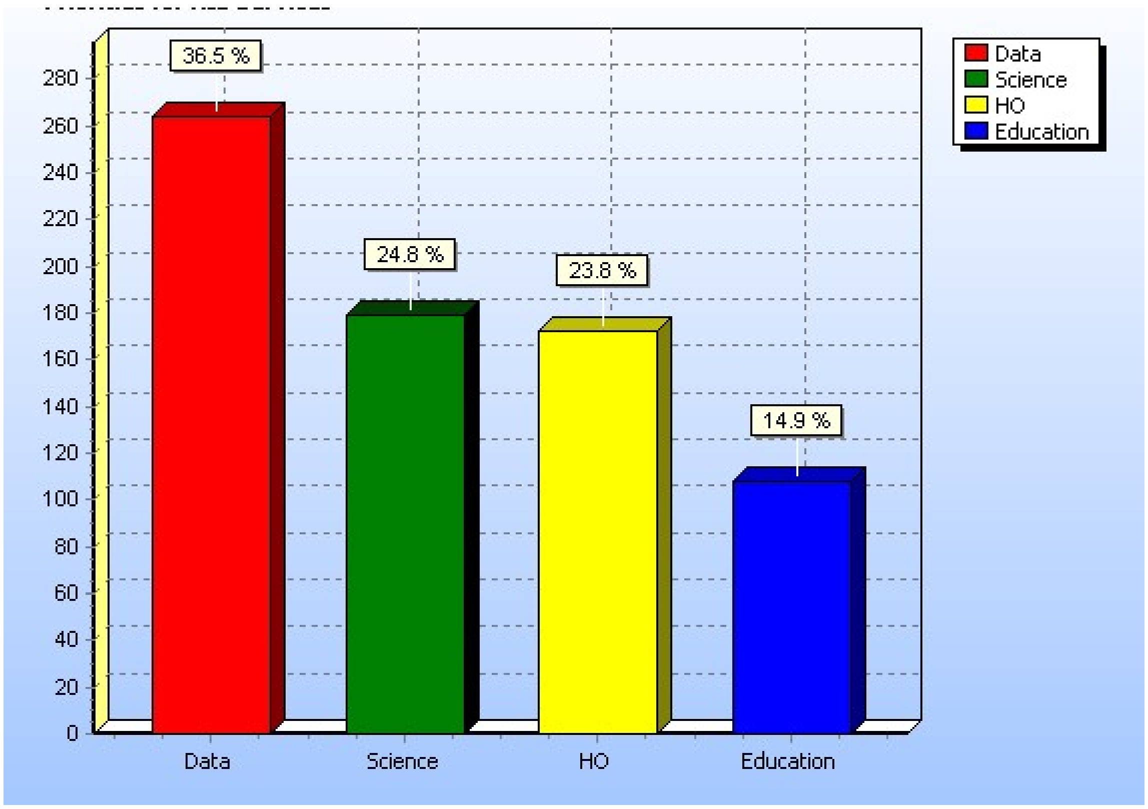


Figure 9. Priorities for general categories of HIS Services. (using a value score where first choice has a score of 4 points, second choice has a score of 3 points, third choice has a score of 2 points, fourth choice has a score of 1 point).

The Need for HIS

Comments about the HIS project show that respondents are excited that the project is moving forward, think it is a great idea and will do much to advance hydrologic science. The HIS is expected to advance the quality of the hydrologic sciences and determine the nature of future research projects. When asked if a CUAHSI HIS was developed with the priorities the respondent had listed for a watershed where they conduct research, 77% responded ‘Yes, a user-friendly digital watershed for data access is what I need’. One would expect that the majority of respondent might answer ‘Yes’ since they are interested enough in the topic to take the time to answer the questionnaire. The important result shown in Figure 10 is that there are many who do not know enough about the CUAHSI HIS project to determine its expected utility. Almost 18% of respondents did not feel that had enough information to know if they would use the HIS or not. This speaks to a need for better communication and articulation to the community of the potential capabilities of the HIS.

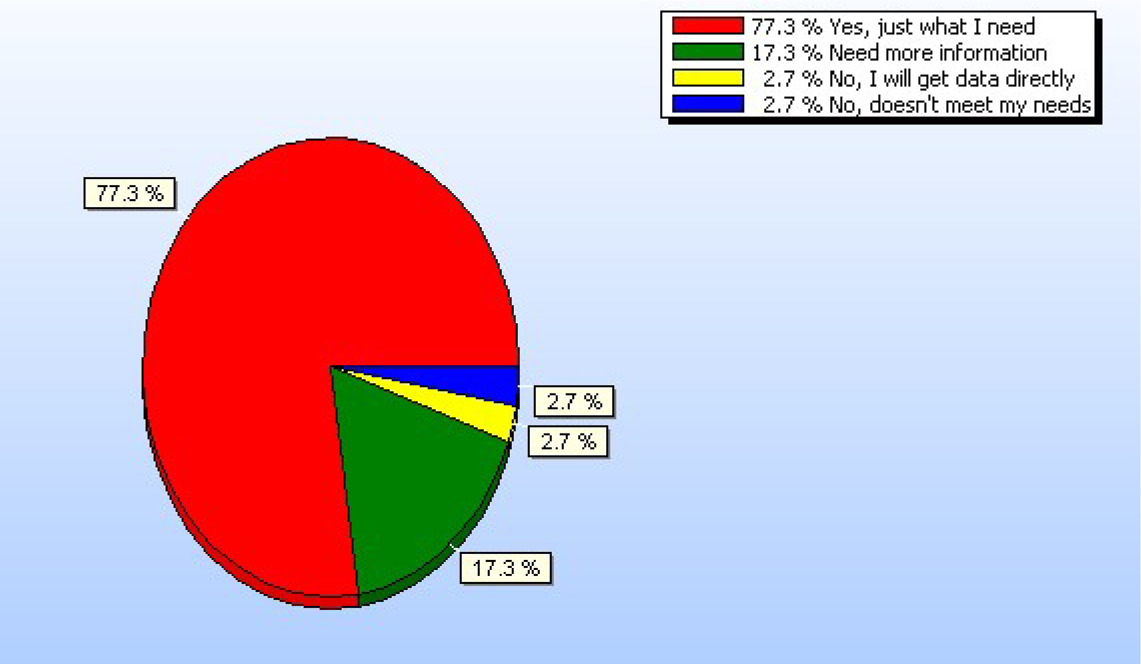


Figure 10. Anticipated use of CUAHSI HIS and the need for more information dissemination.

Conclusions and Recommendations

Our results verify the need for an HIS. Most researchers surveyed spend a significant amount of time preprocessing data for their research and expect that the HIS will be helpful and relevant to their work. Data services are the most important services for the HIS to provide while addressing critical data use difficulties such as inconsistent data formats, the existence and consistency of metadata, and irregular timesteps. Most agree that the HIS client software should work on all operating systems and that it should leverage commercial software as well as be open source. The cost related to the use of the HIS, and its long-term stability and functionality are critical issues to be addressed in the HIS development.

In hydrologic research, FORTRAN is the most popular programming language, Microsoft Excel most used for data management, Matlab is most used for mathematics and statistics work. Most respondents use GIS for their research (93%) and ESRI ArcGIS is the most used software. The survey shows that 80% of respondents use hydrologic modeling in their research. Modflow is the most popular for groundwater modeling, but there is no predominant surface water model of the more the 25 hydrologic models listed in the survey questions and mentioned by respondents. A general, simple, standard, and open interface that could connect with many systems would be the only way to accommodate all of the models used.

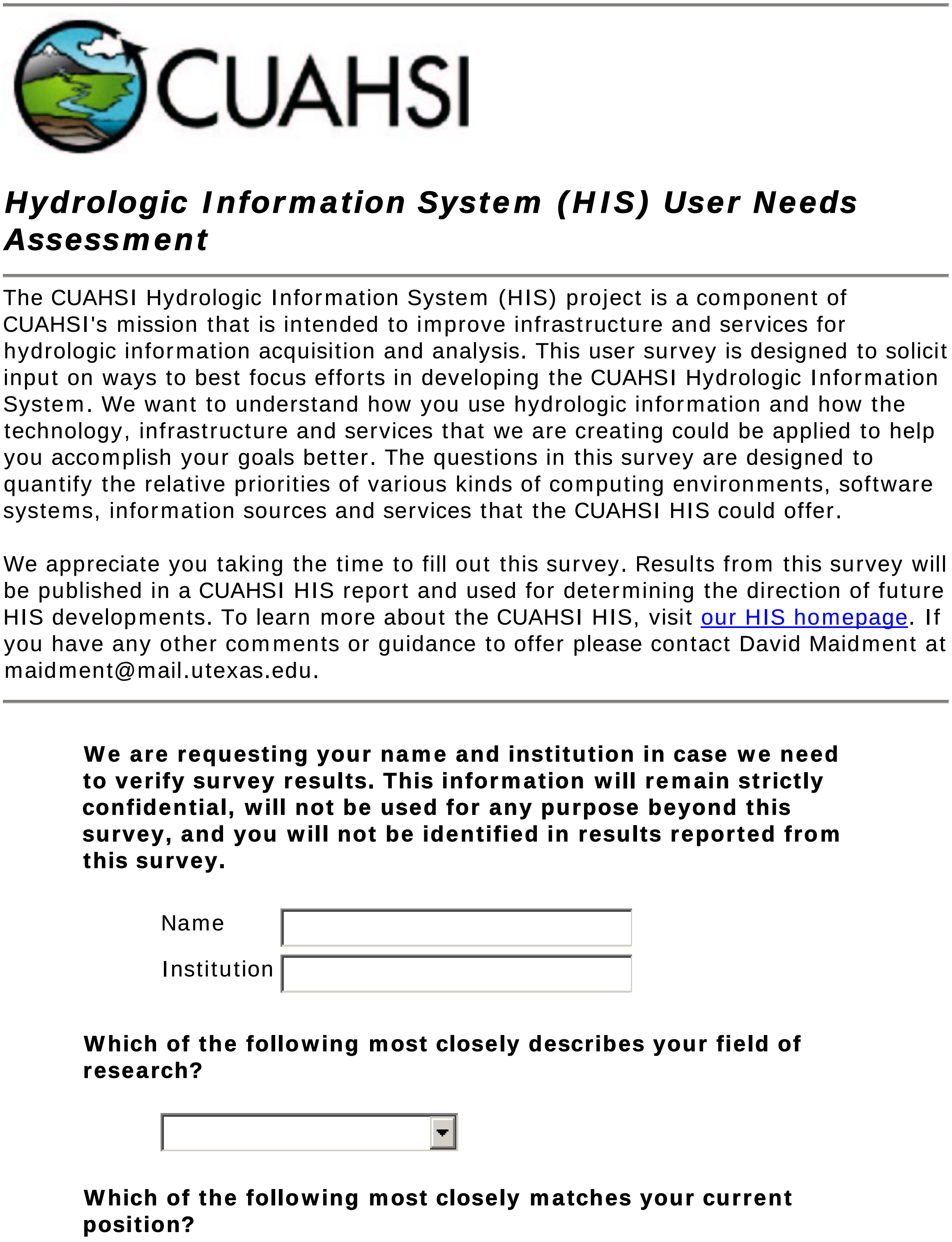
USGS Streamflow and NCDC Precipitation are viewed as the most essential datasets for inclusion in the HIS. Remote sensing data (e.g. LANDSAT, GOES, AVHRR), National Elevation Dataset and derivatives (EDNA), Other NCDC Weather and Climate Data, and USGS Groundwater levels also scored as high priorities. Datasets considered the most difficult to access and use that would most benefit from inclusion in the HIS include EPA STORET Water Quality Data, USGS Streamflow, Remote Sensing data, and NEXRAD Radar precipitation. Results show that most researchers work at multiple spatial scales, predominantly at the field, sub-basin, and watershed levels. The HIS should provide datasets in formats that can be easily utilized at a variety of spatial scales.

During the process of developing the HIS User Needs survey, three levels of HIS user information were outlined: 1) how people are currently conducting hydrologic analysis 2) how researchers believe an HIS could help them conduct hydrologic analysis with detailed HIS service prioritization, and 3) how data providers from other public and private institutions are currently assisting in HIS-like endeavors as well as how CUAHSI can coordinate HIS efforts with them. This paper focused on a web survey intended to collect the first level of information or a 'client-based' survey of software and data use patterns and current practices in hydrologic analysis by researchers at CUAHSI institutions. Future HIS user surveys should focus on prioritization of specific HIS services and how to best coordinate with other public and private providers of hydrologic data.

Acknowledgements

We would like to thank Richard Hooper, Jon Duncan, Xu Liang, Yao Liang, Chunmaio Zheng and Leroy Poff for their contribution and assistance with CUAHSI HIS User Needs preliminary surveys that led to the development of the web survey on which this paper is based.

Appendix 1. Web Questionairre used for HIS User Needs Assessment, May 2005. Values presented in drop down boxes are given with the presentation of results in Tables, Appendices and Figures.



CUAHSI

Hydrologic Information System (HIS) User Needs Assessment

The CUAHSI Hydrologic Information System (HIS) project is a component of CUAHSI's mission that is intended to improve infrastructure and services for hydrologic information acquisition and analysis. This user survey is designed to solicit input on ways to best focus efforts in developing the CUAHSI Hydrologic Information System. We want to understand how you use hydrologic information and how the technology, infrastructure and services that we are creating could be applied to help you accomplish your goals better. The questions in this survey are designed to quantify the relative priorities of various kinds of computing environments, software systems, information sources and services that the CUAHSI HIS could offer.

We appreciate you taking the time to fill out this survey. Results from this survey will be published in a CUAHSI HIS report and used for determining the direction of future HIS developments. To learn more about the CUAHSI HIS, visit our HIS homepage. If you have any other comments or guidance to offer please contact David Maidment at maidment@mail.utexas.edu.

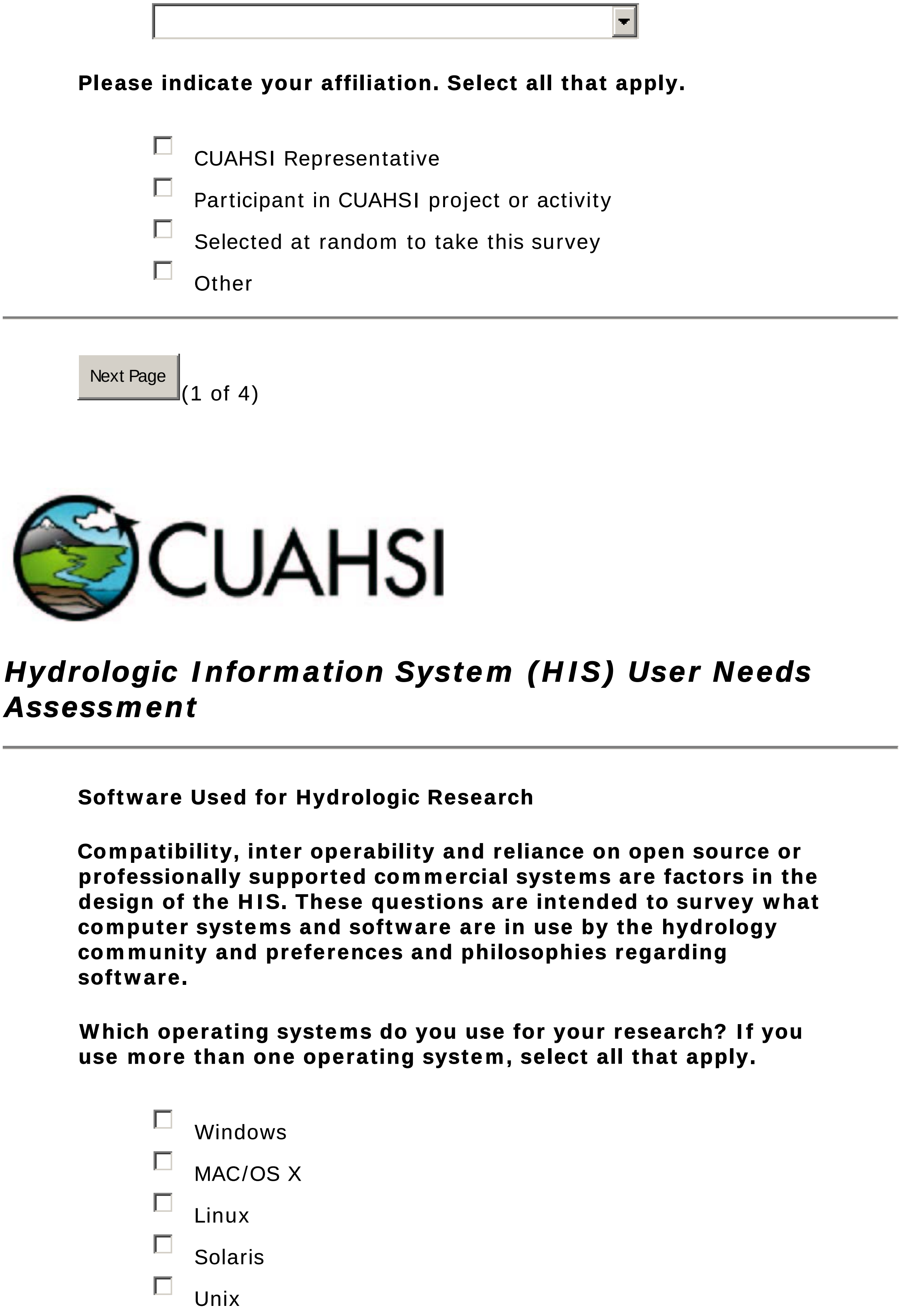
We are requesting your name and institution in case we need to verify survey results. This information will remain strictly confidential, will not be used for any purpose beyond this survey, and you will not be identified in results reported from this survey.

Name

Institution

Which of the following most closely describes your field of research?

Which of the following most closely matches your current position?



▾

Please indicate your affiliation. Select all that apply.

□ CUAHSI Representative

□ Participant in CUAHSI project or activity

□ Selected at random to take this survey

□ Other

Next Page (1 of 4)

CUAHSI

Hydrologic Information System (HIS) User Needs Assessment

Software Used for Hydrologic Research

Compatibility, inter operability and reliance on open source or professionally supported commercial systems are factors in the design of the HIS. These questions are intended to survey what computer systems and software are in use by the hydrology community and preferences and philosophies regarding software.

Which operating systems do you use for your research? If you use more than one operating system, select all that apply.

□ Windows

□ MAC/OS X

□ Linux

□ Solaris

□ Unix

□ Other (please specify)

If you selected other, please specify:

In the questions below we are asking that you rank selections picking a few that are highest priority to you. This form of question is designed to have you to tell us which is most important among the choices presented.

The HIS may include the capability to interact with other software. Please choose up to two software packages in each category that are most important for hydrologic analysis in your research.

Programming

1	
2	

Data Management

1	
2	

If you program database client software to access data, which programming language do you use?

If you selected other, please specify:

GIS (Geographic Information Systems)

1	
2	

Mathematics/Statistics

1	▾
2	▾

Hydrologic Models

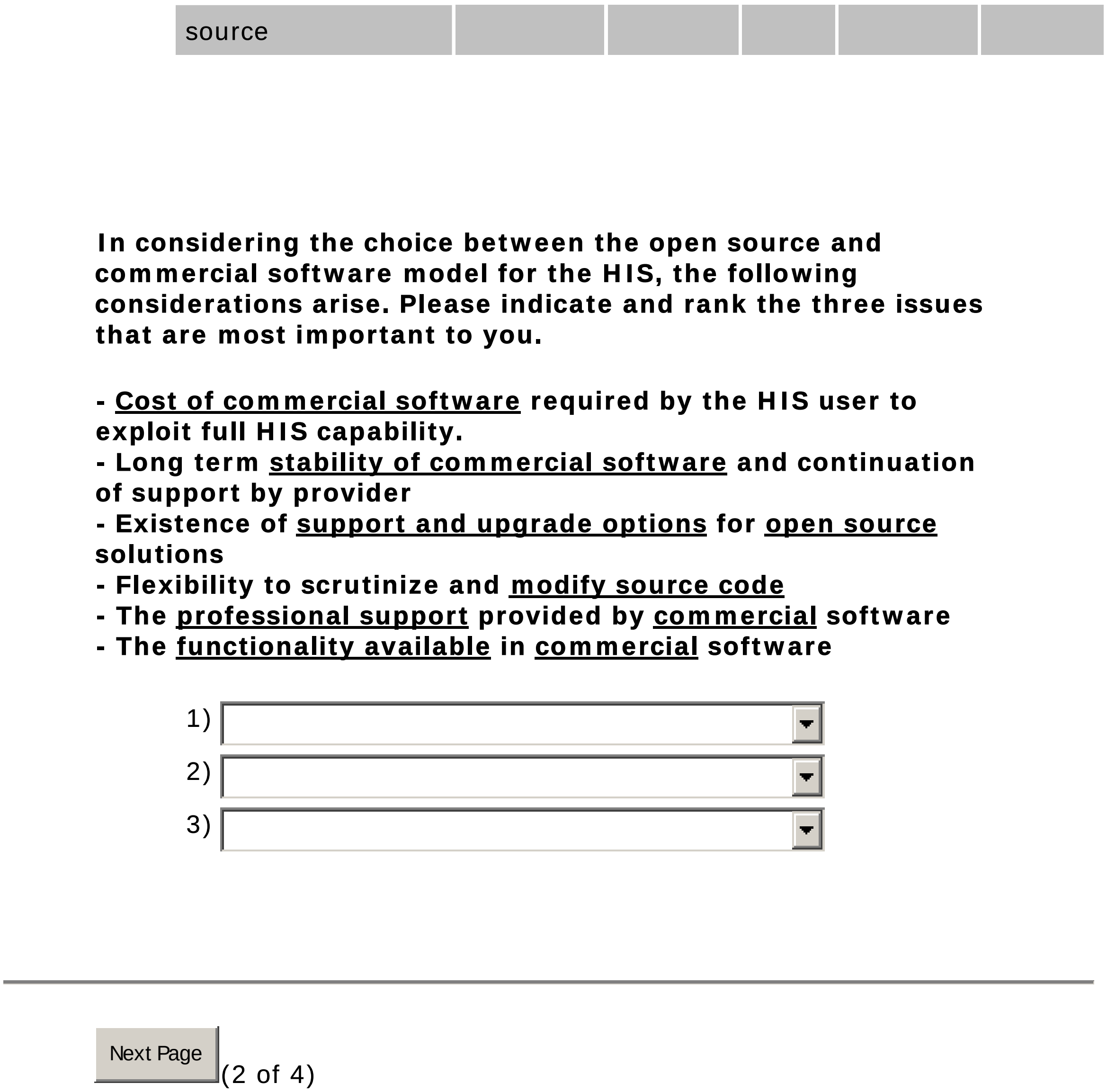
1	▾
2	▾

Are there any other software packages we should consider for interfacing with the CUAHSI HIS?

In building the CUAHSI HIS choices need to be made with respect to reliance on the capability of existing software, both proprietary and open source. Reliance on other software takes advantage of existing technology, avoids the need to repeat existing capability and comes with professional reliability and support. Disadvantages include the costs to users (e.g. for a needed commercial system like Windows or GIS), lack of flexibility (e.g. to change something and access the code) and dependence on the business strategies of the provider.

Please indicate your opinion in the selection of commercial/open source software platforms for the CUAHSI HIS.

	Strongly Disagree	Disagree	Agree	Strongly Agree	No Opinion
HIS Client software should work on all computer operating systems	○	○	○	○	○
HIS Software should leverage commercial software systems	○	○	○	○	○
HIS Software should be open	○	○	○	○	○



source					

In considering the choice between the open source and commercial software model for the HIS, the following considerations arise. Please indicate and rank the three issues that are most important to you.

- Cost of commercial software required by the HIS user to exploit full HIS capability.
- Long term stability of commercial software and continuation of support by provider
- Existence of support and upgrade options for open source solutions
- Flexibility to scrutinize and modify source code
- The professional support provided by commercial software
- The functionality available in commercial software

1) ▾

2) ▾

3) ▾

(2 of 4)

CUAHSI

Hydrologic Information System (HIS) User Needs Assessment

Hydrologic Data Acquisition and Preparation

What proportion of your research time do you spend on preparing or preprocessing data into appropriate forms needed for research purposes?

○ Less than 10%

○ 10%-25%

○ 25%-50%

○ 50%-75%

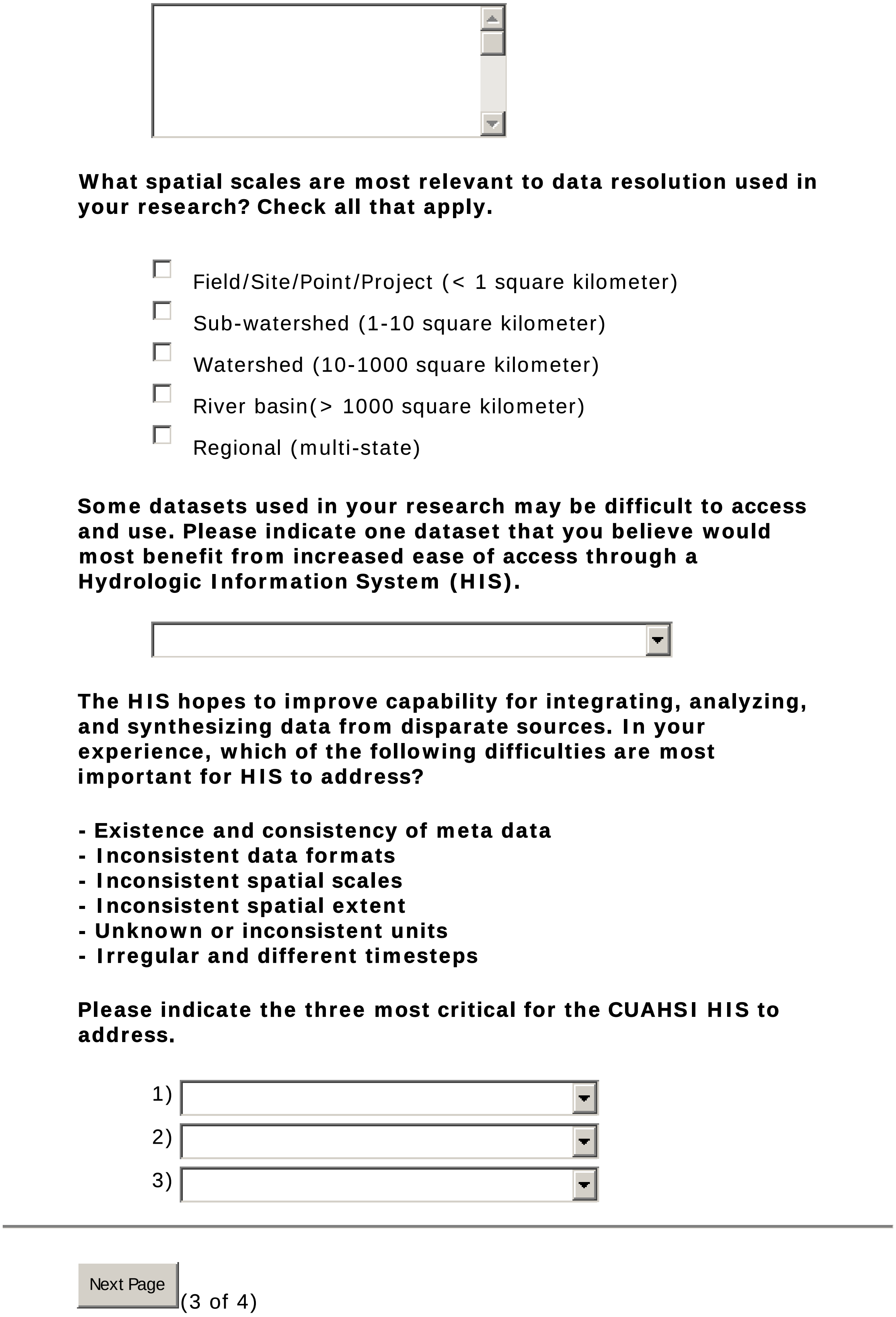
○ More than 75%

Consider the following datasets that the CUAHSI HIS may incorporate. Please rate each dataset for its priority for inclusion in the HIS.

	I have not heard of this dataset	I am aware of this, but not likely to use it	Am likely to use in my research	Essential to my research
USGS Streamflow	○	○	○	○
USGS Groundwater levels	○	○	○	○
USGS Water Chemistry (NASQAN, HBN, Cooperative data)	○	○	○	○
National Water Quality Assessment (NAWQA), Biological Data	○	○	○	○
EPA STORET Water Quality	○	○	○	○

NCDC Precipitation	○	○	○	○
NCDC Pan Evaporation	○	○	○	○
Other NCDC Weather and Climate Data	○	○	○	○
SNOTEL	○	○	○	○
National Elevation Dataset and derivatives (EDNA)	○	○	○	○
National Hydrography Dataset (NHD)	○	○	○	○
Soils Data (STATSGO/SSURGO)	○	○	○	○
USGS National Geology data	○	○	○	○
National Land Cover dataset (NLCD)	○	○	○	○
USGS Hydrologic Landscape Regions	○	○	○	○
NEXRAD Radar Precipitation	○	○	○	○
PRISM Precipitation data	○	○	○	○
Climate Model Reanalysis data (e.g. NARR)	○	○	○	○
Remote Sensing data (e.g. LANDSAT, GOES, AVHRR)	○	○	○	○
Aquatic Ecoregions (AQUAECO)	○	○	○	○
Acidic Surface Waters (A_WATER)	○	○	○	○

Are there any other datasets you think should be included in a HIS?



What spatial scales are most relevant to data resolution used in your research? Check all that apply.

□ Field/Site/Point/Project (< 1 square kilometer)
□ Sub-watershed (1-10 square kilometer)
□ Watershed (10-1000 square kilometer)
□ River basin(> 1000 square kilometer)
□ Regional (multi-state)

Some datasets used in your research may be difficult to access and use. Please indicate one dataset that you believe would most benefit from increased ease of access through a Hydrologic Information System (HIS).

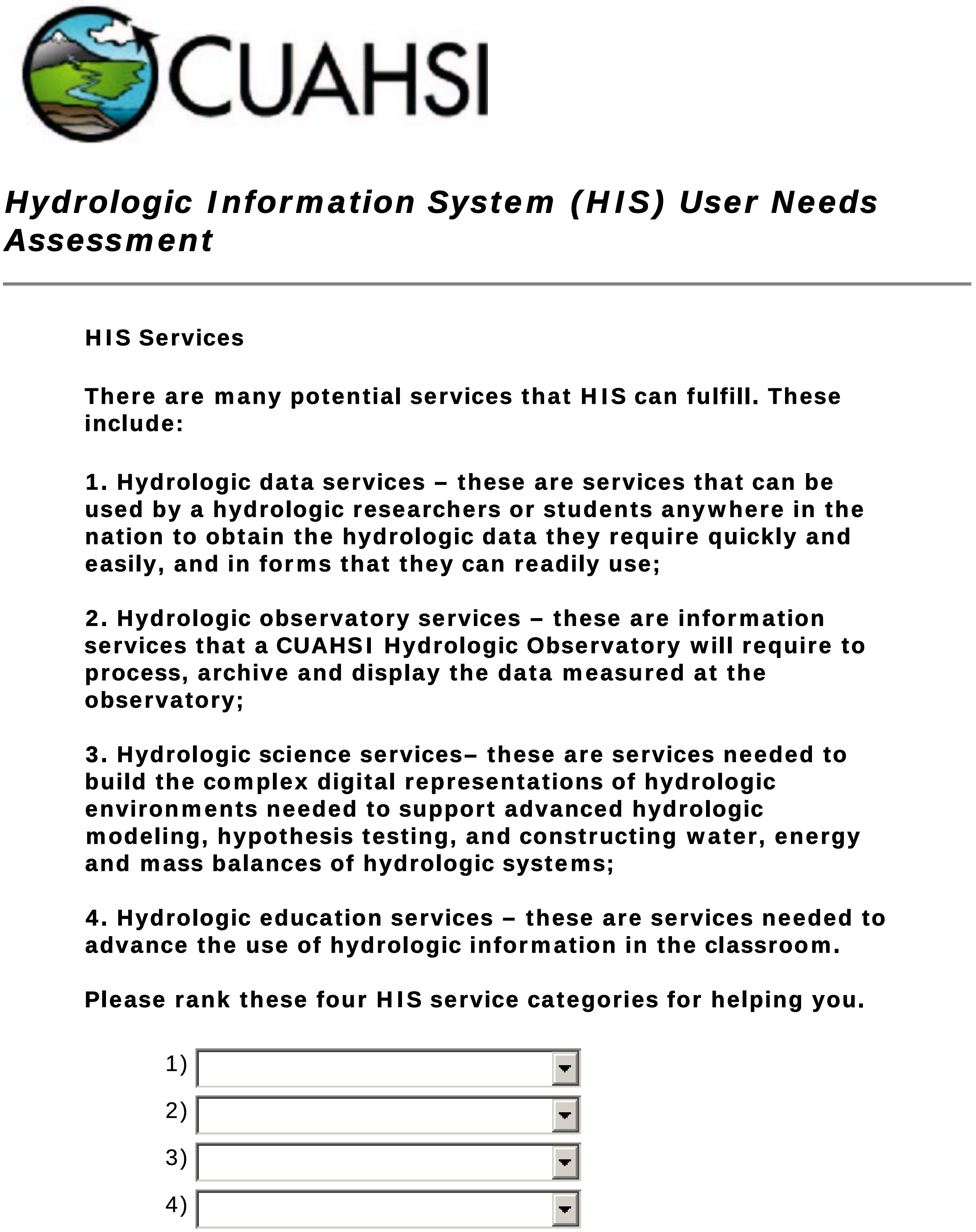
The HIS hopes to improve capability for integrating, analyzing, and synthesizing data from disparate sources. In your experience, which of the following difficulties are most important for HIS to address?

- Existence and consistency of meta data
- Inconsistent data formats
- Inconsistent spatial scales
- Inconsistent spatial extent
- Unknown or inconsistent units
- Irregular and different timesteps

Please indicate the three most critical for the CUAHSI HIS to address.

1)
2)
3)

(3 of 4)



CUAHSI

Hydrologic Information System (HIS) User Needs Assessment

HIS Services

There are many potential services that HIS can fulfill. These include:

1. Hydrologic data services – these are services that can be used by a hydrologic researchers or students anywhere in the nation to obtain the hydrologic data they require quickly and easily, and in forms that they can readily use;

2. Hydrologic observatory services – these are information services that a CUAHSI Hydrologic Observatory will require to process, archive and display the data measured at the observatory;

3. Hydrologic science services– these are services needed to build the complex digital representations of hydrologic environments needed to support advanced hydrologic modeling, hypothesis testing, and constructing water, energy and mass balances of hydrologic systems;

4. Hydrologic education services – these are services needed to advance the use of hydrologic information in the classroom.

Please rank these four HIS service categories for helping you.

1)

2)

3)

4)

If a CUAHSI HIS was developed with the priorities you have listed for a watershed where you conduct research, would you use it?

○ Yes, a user-friendly digital watershed for data access is what I need.

○ No, it doesn't meet my current research needs

○ No, I would rather get data directly from various data providers.

○ I don't have enough information to know.

Please provide any additional comments or suggestions regarding the HIS project.

Thank you for the information you have provided!

Submit Survey (4 of 4)

Appendix 2: Responses to question “Are there any other software packages we should consider for interfacing with the CUAHSI HIS?” (raw data)

- Maple
- Rivertools
- I use JMP for data management and statistics
- PHREEQC,NETPATH
- Rockware Rockworks Geochemists Workbench Surfer ModFlow
- AGWA
- Math: Mathcad, Scientific Notebook Curve fitting: TableCurve Parameter Estimation: Pest, UCODE Hydrologic Modeling: TOUGH and related
- FEFLOW
- Kineros model
- ENVI
- U.S. EPA Stormwater Management Model (SWMM5)
- Chemical equilibrium and speciation software such as PHREEQC
- HSPF
- CUENCAS being developed by our group at the univerrsity of colorado, boulder
- IPW (Frew/UCSB) ISNOBAL many others -- be flexible
- AutoCAD
- Highest priority is an efficient map based web interface. Second priority is a programmable application interface so that data access can be programmed and scripted.
- User developed data analysis, data visualization, or modeling applications.
- GAMS (for solving dynamic systems of equations) LIMDEP (statistical software) STATA (statistical software) GAUSS (statistical software)
- ERDAS Imagine a raster based image processing and GIS software with modeling capabilities.
- Imagine ERDAS for GIS image processing
- Penn State Integrated Hydrology Model
- For my research, I use the statistical software package Statistica (Statsoft, Tulsa, OK) extensively, Argus One
- MLAEM, Split, GFlow, TwoDAN, 3DFlow, Tim

Appendix 3. Important Datasets for inclusion in the CUAHSI HIS

Table A3.1 Datasets CUAHSI may incorporate ranked by priority ratings.

Rank		Essential to my research	Am likely to use in research	I am aware of this, but not likely to use	I have not heard of dataset	Weighted[1] Score
1	USGS Streamflow	60.8%	27.0%	9.5%	2.7%	49.5%
2	NCDC Precipitation	35.1%	44.6%	12.2%	8.1%	38.3%
3	Remote Sensing data (e.g. LANDSAT, GOES, AVHRR)	30.7%	34.7%	32.0%	2.7%	32.0%
4	National Elevation Dataset and derivatives (EDNA)	32.4%	31.1%	24.3%	12.2%	32.0%
5	Other NCDC Weather and Climate Data	22.5%	47.9%	21.1%	8.5%	31.0%
6	USGS Groundwater levels	25.7%	39.2%	28.4%	6.8%	30.2%
7	National Land Cover dataset (NLCD)	21.3%	46.7%	28.0%	4.0%	29.8%
8	Soils Data (STATSGO/SSURGO)	20.0%	44.0%	26.7%	9.3%	28.0%
9	National Hydrography Dataset (NHD)	25.3%	32.0%	22.7%	20.0%	27.5%
10	USGS Water Chemistry (NASQAN, HBN, Cooperative data)	20.0%	38.7%	32.0%	9.3%	26.2%
11	NCDC Pan Evaporation	18.7%	41.3%	29.3%	10.7%	26.2%
12	SNOTEL	26.0%	17.8%	35.6%	20.5%	23.3%
13	EPA STORET Water Quality	18.7%	32.0%	30.7%	18.7%	23.1%
14	National Water Quality Assessment (NAWQA), Biological Data	16.9%	35.2%	36.6%	11.3%	23.0%
15	NEXRAD Radar Precipitation	18.7%	30.7%	38.7%	12.0%	22.7%
16	USGS National Geology data	13.5%	33.8%	36.5%	16.2%	20.3%
17	PRISM Precipitation data	12.0%	29.3%	33.3%	25.3%	17.8%
18	USGS Hydrologic Landscape Regions	9.3%	33.3%	38.7%	18.7%	17.3%
19	Climate Model Reanalysis data (e.g. NARR)	6.7%	18.7%	38.7%	36.0%	10.7%
20	Aquatic Ecoregions (AQUAECO)	1.4%	16.4%	32.9%	49.3%	6.4%
21	Acidic Surface Waters (A_WATER)	1.3%	4.0%	29.3%	65.3%	2.2%

[1] Note: The weighted score is (“Essential”*2 + “Likely to Use”)/3

Responses to question “Are there any other datasets you think should be included in a HIS?” (raw data)

- Isotope data
- A lot of USGS stations are being farmed to state or local resource agencies. This likely will continue. Is there any way to include these stations?
- SRTM! Calif Coop Snow Survey
- Other elevation data sets (state)
- USGS DRG Maps (Terraserver topo and aerial images)
- The many regional datasets that, for a given area, are far superior to the national datasets. The HIS needs the best coverage, not just national coverages, which will be what folks use only if they cannot get something better.
- Quantification of diversions, water use, irrigation
- Local agency data. Diversion flows in managed systems. Reservoir levels and volumes. GIS base layers (administrative boundaries, watershed boundaries, etc.) There are other existing sources of climate information. Many of these are publicly available, and may provide different data than NCDC.
- DEM digital elevation model data

- State data sets on water/use, socio-economic data
- snowcourse (whenever feasible), oceanography data (i.e SEASAT, circulation, etc)

Appendix 4: Preliminary Information Gathering and Pilot Survey

Preliminary Information Gathering

In order to develop a thorough and relevant web-based survey to collect specific information on HIS User Needs, HIS project collaborators were requested to conduct pre-surveys at their home institutions. Four separate surveys were conducted at Utah State University, University of California at Berkley, Virginia Tech University and the University of Alabama. Surveys were independently designed by HIS project collaborators and results were presented at the March 20, 2005 HIS Symposium in Austin, Texas by HIS Project collaborators: David Tarboton (USU), Xu Liang (UCBerkley), Yao Liang (Virginia Tech), Chunmaio Zheng (University of Alabama). LeRoy Poff (Colorado State University) presented perspectives representing Biology and Ecology HIS users.

The pre-surveys served as an initial information gathering effort. The questions developed by collaborators focused on fulfilling project goals, but how the questions were presented was left to their discretion. This created broad information on HIS User Needs, and helped focus the surveys that followed. There were three main goals for the preliminary information gathering:

- Set clear objectives for the web-based information collection; decide what information was necessary to collect from users
- Create a definition of the 'HIS User Community' in order to target the correct population or sample for the web-based survey
- Design clear questions in unambiguous formats

With the use of preliminary data collected from different institutions, we obtained feedback on which open-ended questions are relevant and began development of a survey that probes deeper into how data is used by different groups. In this appendix we present results from the Utah State University and University of California at Berkeley pre surveys as well as the pilot survey conducted at the March 2005 HIS Symposium in Austin.

Utah State University Preliminary Survey

At Utah State University, the mode of data collection was a questionnaire distributed by email to members of the campus water sciences community through the Water Initiative. The Water Initiative is a framework for all the Water Sciences at USU and includes a group of physical, biological, and social scientists, and engineers located in six colleges, multiple departments, and academic units whose careers focus on water-related science, engineering, and policy problems. Surveys were received back by eighteen respondents representing research fields in Hydrology, Watershed Science, Geomorphology, Fisheries, Biogeochemistry, Aquatic Ecology, Water resources engineer (includes irrigation), Environmental Engineering, Meteorology, Ecology, Natural Resources Sociology, and Remote Sensing/GIS. Twelve of the respondents were faculty, five graduate students, and one university professional.

Results

At USU, Windows is the most popular operating system used for research (80%) with some respondents also using Unix, Linux, and MAC/OS. Microsoft Excel and ESRI ARCGIS/Arcview are the most used software programs used by USU researchers; C++, Visual Basic, Matlab, MS Access and Fortran are also highly utilized. USU respondents were presented with a set of 10 software functions that could potentially be included in an HIS and asked to score each on a scale between 1 and 5 (where 1=Never use or do not find useful; 2=Have used but do not rely on this; 3=Use occasionally and am comfortable with its use; 4=Use often; 5=Use frequently and find indispensable). The following software functionality was the most important (score > 4.0) for including in an HIS: data storage and retrieval, visualization of spatial data, visualization of time series data, and building relational links. The software functionality USU respondents felt was less important (score <4.0) included: efficient coupling with 3rd party analysis software, presentation to non-technical audiences, development of publication quality figures, numerical analysis, multivariate statistical analysis, and univariate statistical analysis. On a scale of 1 to 5 (where 1= not important and 5 = essential), USU respondents believe that *CUAHSI HIS software should work on all computer systems (Windows, Linux, Mac, Unix)* (4.5) and that *CUAHSI HIS software should work independently from any 3rd party software (e.g. Matlab, ArcGIS)* (3.8).

When asked about the priority of specific datasets for inclusion in the HIS (where 1 =low and 5 = high), USU respondents believe that the National Elevation Dataset, USGS Historical Streamflow, NCDC Precipitation, and the National Hydrography Dataset are the most important (scores >4.5). When asked about the priority of specific roles for the HIS, 'Retrieval of relevant National, Community, and Hydrologic Observatory datasets' and 'Uploading, archival and sharing of hydrologic data with collaborators and the CUAHSI community' were the highest priority for USU respondents (scores > 4.4). To comment on the quality of the data collection, all datasets presented scored at 3.7 or higher, and all HIS roles presented scored 3.1 or higher, showing a bias towards everything presented as a priority. Information on local data and standards was collected in an open format. Results from these questions were used to develop some of the questions presented at the Texas Symposium. The USU survey took an average of 15 minutes (with a range of 5-30 minutes).

UC Berkeley Preliminary Survey

This survey was conducted by the faculty and graduate students at UC Berkeley and researchers at the Lawrence Berkeley National Laboratory (LBNL). A total of 29 individuals from five departments, who conduct research in the general areas of hydrology or some related research projects, participated in the survey. The five departments include: Civil and Environmental Engineering (CE), Environmental Science, Policy, and Management (ESPM), Earth and Planetary Science (EPS), Landscape Architecture and Environmental Planning (LAEP), and LBNL (see Figure A4.1 below). The contributions from researchers in related fields helped to understand the status of the use of hydrologic information and systems at Berkeley in a more interdisciplinary context.

The specialties of those who participated in the survey were divided into seven categories raging from hydrology to ecology (see Figure A4.2 below). The diversity of participants' specialties revealed that people from related fields use the same or similar data sources and systems and that we have common challenges to cope with for better scientific information use.

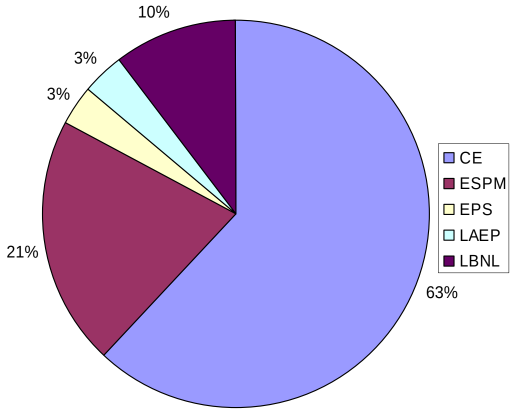


Figure A4.1. Participation by department

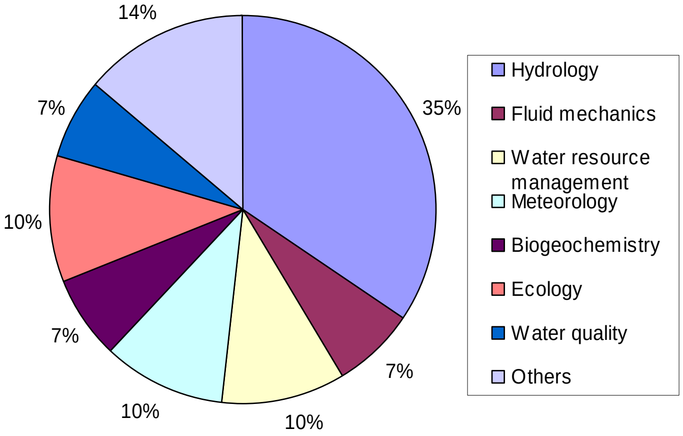


Figure A4.2. Participation by specialty

This survey was implemented using two approaches: Web-based approach and paper survey approach. 23 people used the Web-based version, and 6 people used paper questionnaire. The survey was divided into four sections: 1) systems and software; 2) data and sources; 3) needs of a data system for research, applications, and education; and 4) CUAHSI HIS. Each section included 4-7 questions. The survey was performed from January 20 to February 24, 2005.

Results

More than 70 percent of respondents at Berkeley use Windows (Figure A4.3). Those who work on modeling were found to use either Unix/Linux or a combination of two or three platforms. When asked about software that the participants prefer to use for data analysis, participants indicated that they preferred easy-to-use software with appropriate functions rather than programming languages such as C and Java (Figure A4.4). For this question, people were allowed to provide multiple answers.

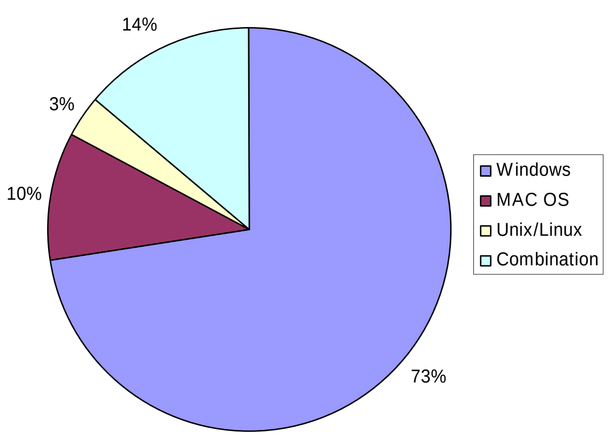


Figure A4.3. Operating systems in use at Berkeley.

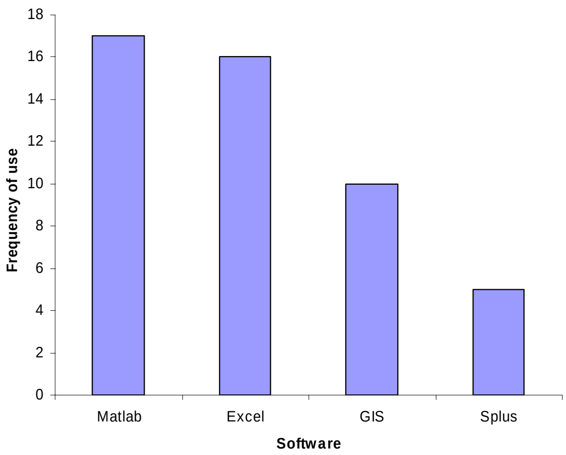


Figure A4.4. Preferred data analysis software.

In a question asking whether the participant developed any software for his/her research, 8 out of 29 answered "yes." The types of software development ranged from Matlab programs to a Web-based data analysis system. The participants described that most of their system administrators have multiple skills such as server management and hardware maintenance. Also, the participants indicated their preference of being provided by an effective and easy-to-use interface so that they can easily connect their own programs (e.g., Matlab or other software programs) to HIS data. Ten of the participants also indicated that they preferred to use GIS software and have a GIS type of components in the HIS system in a user-friend way.

In the section on data and sources, the participants indicated that they spent, on an average, about 30 % of their total research time on data processing (Figure A4.5(a)). When they were asked about questions of what data sources they often used and what kind of difficulties there were in using these data, the participants indicated the following main concerns and would like see they are to be addressed by HIS. These include:

- Necessity of assess to many different data sources with very different interfaces
- Lack of data visualization tools
- Large uncertainties associated with data
- Lack of basic functions to conduct data analysis (e.g., checking consistency, basic statistics, etc.) before downloading the data

When asked which data sources participants would like to use, participants replied that they preferred to use data that were from more established data providers such as UGSG and NCDC and expected to continue using them (Figure A4.5(b)). Thus, the participants suggested that CUAHIS HIS provide easier access to all of the existing popular data sources.

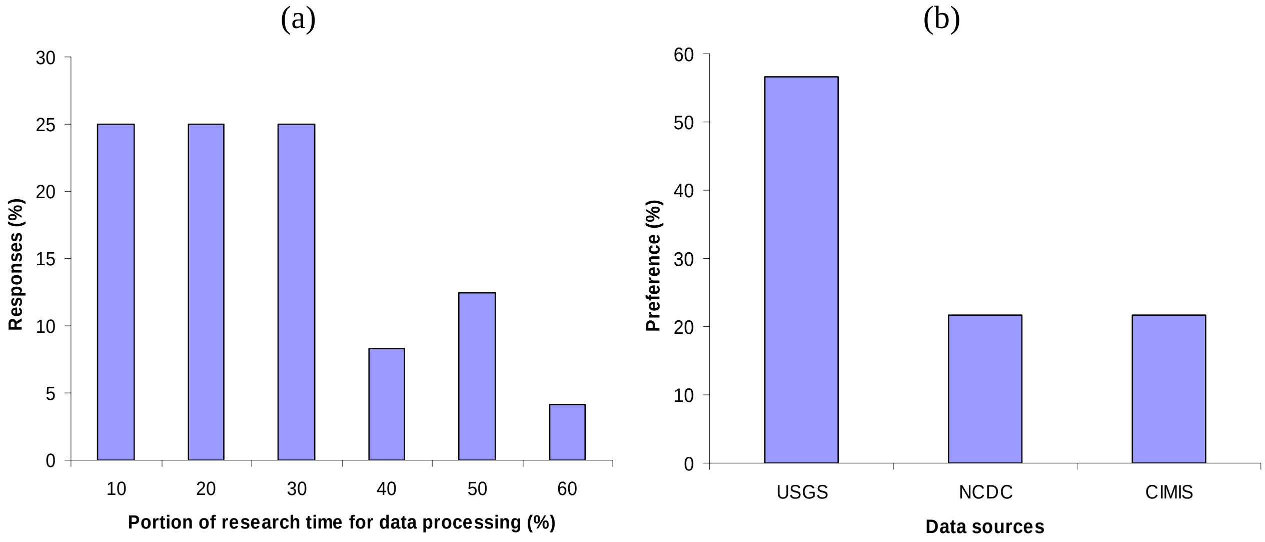


Figure A4.5. Results in percentage (%) of (a) the total research time for data processing and (b) preferred data sources.

In the section regarding needs of a data system for research, applications, and education, answers can be summarized through the identifications of the following needs:

- to address common problems that people encounter
- to provide quick and easy-to-use visualization and basic statistic functions to check the datasets before the user downloads the data
- to integrate various data sources in a single Web system
- to provide easy access to various data sources
- to provide assess to existing popular data providers
- to provide a user-friend connection to popular software programs for further in-depth data analysis

When asked about the common problems that they encounter, the participants indicated the lack of basic functionalities in most of the current data sources. Also, 100 % of the participants express their needs to have an ease of getting data, and 40% indicated that a complicated data system would be helpful, but not necessary (Figure A4.6). Most people wanted to see improvement in data visualization, in particular 3-D visualization and contour plots, followed by statistical analysis (Figure A4.7). Regarding the question of how people use hydrologic information in their research, main responses included:

- Modeling such as hydrological, atmospheric, groundwater, and water quality modeling
- Calibration and validation of numerical models
- Ecosystem modeling (e.g., climate/plant interactions, relationship of species meta-population with water management, wetland dynamics, etc.)
- Watershed and river restoration

From the survey results, it was clearly indicated that the hydrologic information be required by a broad range of research applications. Therefore, CUAHSI HIS needs to consider researchers in related fields.

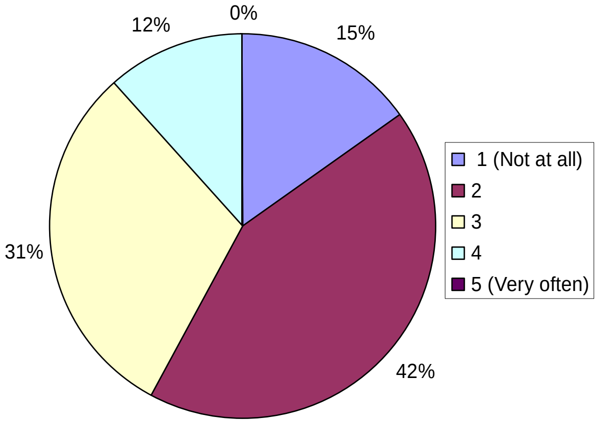


Figure A4.6. Survey participants' recognition of the need for a complicated hydrologic information system.

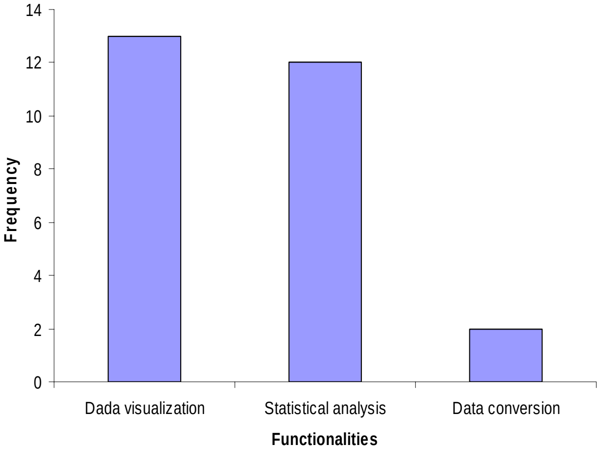


Figure A4.7. Most needed functionalities.

In the final section, we asked the participants about CUAHSI HIS to find its potential applicability at Berkeley. When asked whether they heard about CUAHSI HIS, one-third answered "yes." When asked about the expected infrastructure and services from CUAHSI HIS, main expectations from the participants were:

- Capability of data sharing (e.g., easy to ingest data into and to retrieve data from HIS)
- Standard data transferability (e.g., temporal and spatial resolution conversion)
- Support of various data formats (e.g., Ascii, Bin, HDF, etc.)
- Easy data configuration
- User-friendly cataloguing and indexing
- Service for people in other fields (e.g., ecology)
- Single interface, web-based data system
- Data visualization
- Basic statistical analysis functions
- Easy connection to other popular software programs (e.g., Matlab, Excel, GIS, Splus, etc.) for further in-depth analysis
- Open source approach
- Complicated data system is helpful but not necessary
- Prototype its integrated system, and receive feedbacks

From this list, the need for user-friendly cataloguing and indexing is notable which suggests that the user be able to know what is in the HIS when they access it.

A4.3 Pilot Survey at HIS Symposium 2005

The pilot survey was directed at those attending the HIS Symposium 2005 to focus on a subset of the CUAHSI membership most interested in the development of the HIS. The aim of the pilot survey was to refine questions based on internal surveys in order to improve the effectiveness of the web survey, which was later presented to the entire CUAHSI membership.

A paper survey questionnaire was distributed to participants in the conference information packet. Attendees were requested to fill out the survey during a break in the first day of the symposium.

Results

The pilot survey had 38 respondents from 23 different Universities and 3 different government institutions. A wide range of disciplines was represented at the Symposium (Figure A4.8). About half of the respondents represent disciplines outside hydrology or engineering, but use hydrology data for their research. The high number of computer scientists in attendance was due to interest in the development of the computer specifications of the HIS, a main thrust of the Symposium. The majority of the respondents were University faculty (57%), followed by graduate students (16%), university professionals (14%), working professionals (8%), and others (5%).

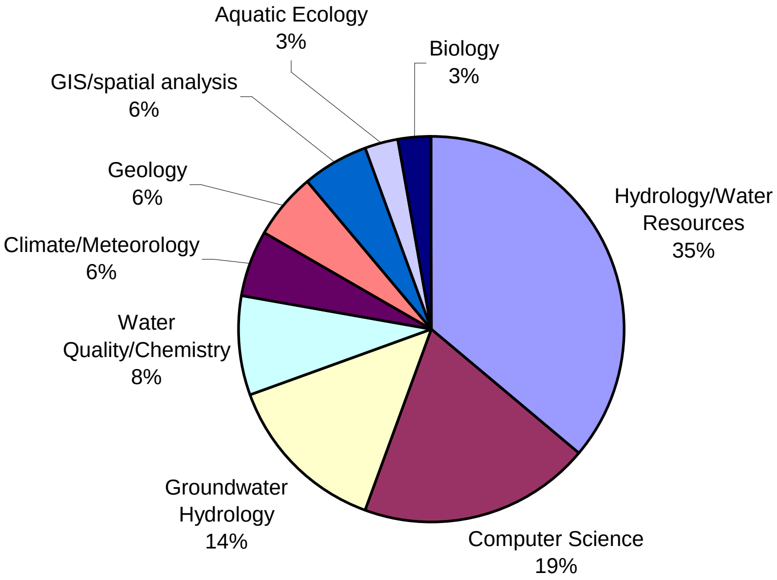


Figure A4.8. Specialties represented at the HIS Symposium 2005

Respondents were asked which operating systems they use for hydrologic research and what percentage of their time they spend using different operating systems.

Microsoft Windows was the preferred operating system: 36 of the 38 respondents use Windows an average of 75% of the time they are conducting their research. The remaining one quarter (25%) of research time is spent on using other operating systems (Table A4.1). Interestingly, only 11 of the respondents (30%) reported using Windows operating system 100% of the time.

Table A41. Operating system use by average percent of time used for research

Operating System	Number of Respondents	Average percent of research time
Windows	36	75%
Linux	15	25%
Unix	11	20%
Solaris	7	14%
MAC/OS X	5	54%
Other	2	18%

In order to prioritize software and programming languages utilized by the HIS, respondents were asked to rate the importance using a scale of 1 to 5 (where 1= Never use and/or do not find useful to your research, 2= Rarely use and/or do not rely on this, 3= Occasionally use and/or am comfortable with its use, 4= Often use and/or rely on for your research, 5= Frequently use and/or find indispensable). Microsoft Excel (4.0) and ESRI ArcGIS/Arcview (3.9) were the most used software and FORTRAN (3.3) and C/C++ (3.2) the most used programming languages. The other software and programming languages[4] we asked about averaged below the "occasionally use" range. The average results actually reflect the fact that most people "Never use" most of the software listed, while a few others find that same software "Indispensable".

Considering the datasets that the HIS could incorporate, respondents were asked the priority for including in the HIS on a scale from 1 to 5 where 5 is a high priority. Almost all of the datasets listed averaged above 4.0 as a priority for inclusion. Those that scored 4.5 or higher include: National Land Cover Data (4.7), Groundwater level (4.7), NCDC Precipitation (4.6), USGS Historical Streamflow (4.6), Water quality/Chemistry (4.6), National Hydrography Dataset (4.6), NEXRAD Radar precipitation (4.6), EPA STORET Water Quality Data (4.5), USGS Real Time Streamflow (4.5), SNOTEL data (4.5), and USGS National Geology data (4.5)[5].

For all the datasets listed, a score was given for ease of use. To help define the niche where the HIS can help researchers the most, we were interested to understand the intersection between the data priorities and the difficulty associated with using data. EPA

[4] Others listed in question: Java, MS Access, Visual Basic, Matlab, SQL/Server, Modflow, Adobe Illustrator, HEC models, GMS, WMS, SMS, R, SWAT, Sigma Plot, Surfer, SPSS, SAS, HSPF, S-Plus, GRASS, Visual Modflow, Mathematica, PostgreSQL, Tecplot, Groundwater Vistas, Kaleidagraph,

[5] Others listed included: Score 4.4 [National Elevation Dataset, Water use,Evapotranspiration, PRISM Precipitation Data, USGS Hydrologic Landscape Regions], Score 4.3 [SSURGO soils data, STATSGO soils data], Score 4.2 [LANDSAT Satellite Imagery], Score 4.0 [NCEP North American Regional Reanalysis (NARR) climate data, Real-time weather and Nexrad data from Unidata], Score 3.8 [University of Washington Gridded Meteorological Data]

STORET Water Quality Data (2.6) and SNOTEL data (2.6) were the most difficult to use of the high priority datasets. Those datasets scoring low on ‘ease of use’ (score of 2.5 or lower) included: Evapotranspiration (2.3), NEXRAD Radar precipitation (2.4), NCEP North American Regional Reanalysis (NARR) climate data (2.5), Water use (2.5), and Real-time weather and Nexrad data from Unidata (2.5), none of which were listed as the highest priorities for including in the HIS.

When asked to rate the priority of a list of HIS roles and system functionalities between 1 and 5 where 5 is a high priority, respondents did not select any of the options as a low priority. The highest priority roles for the HIS included (score above 3.5 average):

1. (4.7) Retrieval of relevant National, Community, and Hydrologic Observatory datasets
2. (4.6) Uploading, archival and sharing of hydrologic data with collaborators and the CUAHSI community
3. (3.8) Interfacing of hydrologic datasets in standard format with third party analysis software
4. (3.8) Development of community data models and standards for data representation

The highest priority HIS functions included (score above 4.0 average):

1. (4.5) Store and retrieve digital products from a hydrologic digital library
2. (4.4) Include GIS data on terrain, soils, land cover, geology, stream networks
3. (4.2) Allow connection to hydrologic models
4. (4.1) Include information from weather and climate models, and Nexrad
5. (4.1) Design metadata and develop tools for preparing it
6. (4.1) Support intelligent searching for hydrologic data, models, reports and papers
7. (4.1) Include remote sensing information
8. (4.0) Automatically harvest hydrologic observation data from agency websites

Interesting comments we received about the HIS project at the Texas Symposium included concern about data uncertainty, working with datasets at different scales, and including anthropogenic influences on the landscape. Others highlighted the need to create a system that is easy to use, has intuitive interfaces and is responsive to users. Comments also reflected that it is currently unclear to the community whether the HIS will be a data storage system, possibly replicating work of other agencies, or a data dissemination system that includes capabilities for data visualization, manipulation and analysis.

The pilot survey conducted in Austin was critical to identify problems with the question design and to focus the questions planned for inclusion in the web survey. One problem was the tendency for selection of ‘everything is important’ when respondents were asked which issues HIS should address and which datasets should be included. A change of format in the web version to a ranking question forced respondents to select the most important issues and datasets. The distribution of ranked responses gives much more information for HIS planning than a simple result that everything is important. Using an importance scale for software use (1=Never use or do not find useful; 2=Have used but do not rely on this; 3=Use occasionally and am comfortable with its use; 4=Use often; 5=Use frequently and find indispensable) was difficult to interpret. Averaged results did not represent the popular use of the software as well as a ranked choice of software most used by researchers. We also learned that not everyone was familiar with all of the long lists of datasets and software tools presented. Our attempt to simplify the questionnaire by presenting too wide a range of options resulted in off-putting some respondents. The information gathering and Symposium pilot surveys took an average of 15 minutes (with a range of 5-30 minutes), which was deemed too long for the web survey.

Chapter 5

Hydrologic Metadata

By Michael Piasecki, Luis Bermudez, Bora Beran, Saiful Islam and Yoo-Ri Choi
Department of Civil, Architectural & Environmental Engineering
Drexel University

Xu Liang and Seongeun Jeong
Department of Civil Engineering
University of California at Berkeley

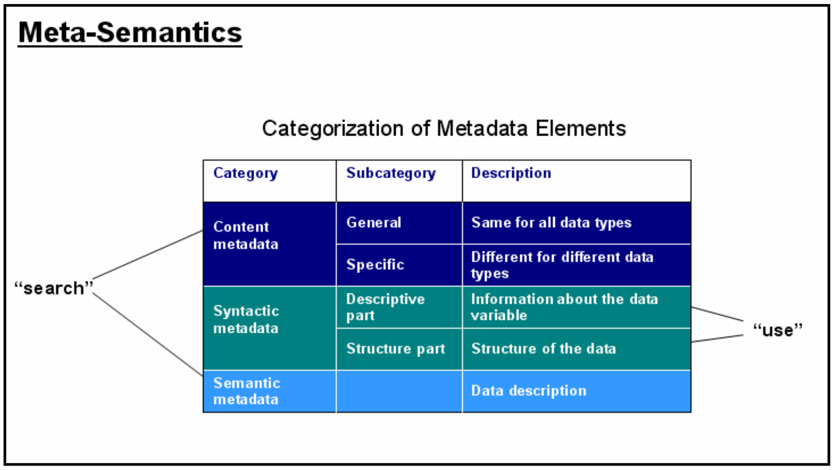
According to the task list in the HIS proposal Drexel University is responsible for developing a concept for a hydrologic community metadata profile. More specifically, the task was formulated as:

The task of the Task 3.1 Inventory and Metadata for Internet Data Sources in Hydrology

Task Leader: Michael Piasecki (Drexel University)
Collaborators: Ilya Zaslavsky (SDSC)
Duration: Month 1 to Month 24
This task will result in a CUAHSI research monograph available in printed and internet form which

- lists the variables describing water and the water environment,
- defines a metadata system which can be used to describe these variables for automated retrieval using the Storage Resource Broker system
- gives URL links to internet data sources describing those variables, and
- contains for each source and variable a summary of the data available at that source,
  - its extent in space and time,
  - the data format,
  - some history of how the data were developed,
  - a short description of how these data can be used in hydrology, and
  - a listing of published studies where this has been done.

Three quarters of the way through the project duration, it is an opportune time to identify the current status of this project task, point out what will be worked on and developed over the remainder of the project, and also indicate metadata related components that CUAHSI HIS needs to address in the future.


Meta-Semantics

Categorization of Metadata Elements

Category	Subcategory	Description
Content metadata	General	Same for all data types
	Specific	Different for different data types
Syntactic metadata	Descriptive part	Information about the data variable
	Structure part	Structure of the data
Semantic metadata		Data description

Figure 5.1 Metadata Categories

5.1 Introduction

5.1.1 Rationale

Metadata are used to describe any type of data set or, in a more general fashion, any type of Arbitrary Digital Object (ADO), of which data stored in files are a subgroup. Metadata typically need to fulfill a number of tasks which has led to a classification of metadata blocks used for different purposes, as shown in Figure 5.1. Typically the metadata description is divided into metadata components that help to search for a given data set, and components that are used to use the associated dataset for further processing. For example, it is important that a researcher is able to find specific data sets by using metadata elements that contain keywords describing the data set and also elements that describe what the time coverage of the data set is. On the other side, it is unlikely that that a dataset will be searched via elements that describe what units have been used or in what format the data is stored. Careful consideration must be given to how many metadata elements of each of the categories are included into the metadata description of the data set so as to fulfill all needs that exist with regard to finding and using the data set.

It is clear that there will be a large quantity of legacy data sets (for each HO) that need inclusion into the DLS. These datasets too need to be accommodated by the CUAHSI metadata profile. Also, there exists a large data world of data sets that have been and are in the process of being collected by federal and state agencies. While these data sets are of vital importance to the hydrologic community the stewardship of these data sets rest with others, and with this also the metadata descriptions used to describe these data sets. In essence, metadata will have to accommodate a large and diverse set of data that are collected by different agencies, are produced by numerical models and are collected by individual researchers in the field, as shown in Figure 5.2.

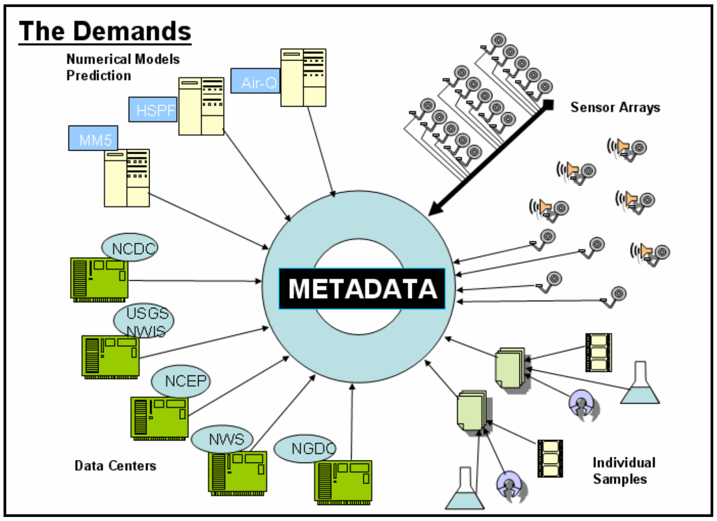


Figure 5.2 Metadata as Key to Data Access

Metadata descriptions are a crucial HIS CyberInfrastructure (CI) development task. In fact, metadata play a crucial role in the implementation and subsequent use of the Digital Library System (DLS) developed at the San Diego Supercomputer Center (SDSC). The application that permits access to the DLS, the HydroViewer interface, makes use of the metadata descriptions to identify appropriate data sets that have been requested by the user in the DLS through predefined search criteria like time brackets and data types. Given the fact that the

5.1.2 Objectives

The objectives of the metadata development are twofold. On the one side the creation of the hydrologic community metadata profile is a necessity that is based on the fact the future

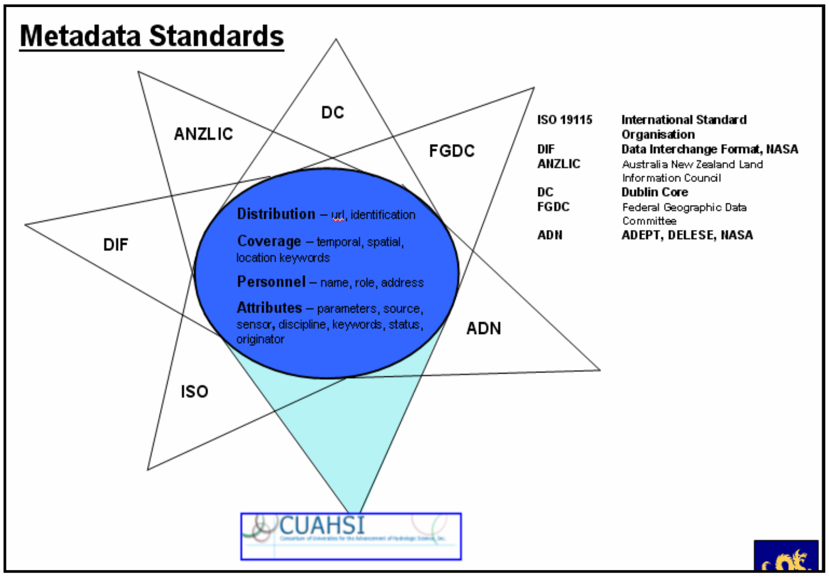
Hydrologic Observatories will collect a potentially large number of new data in the field that will originate from a large group of diverse sensors and sensor arrays, from grab samples, and from data that will be generated based on computations. These include forecasts, added value data based on numerical analysis of collected data, and data that is used to "fill" gaps in gridded data arrays. It is clear that the community must be provided with these metadata descriptions as otherwise newly generated data can neither be stored in the DLS, nor further processed to be included into the Digital Watershed database. In short, the community must be given a specific and common hydrologic metadata profile that will be used by all Digital Hydrologic Observatories that will emerge in the future.

The second objective is to implement this standard such that it has a chance to grow and develop in the future. What we mean by this is that the version_1.0_profile is unlikely to be the final word on the profile content, because i) the community will need to agree and gain experience (and modify as a result) on descriptions for a certain data set, and ii) will need to have the ability to expand the profile as new data sets become available, for example through deployment of new sensors and sensor arrays. But this is not all that is demanded from the implementation. For the idea of a one-stop data shopping to exist the CUAHSI metadata team will have to resolve the conflicts that result from disparate metadata profiles used by the large number of different data collection entities. In short, whatever the CUAHSI metadata profile will look like, it will look different from those being used elsewhere and as such need the deployment of technologies will prepare the CUAHSI metadata profile to interoperate with those existing elsewhere.

5.2 Metadata Technologies

5.2.1 Standards

The use of a metadata standard for the development of the CUAHSI profile is one step towards attaining uniformity for the metadata profile. The selection of a standard will permit the continued growth and expansion of the profile such that it can reach a level of maturity of time. The selection of a specific standard is driven by a number of aspects which include the coverage, extensibility, national and international acceptance and implementation. There are a umber of metadata standards in circulation worldwide as shown in Figure 5.2. Without going into specific details about the pros and cons of each standard, the CUAHSI HIS team at Drexel selected the ISO 19115


Metadata Standards

ISO 19115	International Standard Organisation
DIF	Data Interchange Format, NASA
ANZLIC	Australia New Zealand Land Information Council
DC	Dublin Core
FGDC	Federal Geographic Data Committee
ADN	ADEPT, DELESE, NASA

DC
ANZLIC
FGDC
DIF
ADN
ISO
Distribution – url, identification
Coverage – temporal, spatial, location keywords
Personnel – name, role, address
Attributes – parameters, source, sensor, discipline, keywords, status, originator
CUAHSI

Figure 5.3 Metadata Standards

standard because of four reasons:

1) it is the most comprehensive standard with the largest number of elements to choose from providing a very detailed coverage for a large number of diverse data sets. It also has provisions for extending the metadata descriptions in case the norm does not provide coverage for a specific area sought. If the extension rules are followed then these extensions are still ISO 19115 compliant.
2) It the most widely accepted standard internationally, with most of the other standards providing cross-walks fro the ISO 19915, signaling that the ISO norm is on its way to become the internally recognized benchmark standard.
3) For the US the FGDC sets a content standard, which in its 3 version is planed to either be the ISO 19115 norm or framework that is very close to the ISO 19115 standard. This means that many of the governmental data sets are likely to be described by metadata that either is fully or almost fully compliant and compatible with the ISO 19115 norm. It is a prudent choice for CUAHSI to acknowledge this fact as many of the data sources that are relevant for the CUAHSI community are in fact governmental entities.
4) Last but not least, the ISO 19115 is provided using the Unified Modeling Language (UML). The use of UML permits the transfer into other machine readable formats, like XML or XML schema, in an automated fashion. Also, many of the logical connections between metadata elements and their properties as realized in UML are lending themselves to be transferred to other machine readable formats like the Web Ontology Language (OWL).

It should be mentioned that there are also a number of markup languages (i.e. metadata profiles that have machine readable implementations in XML) in circulation, like the Ecological Markup Language (EML), the Geography Markup Language (GML), the Earth Science Markup Language (ESML), SensorML, and HydroML, all of which are suitable sources for developing the CUAHSI profile. In fact, the Drexel team has been using elements from the SensorML set to supplement the “use” metadata category for data collected from field stations.

Finally, participation in several initiatives and efforts to bring some order and interoperability to the disparate metadata descriptions currently in use or in development clearly point towards the need to use standards to overcome some of the main hurdles for exchanging data between and among communities. In this regard, the CUAHSI HIS is on the right rack by using an internationally accepted standard for metadata descriptions.

5.2.2 Implementation

All published metadata standards appear in a form that is not readily usable for the development of a community profile, i.e. they are in written form, in plain ASCII, or in UML, none of which is suitable for machine readability or the ability to be parsed for information. Hence, there is the need to expend some effort to identify what implementation strategy should be used for casting the CUAHSI metadata profile.

There are several options to implement the profile in digital format. They reach from plain ASCII files with no specific format (like the DIF) to documents that use the Extensible Markup Language (XML). Most of the currently implemented metadata profiles are using the XML

Schema as a means to provide a template for metadata instance creation, i.e. metadata files that are written and stored in XML. The Drexel team has looked at this alternative quite carefully for its applicability to the CUAHSI metadata profile. However, there are several shortcomings for using XML and XML Schema that have let the team to choose a mixed path for implementation.

The Drexel team selected both the XML schema and the OWL language as an appropriate means to implement the CUAHSI profile. There are several reasons for doing this.

1) XML schema offers itself as an appropriate means o define a reliable syntactic structure of the metadata profile. While there are shortcomings in the schema structure like the lack of the inclusion of the Unique Resources Identifier (URI) concept by which information, like the CUAHSI metadata profile, can be made available on the web as a resource that can be used by others in machine readable format, it can be overcome with concepts and applications like Xlink. The use of a schema ensures a repeatable process by which compatible instantiations of the metadata profile are ensured.
2) The use of OWL permits the inclusion of a much richer semantic environment then is possible in a XML schema alone. This has ramifications for how controlled vocabularies can be incorporated into a metadata profile. OWL documents can be linked into XML schema so as to provide for a specific set of terms or keywords as the only permissible entry that should be used with a given metadata element. This provides an environment that can restrict entries to specific elements, i.e. effectively limiting the range of a metadata element of attribute.

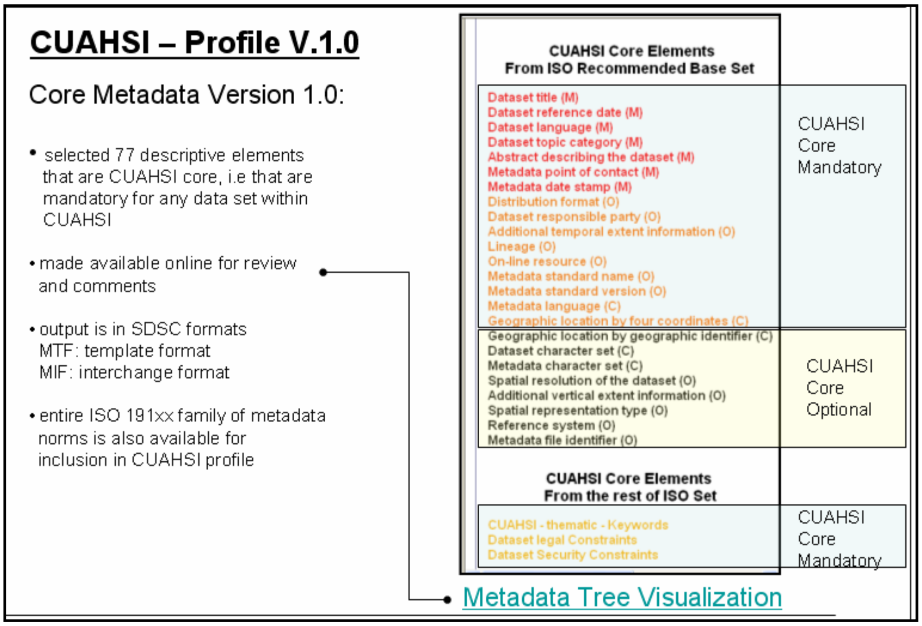
Finally, with the advent and continued growth and acceptance of the Semantic Web, which promotes the idea of making any type of complex information available on the WWW via registering this information as resources, the Drexel team took on the view that the CUAHSI metadata descriptions should become part of this system in the future. We believe that this is of particular importance for achieving interoperability among different communities providing a first step towards overcoming the heterogeneity in data descriptions that currently exists.

5.3 Current Status of Profile

The CUAHSI metadata profile version_1.0 is currently being developed and posted at the Drexel project website at http://loki.cae.drexel.edu:8080/web/how/me/metadatacuahsi.html together with some of the ontologies (in OWL) that we have created. These include some of the standards that need to be re-written as well as some conceptual representations that are relevant to the CUAHSI community like the USGS hydrologic unit system. In view of the requirements set forth by SDSC that metadata should be organized in a canonical system, the Drexel team devised a system that categorizes different metadata blocks into functional elements. This is also very much in tune with the ISO recommendation of identifying specific blocks of metadata that should be mandatory or optional. In fact, the CUAHSI metadata profile consists of a number of sub-profiles that can be combined to yield a specific set of metadata elements for a given data set. The files can be saved in any desired format. In fact, any format is possible reaching from comma or tab delimited formats, to XML documents, to any proprietary format (like the DIF format by NASA), to OWL documents, to the Resource Description Framework (RDF) format, the latter being ideally suited for exposure and accessibility on the WWW.

5.3.1 Core Elements

The ISO 19115 norm suggests a base set of elements that should be considered when describing a geospatial data set or feature. It consists of 24 metadata elements that have a varying number of properties. It is these properties that eventually need an entry whenever an instance for a metadata description is created. In the ISO base, 7 elements are set as "mandatory", i.e. these are not negotiable and must be used whenever the ISO 19115 norm is deployed to describe a geospatial data set. The base set also contains 12 "optional" elements, i.e. elements that may or may not be used. Finally, there 5 "conditional" elements which are of the type: either use this one or use another one, i.e. they are in nature somewhat closer to mandatory elements.



CUAHSI – Profile V.1.0

Core Metadata Version 1.0:

- selected 77 descriptive elements that are CUAHSI core, i.e that are mandatory for any data set within CUAHSI
- made available online for review and comments
- output is in SDSC formats
  MTF: template format
  MIF: interchange format
- entire ISO 191xx family of metadata norms is also available for inclusion in CUAHSI profile

CUAHSI Core Elements
From ISO Recommended Base Set

Dataset title (M) Dataset reference date (M) Dataset language (M) Dataset topic category (M) Abstract describing the dataset (M) Metadata point of contact (M) Metadata date stamp (M) Distribution format (O) Dataset responsible party (O) Additional temporal extent information (O) Lineage (O) On-line resource (O) Metadata standard name (O) Metadata standard version (O) Metadata language (C) Geographic location by four coordinates (C)	CUAHSI Core Mandatory
Geographic location by geographic identifier (C) Dataset character set (C) Metadata character set (C) Spatial resolution of the dataset (O) Additional vertical extent information (O) Spatial representation type (O) Reference system (O) Metadata file identifier (O)	CUAHSI Core Optional

CUAHSI Core Elements
From the rest of ISO Set

CUAHSI - thematic - Keywords Dataset legal Constraints Dataset Security Constraints	CUAHSI Core Mandatory

Metadata Tree Visualization

Figure 5.4 CUAHSI Core Metadata Elements

The ISO recommended base set is, excluding the mandatory elements, not more than a pre-selected group of elements that are highly recommended for consideration. As a result, the CUAHSI core set contains all 7 mandatory ISO elements in addition to 9 elements that have been selected from the conditional&optional group. These have been selected and added based on extensive discussions among group members, discussions with other metadata profile developers that have used the ISO 19115, and what the Drexel team saw as necessary given the data sets the CUAHSI community is likely to be interested in. There are also 3 additional metadata elements that have been selected from other branches of the ISO 19115 to complete the current CUAHSI core set, i.e. CUAHSI keywords, legal constraints, and security constraints. These were added based on discussions that concerned the timing of access rights for all versus the collectors, and also because of raised concerns regarding the security status certain data should have. All other elements are optional within the CUAHSI framework, as shown in Figure 5.4. The above system results into 88 attributes, i.e. metadata entries that need to be supplied for every data file that is to be added to the DL system. While this may seem to be excessive, notice that about 70% of the information required is contact information that is associated with the individuals who collected the data, curated it, or supplied the metadata for it. This type of information can easily be supplied via an automated metadata system that makes use of login profiles and as such will greatly diminish the associated work of the individual researcher.

5.3.2 Minimum Description Set(MDS)

It is clear from the above number (88) that this core set is likely to be too large for a single user to supply this information by hand. In addition, discussion within the community revealed that the entire set of metadata is likely not to be what a user might want to see at a first glance. As a result of these discussions it was agreed to identify a MDS that would i) serve the need to provide a quick glance on a data set based on information slots that are likely to be high on a list of priority information, and also ii) to provide a small enough number of entry fields that a potential user of this system will not be deterred to use it, minimal amount of hands-on work to get a data set into the DL.

The MDS is comprised of two sections: the first constitutes metadata that is needed for **Search**, and the second contains elements that a user might need for **Identification** and as such should be displayed together with the **Search** components to form the MDS. We have identified 4 search and 6 identification elements:

Table 1 Minimum Description Set (MDS) for CUAHSI Metadata Profile

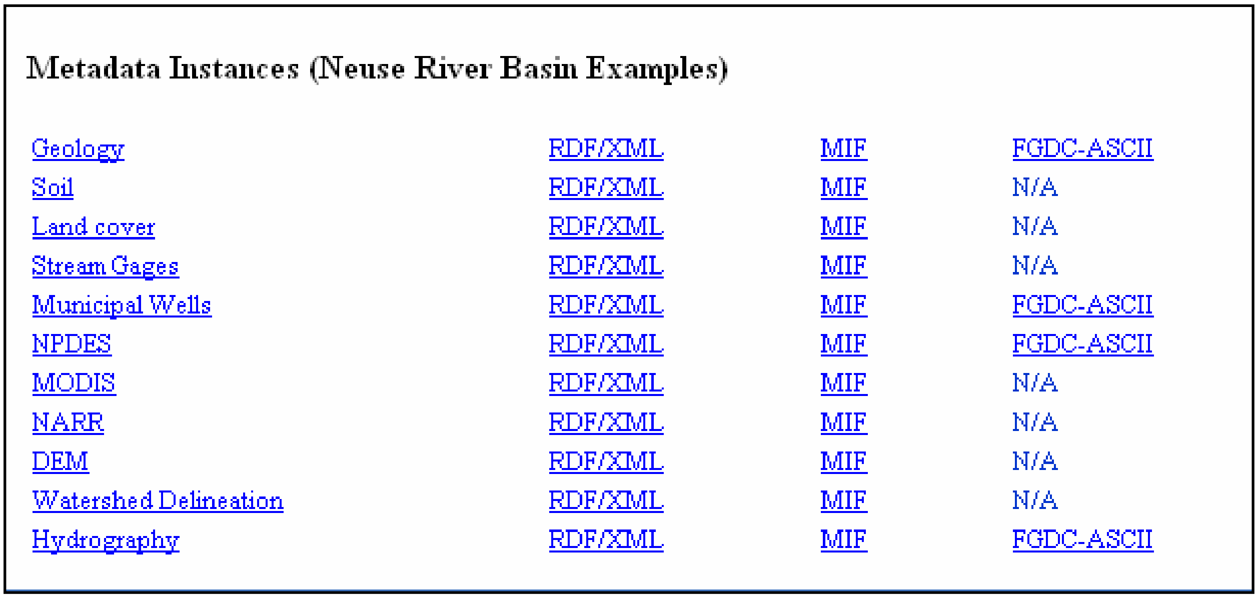
Search	Identification
1) Publisher (e.g. USGS)	**5) Title of Data set**
2) Subject keyword (Streamgauge)	**6) Description**
3) SpatialCoverage (4 Lat/Lon pairs for BBox)	7) Download link of data file (auto)
4) TemporalCoverage (from / to)	8) Download of full metadata set (auto)
	9) File format/resource type (auto)
	10) File size (auto)
	11) Access control of permission (default)
	12) Last successful update (auto)

Notice that the elements in bold are those that have the highest potential of having to be provided via a hand-entry, i.e. constitute a good portion of the real work for a user who wants to register her/his data file in the CUAHSI system. This MDS is currently under review and will be updated as needed. It may also be used as first metadata set to get some of the initial Digital Libraries for the Digital Hydrologic Observatory teams started.

5.3.2 Additional Elements

The canonical metadata system deployed by the SDSC team allows one to minimize the number of metadata elements used to describe any given data set or ADO by combining selected blocks of metadata that are tailored for adequate description of that specific ADO. All metadata descriptions regardless what type of ADO will have to contain the CUAHSI core set at the top level, but can then be supplemented with topic specific metadata blocks stored in the MTF format. Because there are no data sets that are being produced by the CUAHSI community within the Hydrologic Observatories, the HIS team is using the Neuse River in North Carolina as a test-bed for the metadata developments.

Even though the data availability is somewhat limited in scope (not in quantity though) in so far as it addresses data files that are in GIS format or data sets that originate from federal collection efforts and as such already have a metadata description (that of the collecting agency), these descriptions are in many cases incomplete and therefore lend themselves well for being used as test cases. The HIS team agreed early on to use 4 different data types to first populate the Digital Library system: i) GIS type data, ii) point measurement data (e.g. gauges for streamflow and rainfall), iii) flux data (from the NARR program) and iv) remote sensing data (MODIS). The Drexel team has created 12 example metadata description files that have been selected from the above 4 main categories as shown in Figure 5.5.



Metadata Instances (Neuse River Basin Examples)

Geology	RDF/XML	MIF	FGDC-ASCII
Soil	RDF/XML	MIF	N/A
Land cover	RDF/XML	MIF	N/A
Stream Gages	RDF/XML	MIF	N/A
Municipal Wells	RDF/XML	MIF	FGDC-ASCII
NPDES	RDF/XML	MIF	FGDC-ASCII
MODIS	RDF/XML	MIF	N/A
NARR	RDF/XML	MIF	N/A
DEM	RDF/XML	MIF	N/A
Watershed Delineation	RDF/XML	MIF	N/A
Hydrography	RDF/XML	MIF	FGDC-ASCII

Figure 5.5 Example Metadata files for the Neuse River

The list shows metadata description files that have been saved both in RDF/XML and MIF format, which can be downloaded for inspection and subsequent use. The list also shows files that contain the original metadata in case it was provided together with the ADO. These are based on FGDC type metadata and provided in plain ASCII format as well for comparison.

5.3.3 Controlled Vocabulary

Controlled vocabularies (CV) are an essential component for defining good metadata. This is a true semantic problem, i.e. while metadata elements define part of the syntax of a profile, it is the group of permissible entries when creating metadata instances that contributes to a coherent description of data files or ADOs. IN fact, missing or incompatible controlled vocabularies between communities is one of the main causes for lack of interoperability among communities and sometimes even within a community. For example, the keywords used for identifying water elevation measurements in a stream vary between different agencies who deploy keywords like “stream_gauge” and “gauge_height” to label the very same measurement.

Above scenario constitutes a problem of heterogeneity and cannot be readily resolved. As a result CUAHSI will need to develop its own CV to be used within the DLS and also the DWS. It is also clear that the CUAHSI CV will by default not be interoperable with all the other CVs that are currently being used most notably that of USGS-NWIS, NOAA-NWS, NCAR-UNIDATA, or EPA-STORET. As a result the Drexel team will have to address two objectives. The first objective is to develop a CV for CUAHSI that will be used by the DLS to store the data sets generated by the digital hydrologic observatories and those data sets that will be downloaded to the DLS from other sources. The second objective is to use an implementation of the CUAHSI CV that provides the means to make it interoperable with other already existing CVs elsewhere in order to resolve the semantic heterogeneity that will exist as an inevitable fact.

This is a formidable task and will require a diligent and focused effort. The Drexel team has submitted a first version of the CV to the SDSC so a preliminary deployment for the two DLS test sites, i.e. Neuse River and St. Margarita, could be set up. This Controlled Vocabulary is largely based on the UNESCO Glossary of Hydrologic Terms and represents a scaled down version of the Controlled Vocabulary. This particular Controlled Vocabulary was selected from a dozen or so alternatives for its immediate applicability and relatively narrow focus on terms. It is clear however that this is really just a first version that needs further extension.

5.3.4 Review

The current metadata profile is under review by an expert group of collaborators (lead by Dr. Xue Liang) within the extended CUAHSI HIS group. The goals of this effort are:

i) to review existing metadata profiles used by important and relevant data sources for the CUAHSI community in order to examine their practices and relevance for the CUAHSI profile

ii) to review and examine the metadata components defined so far from the ISO 19115 and other related ISO norms with particular emphasis on the completeness of the descriptions for selected sets of data (see above list in 5.3.2)

iii) to review the appropriateness of the selected technologies used, i.e. is the ISO 19115 norm adequate? Is the use of XML , RDF, and OWL an appropriate means to encode the CUAHSI metadata profile? Is the structure of the chosen metadata blocks reflecting the needed components to sufficiently characterize the data sets?

iv) to review concerns the presentation of the metadata to the public user. It is of particular importance that the user community can be exposed to this profile such that it is comprehensible for the non-metadata expert, that the profile is presented such that it can be reviewed and feedback comments provided, and that it is embedded in a framework in which sufficient information can be gleaned as to what an a specific metadata element is for and it will be incorporated into a metadata description for an ADO.

These aspects are very relevant to the usability of the metadata profile for the user community, i.e. the ability to actually supply metadata information in an environment that is easy to understand, user friendly, and designed such that dealing with metadata is a positive experience enhancing the prospect of acceptance by the CUAHSI community. There is no specific time line set for the completion of this review, however item i) in the above list was completed recently and a summary has been provided below.

Summary for Case i):

The four data sources that reviewed provide metadata based on the Federal Geographic Data Committee (FGDC) standards. Even though the FGDC standards are used, each data provider applies the standards differently. Oftentimes, the metadata are not published directly by data providers. Instead, individual owners or groups maintain metadata for the original data providers. For instance, the STORET program of EPA does not provide metadata although it provides supplementary information about the data. Rather, relevant or participating organizations maintain their own metadata.

Most data providers do not elaborate on essential elements of metadata that need consistency. This may be due to the different focus of the data providers who pay more attention to, at present, the basic information, such as who collected the data and how the data are distributed, rather than the data structure and contents to help users understand how data are organized. For example, the /Entity_and_Attribute_Information / element of the FGDC standards can be used not only to convey information about data structure and content, but also to run data systems (e.g., data retrieval system, database, etc.). Most data providers lack the description of this important element. Also, the four data providers do not relate metadata to system operations such as data integration or data mapping. To support flexible use of metadata and utilize metadata for operations such as data integration, it may be desirable to categorize metadata into core metadata and supplementary metadata. The core metadata (i.e., system integration metadata) that contain data identification information and data structure may be used to support system-level operations such as data integration between different systems. Whenever necessary, supplementary elements, such as distribution, metadata reference elements, etc., may be added to the core system.

5.4 Future Needs

The project has achieved a number of significant goals, as pointed in the above sections. However, the development of HIS for the CUAHSI community is a research task in itself and while the tasks for year two are emerging (those are a continuation of what has been started and also newer tasks not previously stipulated in the original proposal) there are a number of major developments whose need have become apparent as a result of the ongoing efforts. These new needs differ in scale and scope and may be addressed in part during the second project year. However, it should be clear that not all of this can be addressed given the current resource allocation and time left on the project. In a sense, the following sections outline a future task list that need to be dealt with in subsequent efforts.

5.4.1 Metadata Access System

During the development of the metadata profile it became clear that tools need to be developed that will aid the developers to visualize and to select the metadata elements necessary for any given data file (ADO). This is a necessary step because the ISO 19115 metadata norm is quite complex in its structure and can only be navigated efficiently if it is cast in a machine readable format, much like a WORD document. Because the norm has been implemented in OWL it offers itself to be viewed via the Protégé 2000 tool. However, this tool is only efficient for creating ontologies and viewing them, not however for marking elements as optional or mandatory. Also it is not designed to accept code lists as possible entries when creating instances.

While developers typically bring enough expertise to the task and can help themselves by going to a steep learning curve how to use a tool like Protégé, it is apparent that the end users who seek to create a metadata description for a given ADO need a much more user friendly environment in which metadata instances can be generated. It is particularly important that users who do not typically deal with the details of metadata generation, storage and so on, are exposed to a positive experience when dealing with metadata as otherwise the real danger exists that metadata generation will be viewed as cumbersome and tedious making it extremely difficult to reach acceptance of the need to provide metadata descriptions. The motto here is: as much automation as is feasible, as little "hands-on" work as possible.

Even though the response from the metadata review group has not yet been generated, it is clear that a considerable effort needs to be invested into how the CUAHSI metadata profile will be presented, can be visualized, will be explained, and can receive community feedback in the future. Also, the maintenance of the profile should preferably take place in specifically designed application that allows easy access, upgrading, and versioning of the profile. This includes an application that would permit to easily create metadata mapper or crosswalks to other metadata profiles elsewhere in the community and across to neighbor communities.

Finally, the access system should include an automated metadata generation application. This application would select pre-defined metadata entries from a sensor profile database and combine those with actual information that arrives from the sensor data harvester (timestamp, location,

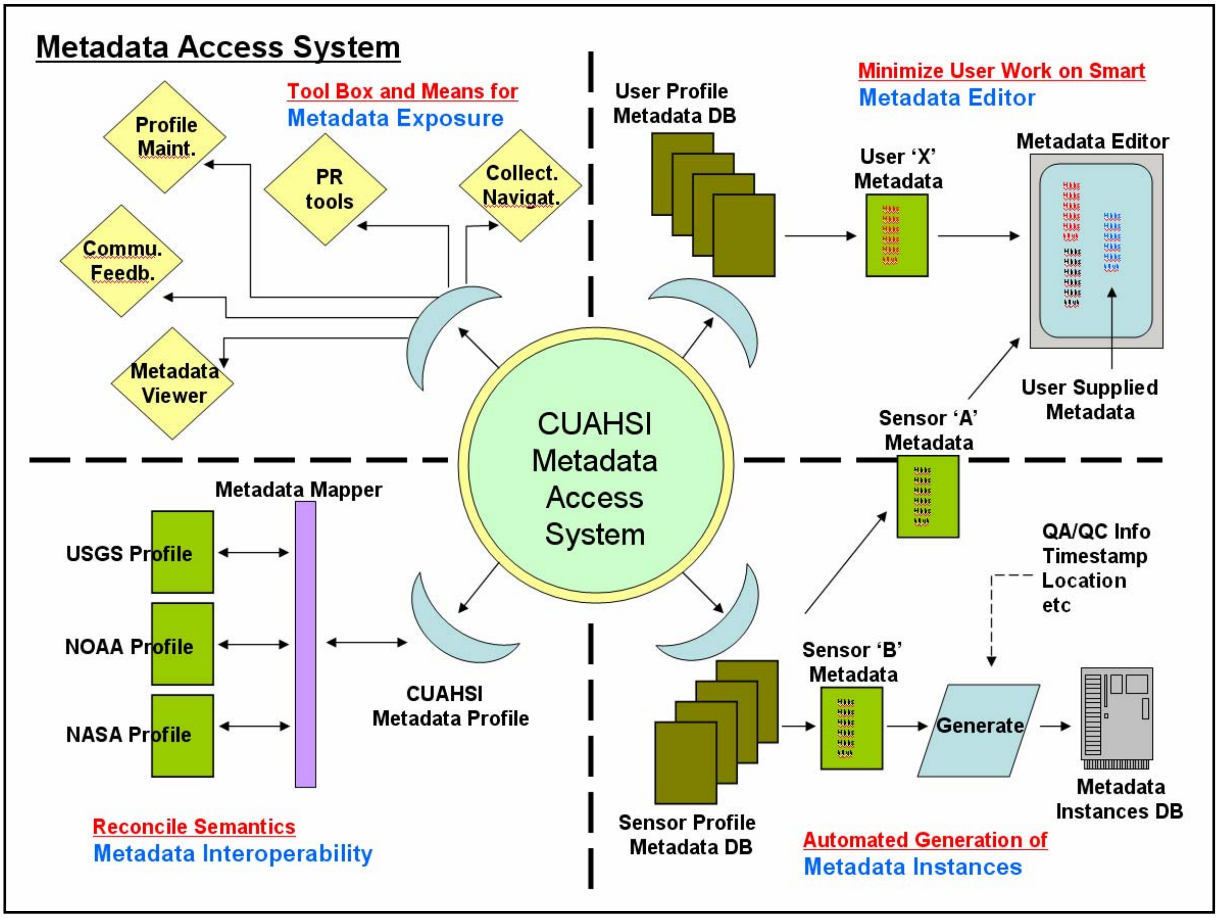


Figure 5.6 Schematic of CUAHSI Metadata Access System

QA/QC, etc) to generate a fully automated metadata instance that can be stored in the metadata database of the DLS. This should greatly reduce the effort of metadata instances generation and also make the maintenance of the sensor profiles a straight forward task. A graphics that summarizes all tasks of the Metadata Access System (MAS) is shown in Figure 5.6.

5.4.2 Controlled Vocabulary

In the remaining project time the Drexel team will as much as possible expend considerable effort to develop a more comprehensive Controlled Vocabulary that will be extracted from the UNESCO Glossary but also based on other more comprehensive collections of area specific terms. Good targets are the Global Change Master Directory (GCMD) for which the team has established contact with Lola Olsson who is the responsible person at NASA for its stewardship. Another very good source is the SWEET effort headed by Rob Raskin at the Jet Propulsion Lab, that cast the GCMD taxonomy of terms into a framework of smaller ontologies (using OWL) and also rearranging the terms according to phenomena, living and non-living things, and properties and entities rather then by topic as is done in the GCMD. We will also examine the usability of other thesauri and glossaries for this purpose and attempt to extract as much useful information as possible from those sources as well.

The ultimate goal of this task is to i) define a comprehensive Controlled Vocabulary for CUAHSI, and ii) cast the taxonomy of terms into an OWL based ontology such that it can be used to interoperate with other OWL based implementations of external Controlled Vocabularies. The team may take on the task of casting other CVs into OWL as well so it can be demonstrated that the chosen approach is viable and establishes the ground work upon which future developments or modifications can be based. It is also of importance to identify an appropriate mark-up language for identifying specific standard names (like the CF in netCDF) and also a vocabulary (keyword lists) that can be used for discovery purposes (more like as in GCMD). The purpose of these vocabularies is slightly different and the Drexel team needs to carefully examine what the needs are for the metadata descriptions as well as the needs for the HydroViewer environment. It must be stressed that this work represents cutting edge research, which requires a considerable amount of effort and is likely not to be completed within the current project duration. This topic however is of crucial importance as it is emerging as a possible path to overcome interoperability problems.

5.4.3 Hydrologic Ontologies for the WEB (HOW)

The development of the metadata profile prompted the Drexel team to use ontologies as a base technology to ensure the potential of future interoperability between metadata profiles. We have used these ontologies to implement the metadata profile by using both OWL implementations of the metadata standard and also for the CV or keyword list that define the permissible entries for the metadata descriptors. During the course of this work it became apparent that ontologies have the potential of providing a means to represent knowledge in a much more comprehensive way than currently envisioned, i.e. through the use of the DLS and the Digital Watershed. While some concepts of the Digital Watershed lend themselves to be mapped into an ontology, the goal of creating a hydrologic ontology framework requires a much more concentrated effort.

In fact, one of the components of the funded effort is to establish a formal interface between the HydroViewer system (DSL) and the GEON system. GEON is based on an innovative concept that has the potential to provide the GeoSciences with an umbrella framework. Hence the hydrologic community should undertake an effort to register a specifically designed ontology framework with GEON that would enhance interoperability between hydrology and neighboring sub-disciplines like geology and atmospheric sciences, beyond what is currently being envisioned in the CUAHSI-HIS project.

One crucial task is to define what a “top level” or “upper” hydrologic ontology could look like. This “upper ontology” would serve as a backbone for other more detailed ontologies that would link into the backbone. Several alternatives come to mind. One of them is to divide the hydrologic realm vertically into an atmospheric layer, a surface water layer and a sub-surface (groundwater) layer. Another alternative is to identify spatial features along which the water travels, for example a vertical water budget. Other topics include the idea to use a taxonomy of hydrologic terms to construct the hydrologic ontology. Another alternative is to sub-divide the hydrologic realm into a small number of key areas of interest and then expand these key areas with an appropriate number of detailed ontologies to incorporate various aspects and concepts.

As a result, the HIS group should embark on this track and try to continue to explore technologies that not only help define how the hydrologic community can access data but also to implement concepts of discovery and deduction into these data sets that will help the hydrologic researcher in his task to identify appropriate or necessary data, information, and even knowledge.

5.5 Summary

Drexel University is tasked with developing a first version of the CUAHSI metadata profile that is intended to be used by the Hydrologic Observatories (and other researchers as well who want to deposit their data files with the CUAHSI organization) to accurately describe the data sets generated in the HOs. To tackle this task the team first explored what metadata norm to use and also investigated similar mark up languages for their applicability to the given task The team decided on the use of the ISO 19115:2003 metadata norm because of its widespread acceptance, broad coverage, implementation in UML, as well as its ability to be extended quite easily while still being compliant with the overall norm.

The metadata profile has been established in its first version to cover a number of base descriptions that are either mandatory or optional for any data set that will be described as part of the CUAHSI metadata world as well as data set specific blocks that will to be added to the base set of descriptors. In lieu of a set of HO data sets that could be used to test the profile (the HOs are not yet in existence) the group focused on the description of data sets that can be found at a national level regardless of where an HO might be established in the future. These data sets are taken from the prototype HIS watershed (Neuse River) in order to demonstrate their applicability. These data set descriptions have been made available on the project website and are also have been transmitted to the DLS developers at the SDSC so they can be used to populate the prototype DLS. These descriptions are stored in MTF and MIF formats as requested by the SDSC.

The group is also working on the development of a hydrologic CV that can be used in conjunction with the DLS to better discover, navigate, and uniformly describe the data sets in the hydrologic realm. The development of the CV and its subsequent implementation in an OWL ontology will also set the stage for future interoperability between different data set descriptions, i.e. they will help to resolve the semantic heterogeneities that currently exist between the many entities that collect and disseminate hydrology relevant data. This is work is also preparing the HIS group to scope out the future task of developing a Hydrologic Ontology for the WEB, or HOW.

Chapter 6

Hydrologic Observations Data

Jeffery S. Horsburgh[1], David G. Tarboton[1] and David R. Maidment[2]

Abstract

The CUAHSI Hydrologic Information System project is developing information technology infrastructure to support hydrologic science. Part of this includes a data model for the storage and retrieval of hydrologic observations in a relational database. The purpose for a hydrologic observations database is to store hydrologic observations data in a system designed to facilitate data retrieval for integrated analysis of information collected by multiple investigators. It is intended to provide a standard format to facilitate the effective sharing of information between investigators and to facilitate analysis of information within a single study area or hydrologic observatory, or across hydrologic observatories and regions. The hydrologic observations data model is designed to store hydrologic observations and sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and used and provide traceable heritage from raw measurements to usable information. A relational database format is used to provide querying capability to facilitate data retrieval in support of a diverse range of analyses. An initial data model design was presented at the CUAHSI Hydrologic Information System Workshop held in Austin during March, 2005. An independent review of this initial design identified significant issues that needed to be addressed. This paper presents a redesign of this data model that addresses these issues, to the extent possible within the scope of a relational database model, for the storage and retrieval of point observations.

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) is an organization representing more than 100 universities sponsored by the National Science Foundation to provide infrastructure and services to advance the development of hydrologic science and education in the United States. The CUAHSI Hydrologic Information System (HIS) project's purpose is to improve infrastructure and services for hydrologic information acquisition and analysis. The project is examining how hydrologic data can be better assembled and analyzed to support hydrologic science and education. As presently conceived, the CUAHSI Hydrologic Information System has four components (Figure 1):

- a *Hydrologic Observations Database*, which is a relational database containing observational data on streamflow, climate, water quality, groundwater levels, and other data measured at monitoring points;

[1] Utah Water Research Laboratory, Utah State University
[2] Center for Research in Water Resources, University of Texas at Austin

- A *Digital Watershed*, which synthesizes the Hydrologic Observations Database with GIS data, weather and climate grids and remote sensing data to form a comprehensive depiction of the water environment of a hydrologic region;
- A *Hydrologic Analysis System*, which supports analysis of fluxes, flow paths, residence times, and mass balances on the Digital Watershed;
- A *Hydrologic Digital Library*, which stores and provides internet access to digital products from all parts of the Hydrologic Information System.

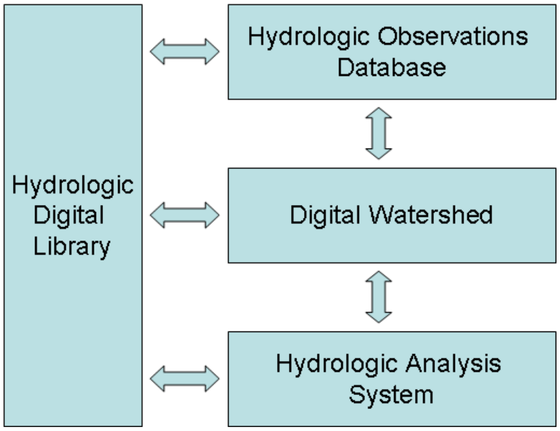

Hydrologic Digital Library
Hydrologic Observations Database
Digital Watershed
Hydrologic Analysis System

Figure 1. CUAHSI Hydrologic Information System Components

The hydrologic observations data model is the template for the Hydrologic Observations Database and is designed to store hydrologic observations and sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and used. The metadata will also provide traceable heritage from raw measurements to usable information. A relational database format is used to provide querying capability that facilitates data retrieval in support of a diverse range of analyses. Reliance on databases, and tables within databases also provides the capability to have the model scalable from the observations of a single investigator in a single project, through the multiple investigator communities associated with a hydrologic observatory ultimately to the entire set of observations available to the CUAHSI community.

The hydrologic observations data model is focused on hydrologic observations made at a point. Remotely sensed image or grid data is explicitly excluded as it is handled separately as part of a digital watershed distinct from the hydrologic observations database. Furthermore, information synthesized or derived from raw observations is also excluded, except for simple transformations essential to get the data into a useable form, such as conversions from water level to discharge through a rating curve at a stream gage, transformations from measured voltage to a physical quantity at a probe or instrument, or aggregations from high frequency observations to a desired time step. Synthesis and the derivation of other information and products from hydrologic observations is the role of the Hydrologic Analysis System.

Hydrologic Observations

Many organizations and individuals measure hydrologic variables such as streamflow, water quality, groundwater levels, and precipitation. National databases such as USGS' National Water Information System (NWIS) and USEPA's data Storage and Retrieval (STORET) system contain a wealth of data, but, in general, these national data repositories have different data formats, storage, and retrieval systems, and combining data from disparate sources can be difficult. The problem is compounded when individual investigators are involved (as would be the case at proposed CUAHSI Hydrologic Observatories) because everyone has their own way of storing and manipulating observational data. There is a need within the hydrologic community for an observations database structure that presents observations from many different sources and of many different types in a consistent format.

Hydrologic observations are identified by the following fundamental characteristics:

- The location at which the observations were made (space)
- The data and time at which the observations were made (time)
- The type of variable that was observed, such as streamflow, water surface elevation, water quality concentration, etc. (variable)

These three fundamental characteristics have been represented by Maidment (2005) as a data cube (Figure 2), where a particular observed data value (D) is located as a function of where it was observed (L), its time of observation (T), and what kind of variable it is (V), thus forming D(L,T,V).

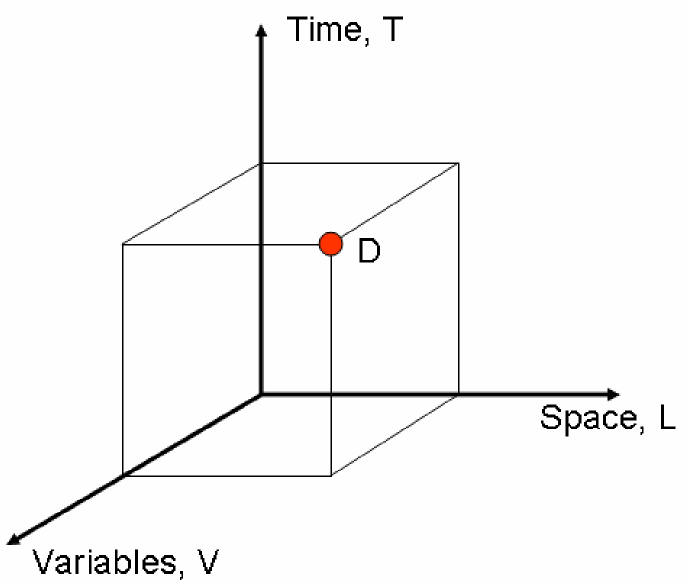


Figure 2. A measured value (D) is indexed by its spatial location (L), its time of measurement (T), and what kind of variable it is (V) (Maidment, 2005).

In addition to these fundamental characteristics, however, there may be many other distinguishing attributes that accompany the observational data. Many of these secondary attributes provide more information about the three fundamental characteristics mentioned above. For example, the location of an observation can be expressed as a text string (i.e., “Bear River Near Logan”) or as latitude and longitude coordinates that accurately delineate the location of the observation. Other attributes can provide important context in interpreting the observational data. These include data qualifying comments and information about the organization that collected the data. One of the fundamental design decisions associated with the HOD is how much supporting information to include in the database. This will be discussed in further detail in subsequent sections of this paper.

The ArcHydro Time Series Data Model

In March of 2005, the ArcHydro Time Series Data Model was proposed as a starting point for the HIS HOD structure (Maidment, 2005). This was closely modeled after the time series data model used in ArcHydro (Maidment, 2002). An independent review of this design was undertaken to evaluate whether the ArcHydro Time Series Data Model is adequate to meet the needs of the CUAHSI community and serve as the HIS HOD structure (Tarboton, 2005). Review comments and input were widely requested from scientists familiar with the CUAHSI HIS, from CUAHSI hydrologic observatory planning groups as potential users of the HIS, and from others knowledgeable in database design and dissemination of data. A total of 22 individual sets of review comments were received, and in general the respondents believed that the ArcHydro Time Series Data Model was a good starting point, but that it fell short of providing adequate information to serve as the CUAHSI HIS HOD structure. In addition to comments about the organization and content of the tables in the database, the following is a summary of some of the most important comments and observations that were received as part of the review:

1. In the ArcHydro Time Series Data Model, there is inadequate information to identify the source, heritage, or provenance and give exact definition of the data.
2. The ArcHydro Time Series Data Model does not provide enough information to fully spatially locate a measurement.
3. It is important that the scale of the measurements, defined in terms of their support (averaging domain), spacing, and extent be quantified and associated with measurements.
4. The ArcHydro Time Series data model does not include depth or vertical offset information associated with observations.
5. The ArcHydro Time Series Data Model does not account for censored observations.
6. The classification of time series data types needs to be extended and modified to provide information that guides appropriate interpretation of the data, such as whether the measurements are continuous so that operations such as aggregation or interpolation are meaningful.
7. The focus of the ArcHydro Time Series Data Model on a favored set of proprietary software raised concerns with some reviewers.
8. The ArcHydro Time Series Data model does not include an indication of the quality of the data.

Many of the observations and comments from the review dealt with the general absence of secondary descriptive attributes associated with hydrologic observations within the ArcHydro Time Series Data Model. In order to address these issues and in an effort to meet the needs of the hydrologic community and the CUAHSI HIS for an adequate HOD structure, we have explored alternatives to the ArcHydro Time Series Data Model.

Design Considerations for a Hydrologic Observations Database

In developing a revised HOD structure, we began by extracting from the review comments the design considerations that were considered important by the reviewers. These considerations are:

1. The design should be generic and not rely on unique capabilities of proprietary software. It should be possible to implement the hydrologic observations database in a variety of relational database management systems, including Microsoft Access, Microsoft SQL Server, MySQL, Postgres, and others.
2. The hydrologic observations database should contain at a minimum the important information identified in the reviews of the ArcHydro Time Series Data Model (refer to the section above and to the review document, Tarboton, 2005)
3. The hydrologic observations database should be intuitive enough that users can understand how the data is stored and how to get data into and out of the database.
4. The hydrologic observations database should be capable of storing all information needed to populate a Time Series Object for interfacing with client software designed to view, manipulate, or analyze the data stored in the database.
5. Since the HOD will be the repository for hydrologic observations collected within the proposed hydrologic observatories, it is important that the database be capable of storing not only observations collected by researchers within the observatories, but in addition the HOD should be capable of storing data from the national databases and data collected by state and local agencies or other sources.

These considerations were used in the redesign of the HIS HOD structure.

Alternative Structures

In considering a revised database structure, we asked: **What are the basic attributes to be associated with each single observation and how can these best be organized?** The responses from the review of the originally proposed data model have provided a list of the important attributes to include in the database; however, fundamentally different database structures result from the choice of how much information to associate directly with each observation at the level of a single record, versus how much information is common to a set of observations and can be stored in a linked table. This consideration is important because the structure, number, and nesting of linked tables dictate the efficiency and ease of understanding and use of the data model.

In table 1 we list the attributes associated with each observation that were considered by the reviewers of the originally proposed data model to be necessary parts of the HOD structure. We have attempted to rank these attributes according to how closely they should be associated with the observation value itself, with the presumption that attributes closely associated with the observation value should be stored in the primary observations table while less closely associated information that is common over larger groups of observations should be stored in tables linked to the primary observations table.

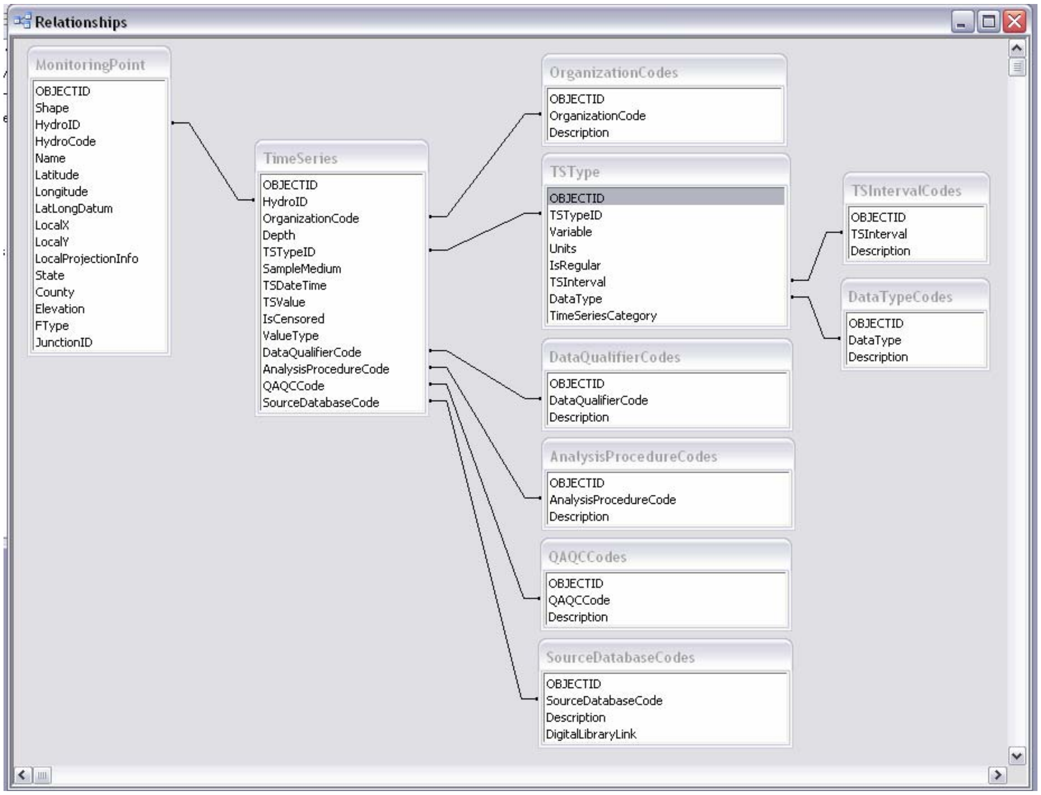
Table 1. Ranking of attributes associated with an observation

Attribute	Notes
Value	The observation itself
DateTime	The date and time of the observation (including time zone in which it occurred or offset relative to UTC)
Variable	The physical quantity that the value is measuring (e.g. streamflow, precipitation, water quality)
Location	The location of the observation (i.e., 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 the observations were 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)
OffsetType/ Reference Point	The reference point from which the offset to the measurement location was measured (i.e., 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 observations 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)
QA/QC	An indication of the quality of the data
Source Database	An indication of the original source of the observation (e.g., USGS NWIS, EPA STORET, local investigator, etc.)
Sample Medium	The medium in which the sample was collected (e.g., water, air, sediment, etc.)
Value Type	An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation

Two fundamentally different database structures were proposed by two different reviewers of the original data model. To evaluate the impact that these different designs have on the characteristics of the observations database, we populated the two different structures with a single dataset. For this comparison, the designs were modified from what the reviewers had suggested so that they both contained the same fields (data attributes), but differed in the way that the tables were organized.

The first structure is very similar to the ArcHydro Time Series Data Model, but it attempts to include much of the additional information requested by many of the reviewers. For the purposes of this example, we considered the first proposed structure to be inclusive of the ArcHydro Time Series Data Model. The second proposed structure is fundamentally different from the first proposed structure in that it stores much of the metadata associated with the observations in a linked table rather than in the same table as the observations themselves.

The following figures illustrate the main differences between the two structures proposed by the reviewers. Structure 1 (Figure 3) proposes direct inclusion of a larger amount of ancillary information as record level metadata in the time series table through identifiers that link in to adjoining tables. Structure 2 (Figure 4) proposes that all metadata information should be referenced through one TSType table linked to the main time series table, with other information linked to the TSType table. The remaining tables were identical in both databases. The first design is intended to facilitate querying directly based on a wide range of attributes at the cost of storing a number of metadata identifiers with each observation. The second design minimizes the number of metadata identifiers to be stored with each observation with the intent of reducing the size of primary time series table, but at the expense of a larger TSType table because there are more unique "type" combinations.



MonitoringPoint
- OBJECTID
- Shape
- HydroID
- HydroCode
- Name
- Latitude
- Longitude
- LatLongDatum
- LocalX
- LocalY
- LocalProjectionInfo
- State
- County
- Elevation
- FType
- JunctionID

TimeSeries
- OBJECTID
- HydroID
- OrganizationCode
- Depth
- TSTypeID
- SampleMedium
- TSDateTime
- TSValue
- IsCensored
- ValueType
- DataQualifierCode
- AnalysisProcedureCode
- QAQCCode
- SourceDatabaseCode

OrganizationCodes
- OBJECTID
- OrganizationCode
- Description

TSType
- OBJECTID
- TSTypeID
- Variable
- Units
- IsRegular
- TSInterval
- DataType
- TimeSeriesCategory

TSIntervalCodes
- OBJECTID
- TSInterval
- Description

DataTypeCodes
- OBJECTID
- DataType
- Description

DataQualifierCodes
- OBJECTID
- DataQualifierCode
- Description

AnalysisProcedureCodes
- OBJECTID
- AnalysisProcedureCode
- Description

QAQCCodes
- OBJECTID
- QAQCCode
- Description

SourceDatabaseCodes
- OBJECTID
- SourceDatabaseCode
- Description
- DigitalLibraryLink

Figure 3. Hydrologic Observations Database Alternative Structure 1.

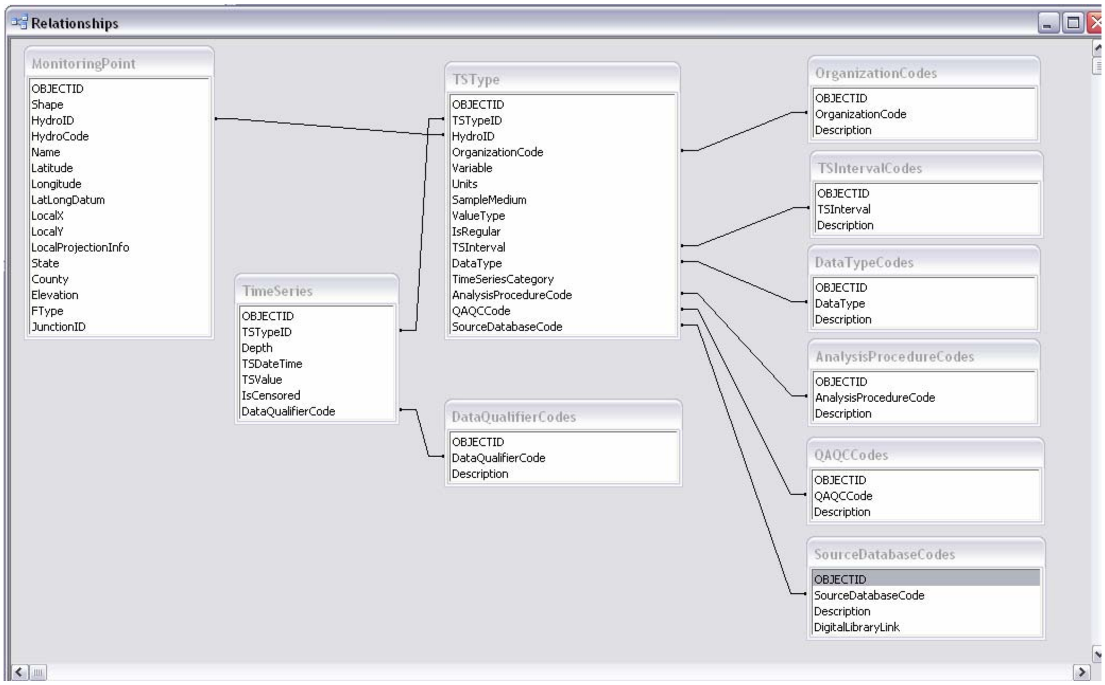


Figure 4. Hydrologic Observations Database Alternative Structure 2.

Both of these proposed structures were considered to be viable designs for the HIS HOD structure, and were, therefore, considered in the redesign of the HIS HOD database structure. It was anticipated, however, that each of these proposed structures would have implications and tradeoffs with regard to the design decisions listed above, and so a series of simple tests were performed to evaluate the two proposed structures. These tests are described in the following section.

Alternative Structure Tests

The two structures described in the previous section were evaluated through a series of simple tests that were designed to provide information about which of the structures was more appropriate to serve as the HIS HOD structure. Both database structures were implemented in Microsoft Access and were populated with USGS water quality data for a single 8-digit HUC (16010203 – Little Bear-Logan[3]). At the time it was downloaded, this dataset included 127 monitoring points, 369 different water quality variables, and 11,885 individual water quality observations.

All of the tables in the two databases are exactly the same, except for the TimeSeries and the TSType tables. In both databases, the TimeSeries table contains 11,885 records (one for each observation), but the TimeSeries table in proposed structure 1 contains metadata information that has been moved to the TSType table in proposed structure 2. The result of this fundamental difference is that the TSType table in proposed structure 1 contains 369 records (one for each unique variable), but the TSType table in proposed structure 2 contains 4359 records (one for each unique combination of location, organization, variable, units, sample medium, value type,

[3] http://nwis.waterdata.usgs.gov/ut/nwis/qwdata?huc_cd=16010203&format=rdb

etc.). In the context of the HOD database, it should be noted that the number of records in the TSTypes table of structure 2 could increase dramatically as more locations, organizations, variables, units, etc. are added to the database.

In terms of size on disk, proposed structure one is approximately 2.3 MB in size, and proposed structure 2 is approximately 6 MB in size. It is anticipated that structure 1 would be smaller than structure 2 as long as there is a relatively small number of observations (records in the TimeSeries table) and a relatively large number of variables (records in the TSType table). Conversely, it is anticipated that structure 1 would likely be larger than structure 2 if there were many observations (records in the TimeSeries table), but few variables (records in the TSType table). No tests were performed to confirm these observations.

Another simple test involved creating a simple query to retrieve data from the databases. This simple query test was not intended to demonstrate completely the differences in querying information out of the two databases. Rather, it is used here to demonstrate what is perhaps one of the most important differences between the two alternative structures. The following query was created so that it could be tested in both databases:

> *"Give me a list of the HydroID, HydroCode, and Name of all sampling locations at which water temperature data has been collected."*

Structure 1 allows the user to create a query to return the requested information by specifying criteria on the TSTypeID field *or* the Variable field to retrieve the requested information. The following are SQL statements used to return the requested information:

```
SELECT DISTINCT MonitoringPoint.HydroID, MonitoringPoint.HydroCode, MonitoringPoint.Name,
TimeSeries.TSTypeID
FROM MonitoringPoint INNER JOIN TimeSeries ON MonitoringPoint.HydroID = TimeSeries.HydroID
WHERE (((TimeSeries.TSTypeID)=10));
```

OR

```
SELECT DISTINCT MonitoringPoint.HydroID, MonitoringPoint.HydroCode, MonitoringPoint.Name,
TSType.Variable
FROM (MonitoringPoint INNER JOIN TimeSeries ON MonitoringPoint.HydroID = TimeSeries.HydroID) INNER
JOIN TSType ON TimeSeries.TSTypeID = TSType.TSTypeID
WHERE (((TSType.Variable) Like "Temperature, water*"));
```

Since there are many records in the TSType table of Structure 2 where the variable is "Temperature, water", this limits the ability to query in that we must specify criteria on the Variable field unless we know all of the TSTypeIDs where the variable is equal to "Temperature, water." The following is the query executed on structure 2 to return the requested information

```
SELECT DISTINCT MonitoringPoint.HydroID, MonitoringPoint.HydroCode, MonitoringPoint.Name,
TSType.Variable
FROM MonitoringPoint INNER JOIN TSType ON MonitoringPoint.HydroID = TSType.HydroID
WHERE (((TSType.Variable) Like "Temperature, water*"));
```

The queries to both database structures are nearly the same, but the criteria (**bold**) are different. In structure 1, we can use TSTypeID = 10 to return water temperature because 10 as the

TSTypeID for water temperature is unique. In Structure 1 we can also put criteria on the variable name because it is unique (i.e., we can do either to return the same information). In structure 2, there are many TSTypeIDs that represent water temperature, so we can only put criteria on the variable name unless we know all of the integer TSTypeIDs that correspond to water temperature (there are 112 of them). It is important to consider that to put criteria on the variable name we must deal with the vocabulary of the Variable issue (i.e., is it "Temperature, water, degrees Celsius" or "Water Temperature, degrees Celsius" or "Water Temperature, deg. C," etc. This can be controlled to some degree through the use of a controlled vocabulary in the variable field.

Revised Hydrologic Observations Database Structure Design

After evaluating the two proposed database structures, we have settled on a structure that falls somewhere in between the two. In general, we preferred structure 1 because it was easier to populate and more intuitive to query. However some changes have been made to proposed structure 1 to meet the needs of the CUAHSI HIS and to address the comments from the reviewers. For starters, some of the metadata will be maintained at the record level in the Observations table (formerly the TimeSeries table), and, where appropriate, some has been moved to the ObservationTypes table (formerly the TSType table). This will avoid what we perceive to be unnecessary duplication in the ObservationTypes table, and it will make it easier to retrieve data from the database based on a variable type. In addition, we have changed the names of some of the tables and fields to reflect that the database is storing hydrologic observations. Figure 5 shows the table schema for the revised HIS hydrologic observations database. Appendix A provides a data dictionary that lists the tables in the database, the names and data types of each of the fields in the tables, and provides a description of the information contained in each of the fields.

In addition to the changes listed in the preceding paragraph, we have made several other modifications to the database structure so that it differs from those that were tested. They are as follows:

1. We have added an ObservationsCatalog table to the database. Although not required to maintain the integrity of the data, this table provides a listing of all of the monitoring point and observation type combinations in the database. This provides a means by which a user can get simple descriptive information about the variables observed at a location, the most common anticipated query, without the overhead of querying the entire time series table, which can become quite large.
2. We have added a UTCOffset field to the Observations table to ensure that local times recorded in the database can be referenced to standard time and to enable comparison of results across databases that may store observations collected in different time zones (i.e., compare observations from one hydrologic observatory to those collected at another hydrologic observatory located across the country). A design choice here was to have UTCOffset as a record level qualifier because even though the time zone and hence offset is likely the same for all measurements at a monitoring point, the offset changes due to daylight savings. Some investigators may run data loggers on standard time, while others may adjust for daylight saving or use universal time. To avoid the necessity to keep track

of the system used, or impose a system that might be cumbersome and lead to errors we decided that if the offset was always recorded the precise time would be unambiguous and would reduce the chance for interpretation errors.

3. We have added an ObservationID to the Observations table to uniquely identify each individual observation and serve as an identifier for use in the definition of logical groupings of observations and sets of observations used to derive other observations.
4. We have added two tables, ObservationGroups and GroupDescriptions, which enable the logical grouping of observations (i.e., assigning all observations from a single reservoir profile to one group). These tables provide a means of grouping together observations that are logically related.
5. We have added a DerivedFromID to the Observations table and a DerivedFrom table to the database. The DerivedFromID points to the DerivedFrom table where the observations from which a quantity was derived are listed (e.g. a daily average discharge value could be linked to the 96 15 minute unit values from which it was derived, or a snow water equivalent value could be linked to the depth and density values from which it was derived).
6. We have combined the AnalysisProcedureCodes and QAQCCodes tables into a single table that indicates the method used to collect the observation and the QAQC associated with that method. The description field in this table would describe both the analysis procedure and the QAQC level.
7. We have converted the DataType field to a text field with a controlled vocabulary (rather than a coded value domain) eliminating the need for a value coding table. We have also added some additional categories to the DataTypes.
8. We have renamed the TSInterval field as ObsTimeSupport to use this field to specifically quantify the time support scale of the measurements. The time units of the observation support are to be listed in a new field called TimeUnit. In addition, we have added a field to the ObservationTypes table called UnitType, which defines the dimensions of the units. The definitions of DataTypes and support scale are given below.
9. We have added a CategoryDefinitions table that stores the categories associated with categorical observations. These observations are encoded as double values in the Observations table.

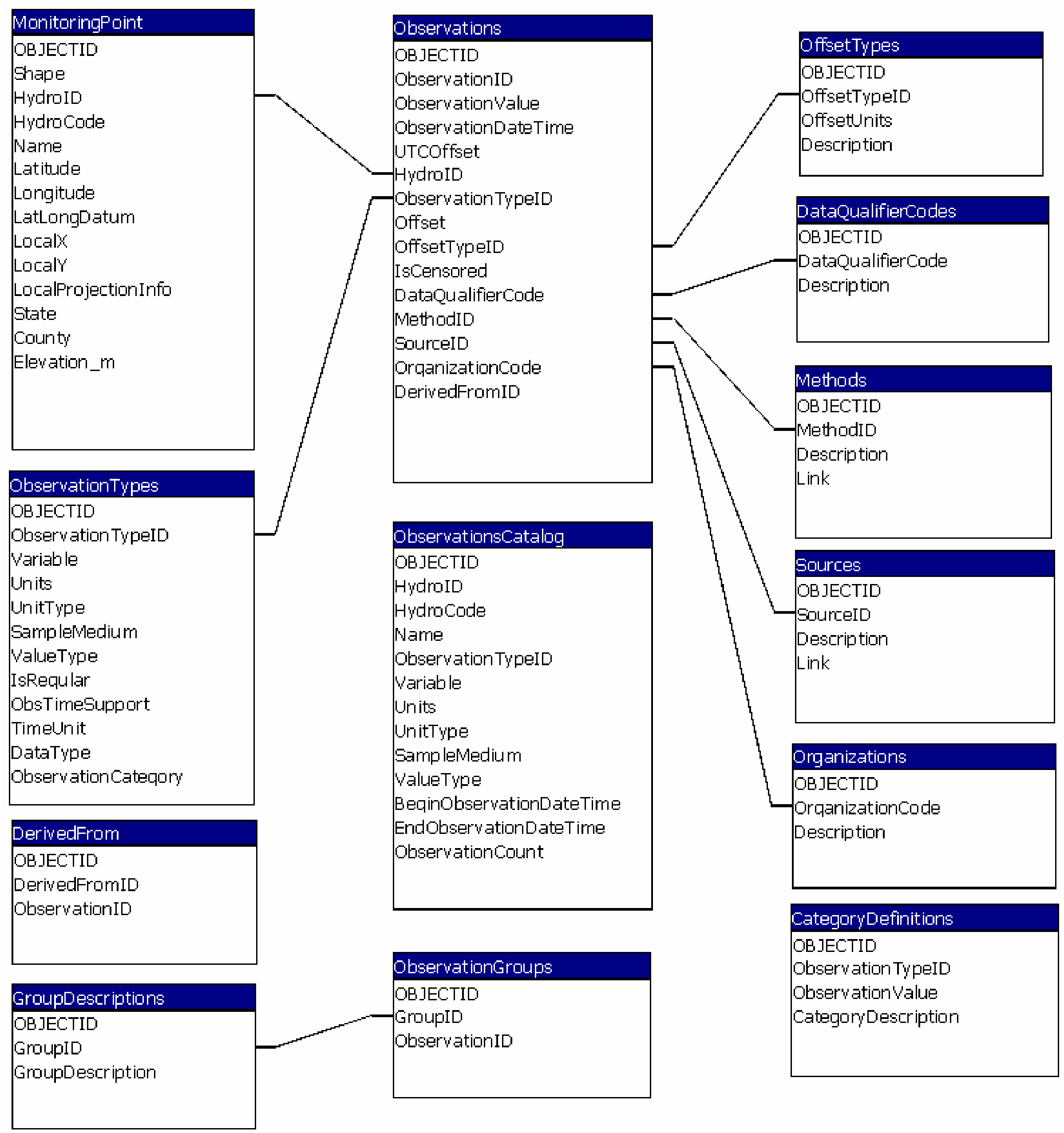


Figure 5. Proposed Hydrologic Observations Database Structure.

DataType and Support Scale

In interpreting observations that comprise a time series it is important to know the scale information associated with the observations. Blöschl and Sivapalan (1995) review the important issues. Any set of observations is quantified by a scale triplet comprising support, spacing and extent, illustrated in Figure 6.

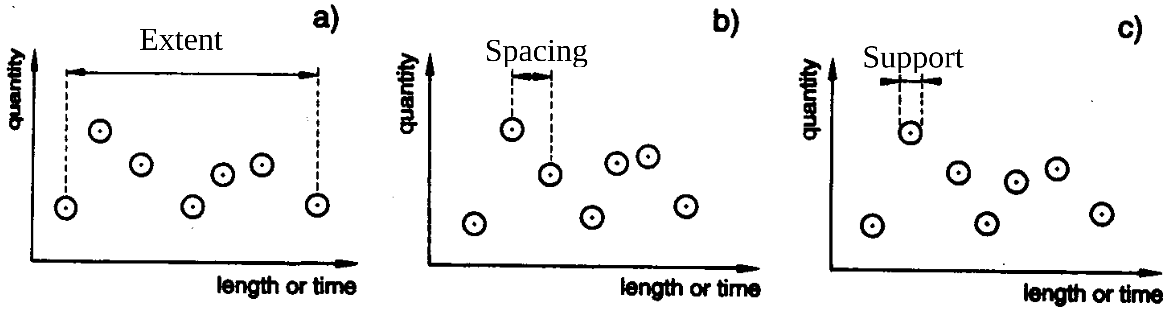


Figure 6. The Scale Triplet of Measurements (a) Extent, (b) Spacing, (c) Support. (from Blöschl, 1996)

Extent is the full range over which the measurements occur, spacing is the spacing between measurements and support is the averaging interval or footprint implicit in any measurement. In the proposed Hydrologic Observations Data model extent and spacing are properties of multiple measurements and are defined by the DateTime associated with observations. Instead of a variable TSinterval that was in the preliminary data model we have included a field called ObservationSupport in the time series table to explicitly quantify support. Figure 7 shows some of the implications associated with support, spacing and extent in the interpretation of time series observations.

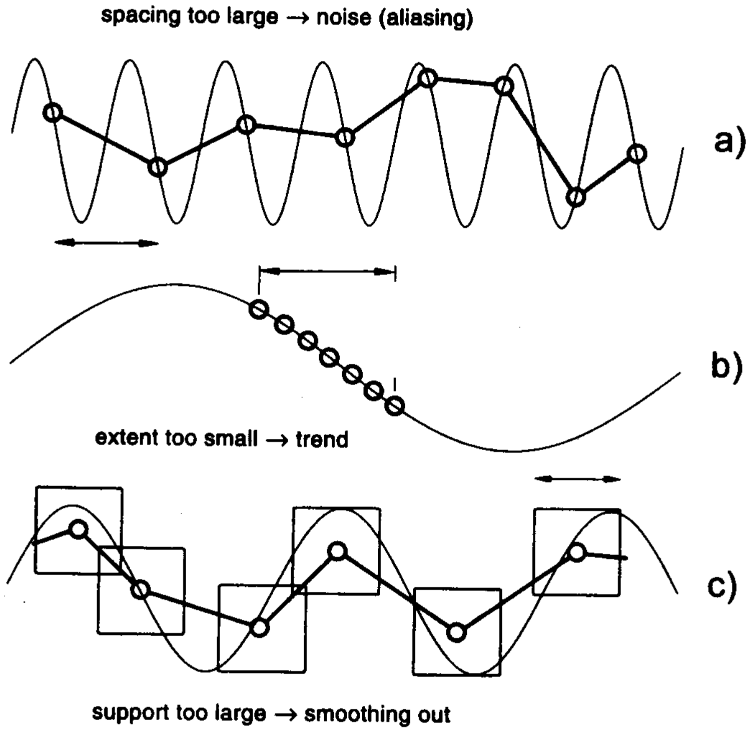

spacing too large → noise (aliasing)
a)
b)
extent too small → trend
c)
support too large → smoothing out

Figure 7. The effect of sampling for measurement scales not commensurate with the process scale. (a) Spacings larger than the process scale cause aliasing in the data; (b) Extents smaller than the process scale cause a trend in the data; (c) Supports larger than the process scale cause excessive smoothing in the data. (from Blöschl, 1996)

In the proposed Hydrologic Observations Data model the following data types are suggested. These are extensions from the initial ArcHydro time series data model.

1. *Continuous* data – 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.
2. *Instantaneous* data – 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, such as for example infrequent water quality samples.
3. *Cumulative* data – the data represents the cumulative value of a variable measured or calculated up to a given instant of time, such as cumulative volume of flow or cumulative precipitation: $\mathrm{V(t)} = \int_0^t \mathrm{Q}(\tau)\mathrm{d}\tau$, where τ represents time in the integration over the interval [0,t]. To unambiguously interpret cumulative data one needs to know the time origin. We suggest the convention of using a cumulative record with an ObservationValue of zero to initialize or reset cumulative data. With this convention cumulative data should be interpreted as the accumulation over the time interval between the DateTime of the zero record and the current record at the same observation position. Observation position is defined by a unique combination of HydroID, ObservationType, Offset and OffsetType. All four of these quantities comprise the unambiguous description of the position of an observation and there may be multiple time series associated with multiple observation positions (e.g. redundant rain gauges with different offsets) at a location.
4. *Incremental* data – the value represents the incremental value of a variable over a time interval Δt such as the incremental volume of flow, or incremental precipitation: $\Delta \mathrm{V(t)} = \int_{t-\Delta t}^{t} \mathrm{Q}(\tau)\mathrm{d}\tau$. As for cumulative data, unambiguous interpretation requires knowledge of the time increment. Here we suggest the convention of using ObservationSupport if this is given, or the time interval from the previous observation at the same position if ObservationSupport is not given or is 0. This accommodates incremental type precipitation data that is only reported when the value is non-zero, such as NCDC data.
5. *Average* data – the 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 ObservationSupport in the case of regular data (as quantified by the IsRegular field) and by the time interval from the previous observation at the same position for irregular data.
6. *Maximum* data – the value is the maximum value occurring at some time during a time interval, such as annual maximum discharge or a daily maximum air temperature. Again unambiguous interpretation requires knowledge of the time interval. We suggest the convention that the time interval is the ObservationSupport for regular data and the time interval from the previous observation at the same position for irregular data.

7. *Minimum* data – the value is the minimum value occurring at some time during a time interval, such as 7-day low flow for a year, or the daily minimum temperature. The time interval is defined similarly to Maximum data.
8. *Constant over interval* data – the value is a quantity that can be interpreted as constant over the time interval from the previous measurement.
9. *Categorical* data – the value is a categorical rather than continuous valued quantity. Mapping from ObservationValue values to categories is through the CategoryDefinitions table.

Examples

To demonstrate the capability of this design to store a diverse set of hydrologic observations Appendix B gives examples of how an illustrative set of observations would be represented in this database design.

Discussion

This data model design was conceived with a number of considerations in mind, some of which came from the review of the initial data model and others of which emerged during discussion of this design. These are reviewed here to give a sense of some of the capabilities envisaged for the data model.

The DerivedFrom and ObservationGroups table fulfill the function of grouping observations for different purposes. These are tables where the same identifier (DerivedFromID or GroupID) can appear multiple times in the table associated with different ObservationIDs thereby defining the associated group of records. In the DerivedFrom table this is the sole purpose of the table and each group so defined is associated with a record in the Observations table (through the DerivedFromID field in that table). This record would have been derived from the observations identified by the group. The method of derivation would be given through the methods table associated with the observation. This construct is useful for example to identify the 96 15 min unit streamflow values that go into the estimate of the mean daily streamflow. Note that there is no limit as to how many groups an observation may be associated with, and observations that are derived from other observations may themselves belong to groups used to derive other observations (e.g. the daily minimum flow over a month derived from daily observations derived from 15 min unit values). Note also that a derived from group may have as few as one observation for the case where an observation is derived from a single more primitive observation (e.g. Discharge from Stage). Through this construct the data model has the capability to store raw observations and simple derivatives preserving the connection of each observation to its more primitive raw measurement.

In the design presented we have represented categorical or ordinal variables in the same table as continuous valued 'double' variables through a numerical encoding of the categorical observation value as a ‘double’ value. The CategoryDefinitions table then associates, for each observation type an observation value with an associated category definition. This is a somewhat cumbersome construct because real valued 'double' quantities are being used as database keys. We do not see this as a significant shortcoming though because typically, in our judgment, only a small fraction of hydrologic observations will be categorical. An alternative approach could have been to have a separate Observations table for categorical observations.

The Methods and Sources tables both contain links that we have indicated as either a URL or reference to a file in a digital library. It will be important as the database grows and is used over time to ensure that links or URL's included are stable. An alternative approach to external links is to exploit the capability of modern databases to store as fields within a record entire digital documents, such as an html or xml page, PDF document or raw data file. The capability therefore exists to instead have these links refer to a Documents table that would actually contain this metadata information, instead of housing it in digital library. There is some merit in this because then any data exported in Hydrologic Observations Data model format could take with it the associated metadata required to completely define it as well as the raw data upon which it is derived. This however has the disadvantage of increasing (perhaps substantially) the size of database file containing the data and being distributed to users. The implications of this idea have not been fully explored. It is mentioned here as a possibility worthy of further consideration.

A considerable portion of hydrologic observations data is in the form of time series. This was why the initial model was based on the ArcHydro Time Series Data Model. The proposed design has not specifically highlighted time series capabilities, nevertheless the data model has inherited the key components from the ArcHydro Time Series Data Model to give it time series capability. In particular one observation DataType is "Continuous," designed to indicate that the observations are collected with sufficient frequency as to be interpreted as a smooth time series. The IsRegular field also facilitates time series analysis because certain time series operations (e.g. Fourier Analysis) are predisposed to regularly sampled data. At first glance it may appear that there is redundancy between the Isregular field and the DataType "Continuous" but we chose to keep these separate because there are regularly sampled quantities for which it is not reasonable to interpret the values as "Continuous". For example monthly grab samples of water quality are not continuous, but are better categorized as having DataType, "Instantaneous". Note that the data model does not explicitly store the time interval between measurements, nor does it indicate where a continuous series has data gaps. Both these are required for time series analysis, but are inherently not properties of single measurements. The time interval is the time difference between sequential regular measurements, something that could be easily computed from DateTime values by analysis tools. The inference of measurement gaps (and what to do about them) from DateTime values we also regard as analysis functionality left for the Hydrologic Analysis System to handle.

Conclusions

This paper has presented the design for a community hydrologic observations database structure that is designed to store hydrologic observations in a flexible, relational database system to facilitate data retrieval for integrated analysis of information collected by multiple investigators. The design represents an evolution of the initial ArcHydro time series database design, to address the specific needs of the CUAHSI community identified by reviewers of the initial design. The data model is focused on storing the original observations, simple derived quantities, and ancillary information (metadata) sufficient to allow unambiguous interpretation of data, while at the same time providing traceable heritage from raw measurements to usable information. It is recommended that this data model be implemented and tested in a number of database systems to fully evaluate its suitability for adoption as a CUAHSI hydrologic observations data model standard.

Acknowledgements

This work is supported by the National Science Foundation under grant #EAR-0413265 to CUAHSI and the University of Texas at Austin. The views and conclusions expressed are those of the authors and do not necessarily reflect the views of the U.S. Government.

References

Blöschl, G. and M. Sivapalan, (1995), "Scale Issues in Hydrological Modelling: A Review," Hydrological Processes, 9(1995): 251-290.

Blöschl, G., (1996), Scale and Scaling in Hydrology, Habilitationsschrift, Weiner Mitteilungen Wasser Abwasser Gewasser, Wien, 346 p.

Maidment, D. R., ed. (2002), Arc Hydro Gis for Water Resources, ESRI Press, Redlands, CA, 203 p.

Maidment, D.R. 2005. A Data Model for Hydrologic Observations. Paper prepared for presentation at the CUAHSI Hydrologic Information Systems Symposium. University of Texas at Austin. March 7, 2005.

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

Appendix A
Table and Field Structure for the Proposed HIS Hydrologic Observations Database

The following is a description of the tables in the proposed hydrologic observations database schema, a listing of the fields contained in each table, a description of the data contained in each field and its type, examples of the information to be stored in each field where appropriate, and any additional information about each field. Values in the example column should not be considered to be inclusive of all potential values, especially in the case of fields that will require a controlled vocabulary. We have developed some suggestions for the controlled vocabulary for some fields, but anticipate that these will need to be extended and adjusted.

Table: CategoryDefinitions

Associates observation value with the definition of a category for categorical variables

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
ObservationTypeID	Integer	Integer identifier that references the observation type record of a categorical variable		This identifies the specific type of observations for which a value to category mapping applies and avoids conflicts where the same numerical value may map into different categories for different observation types.
ObservationValue	Double	Numeric value of Observation	1.0	Although a real number represented as a double these are associated with categories defined in the CategoryDescription field
CategoryDescription	Text	Definition of categorical variable value	"Cloudy"	

Table: DataQualifierCodes

Lists the full descriptions of the data qualifying comments that accompany the data. This table serves to define the controlled vocabulary of text codes stored in the observations table.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
DataQualifierCode	Text	Unique code identifying the data qualifying comment	“H”	The following initial controlled vocabulary is suggested: E – Estimated P – Provisional D – Derived H – Holding time for sample analysis exceeded
Description	Text	Full description or text of the data qualifying comment	“Holding time for sample analysis exceeded”	

Table: DerivedFrom

Table that contains the linkage between derived quantities and the observations that they were derived from.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
DerivedFromID	Integer	Unique integer identifying the group of observations from which a quantity is derived		
ObservationID	Integer	Integer identifier referencing observations that comprise a group of observations from which a quantity is derived		This corresponds to ObservationID in the Observations table

Table: GroupDescriptions

Lists the descriptions for each of the observation groups that have been formed.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
GroupID	Integer	Unique integer identifier for each group of observations that has been formed		This also references to GroupID in the ObservationGroups table
GroupDescription	Text	Text description of the group	"Echo Reservoir Profile 7/7/2005"	

Table: Methods

Lists the methods used to collect the data and provides an indication of the Quality Assurance and Quality Control procedures associated with each method.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
MethodID	Integer	Unique integer ID for each measurement/QAQC method.		
Description	Text	Text description of each measurement/QAQC method.	"Total phosphorus measured using EPA procedure XXX with published QAQC plan"	
Link	Hyperlink	Link to a file in digital library or URL that provides a description of the method		

Table: Monitoring Point

Provides information giving the spatial location at which observations have been collected.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
Shape	Binary Object	ESRI geodatabase shape information		
HydroID	Integer	Unique integer ID for each sampling location		Easier to index and query than the HydroCode, which is text
HydroCode	Text	Unique text identifier for each sampling location	“10109000”	This is redundant with HydroID but is retained to provide a recognizable identifier associated with each location useful for error checking
Name	Text	Full name of sampling location	“LOGAN RIVER ABOVE STATE DAM, NEAR LOGAN,UT”	
Latitude	Double	Latitude in decimal degrees		
Longitude	Double	Longitude in decimal degrees		
LatLongDatum	Text	Datum of latitude and longitude	“NAD 83” “NAD 27”	Controlled Vocabulary
LocalX	Double	Local Projection X coordinate		

Field Name	Data Type	Description	Example	Notes
LocalY	Double	Local Projection Y Coordinate		
LocalProjectionInfo	Text	Information describing local projection	“UTMZone12NAD83”	Controlled Vocabulary
State	Text	Name of state in which the sampling station is located	“Utah”	
County	Text	Name of County in which the sampling station is located	“Cache”	
Elevation_m	Double	Elevation of sampling location (in m)		Meters above sea level

Table: ObservationGroups

Lists the groups of observations that have been created and the observations that are within each observation group.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
GroupID	Integer	Unique integer ID for each group of observations that has been formed		
ObservationID	Integer	Integer identifier for each observation that belongs to a group		This corresponds to ObservationID in the Observations table

Table: Observations

Stores the actual hydrologic observations.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
ObservationID	Integer	Unique integer identifier for each observation		
ObservationValue	Double	Numeric value of observation		Categorical information is stored as a number with the categories defined in observation type table?

ObservationDateTime	Date/Time	Local date and time at which the observation was made		Represented as: MM/DD/YYYY hh:mm:ss.sss
				Where MM=Month
				DD = Day
				YYYY=Year
				hh = Hour
				mm = minutes
				ss.sss = seconds with milliseconds

UTCOffset	Integer	Offset from UTC time at the sampling location		Number of hours
HydroID	Integer	Integer identifier of the sampling location at which the observation was made		This links observations to their locations in the MonitoringPoint table
ObservationTypeID	Integer	Integer identifier that references the variable that was measured		This links observations to their type in the ObservationTypes table

Field Name	Data Type	Description	Example	Notes
Offset	Double	Distance from a datum or control point at which an observation was made		
OffsetTypeID	Integer	Unique integer identifier that references the type of measurement offset		This links observation offsets to their type in the OffsetTypes table
IsCensored	Text	Text indication of whether the data value is censored		Controlled Vocabulary “gt”=greater than “lt”=less than “nc” or blank=not censored
DataQualifierCode	Text	Text code that indicates a data qualifying comment		These codes are defined in the DataQualifierCodes table
MethodID	Integer	Integer identifier that references the measurement method/QAQC combination associated with the observation		This links observations to their method description in the Methods table
SourceID	Integer	Integer identifier that references the record in the Sources table giving the source of the observation		
OrganizationCode	Text	Unique text code that identifies the organization that colledted the data		The organization table associates the 'short' organization code with a complete organization description

Field Name	Data Type	Description	Example	Notes
DerivedFromID	Integer	Integer identifier for the group of observations that the current observation is derived from		This refers to a group of derived from records in the DerivedFrom table.

Table: ObservationsCatalog

Lists each of the MonitoringPoint/ObservationType combinations in the database as an index to speed some simple queries. This table contains the necessary fields to uniquely identify each sampling location and each measured quantity at that location for the purposes of identifying or displaying what data are available at each location without querying the main Observations table.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
HydroID	Integer	Unique integer monitoring point or sampling location identifier		
HydroCode	Text	Unique text identifier for each sampling location		
Name	Text	Full text name of sampling location		
ObservationTypeID	Integer	Integer identifier for each ObservationType		
Variable	Text	Name of the variable corresponding to observation type	“Water Temperature”	Controlled Vocabulary
Units	Text	Units of the variable corresponding to observation type	“Degrees Celsius”	Controlled Vocabulary
UnitType	Text	Text value that specifies the dimensions of the units	“Length” “Time” “Mass”	Controlled Vocabulary

Field Name	Data Type	Description	Example	Notes
SampleMedium	Text	The medium of the sample	“Surface Water” “Sediment” “Fish Tissue”	Controlled Vocabulary
ValueType	Text	Text value indicating what type of observation is being recorded	“Field Observation” “Laboratory Observation” “Model Simulation Results”	Controlled Vocabulary
BeginObservationDateTime	Date/Time	Date of the first observation in the series identified by the combination of the HydroID and ObservationTypeID		
EndObservationDateTime	Date/Time	Date of the last observation in the series identified by the combination of the HydroID and ObservationTypeID		
ObservationCount	Integer	The number of observations in the series identified by the combination of the HydroID and the ObservationTypeID		

Table: ObservationTypes

Lists the full descriptive information about what variables have been measured.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
ObservationTypeID	Integer	Unique integer identifier for each ObservationType		

Field Name	**Data Type**	**Description**	**Example**	**Notes**
Variable	Text	Name of the variable that was measured, observed, modeled, etc.	“Water Temperature”	Controlled Vocabulary
Units	Text	Text units of the observation	“Degrees Celsius”	Controlled Vocabulary
UnitType	Text	Text value that specifies the dimensions of the units	“Length” “Time” “Mass”	Controlled Vocabulary
SampleMedium	Text	The medium of the sample	“Surface Water” “Sediment” “Fish Tissue”	Controlled Vocabulary
ValueType	Text	Text value indicating what type of observation is being recorded	“Field Observation” “Laboratory Observation” “Model Simulation Results”	Controlled Vocabulary
IsRegular	Boolean	Value that indicates whether the values are from a regularly sampled time series	“True” “False”	Controlled Vocabulary
ObsTimeSupport	Double	Numerical value that indicates the support (or temporal footprint) for these observations	0, 24	0 is used to indicate a value that is instantaneous. Other values indicate the time over which the observations are implicitly or explicitly averaged

Field Name	**Data Type**	**Description**	**Example**	**Notes**
TimeUnit	Text	Text value that specifies the basic units of the observation support	“Second” “Minute” “Hour” “Day” “Month” “Year”	Controlled Vocabulary
DataType	Text	Text value that identifies the data as one of several types	“Continuous” “Instantaneous” “Cumulative” “Incremental” “Average” “Minimum: “Maximum” “Constant Over Interval” “Categorical”	Controlled Vocabulary
ObservationCategory	Text	General category of the observations	“Climate” “Water Quality” “Groundwater Quality”	Controlled Vocabulary

Table: OffsetTypes

Lists the full descriptive information for each of the measurement offsets.

Field Name	**Data Type**	**Description**	**Example**	**Notes**
OBJECTID	Integer (Autonumber)			
OffsetTypeID	Integer	Unique integer identifier that identifies the type of measurement offset		
OffsetUnits	Text	Units of the offset distance	“m” for meters	Controlled Vocabulary
Description	Text	Full text description of the offset type	“Below water surface” “Above Ground Level”	Controlled Vocabulary

Table: Organizations

Lists the full descriptive information for each data collection organization.

Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
OrganizationCode	Text	Unique text code that identifies the data collection organization		
Description	Text	Full text description of data collection organizations	"United States Geological Survey"	

Table: Sources

Lists the original sources of the data, including a link to the original data files and metadata that should be contained in the digital library.

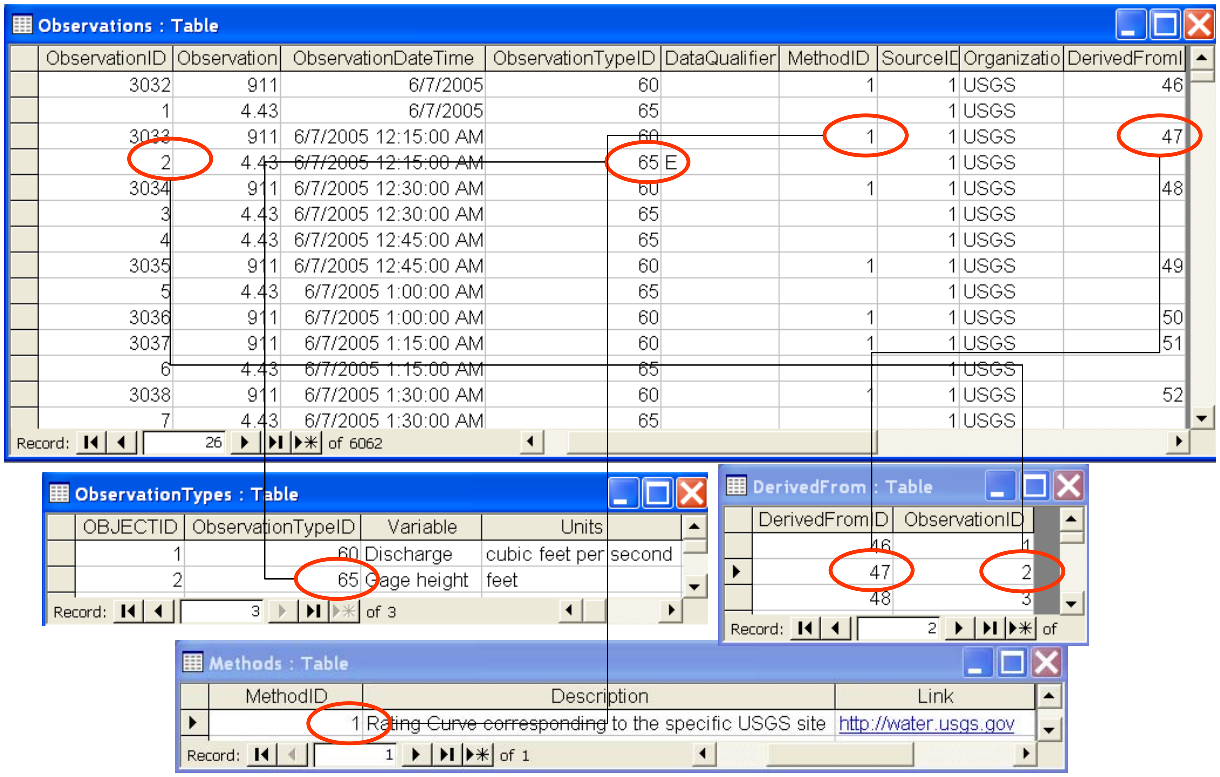
Field Name	Data Type	Description	Example	Notes
OBJECTID	Integer (Autonumber)			
SourceID	Integer	Unique integer identifier that identifies each data source		
Description	Text	Full text description of the source database	"Text file retrieved from the United States Geological Survey National Water Information System"	
Link	Hyperlink	Link to original data file and associated metadata stored in the digital library or URL of data source		

Appendix B
Examples

The following examples show the capability of the proposed data structure to store different types of hydrologic observations.

Streamflow Stage and Discharge

Both stage measurements and the associated discharge estimates derived from the stage measurements can be stored in the proposed observations database (Figure A.1).



Observations : Table

ObservationID	Observation	ObservationDateTime	ObservationTypeID	DataQualifier	MethodID	SourceID	Organizatio	DerivedFromI
3032	911	6/7/2005	60		1	1	USGS	46
1	4.43	6/7/2005	65			1	USGS	
3033	911	6/7/2005 12:15:00 AM	60		1	1	USGS	47
2	4.43	6/7/2005 12:15:00 AM	65	E		1	USGS	
3034	911	6/7/2005 12:30:00 AM	60		1	1	USGS	48
3	4.43	6/7/2005 12:30:00 AM	65			1	USGS	
4	4.43	6/7/2005 12:45:00 AM	65			1	USGS	
3035	911	6/7/2005 12:45:00 AM	60		1	1	USGS	49
5	4.43	6/7/2005 1:00:00 AM	65			1	USGS	
3036	911	6/7/2005 1:00:00 AM	60		1	1	USGS	50
3037	911	6/7/2005 1:15:00 AM	60		1	1	USGS	51
6	4.43	6/7/2005 1:15:00 AM	65			1	USGS	
3038	911	6/7/2005 1:30:00 AM	60		1	1	USGS	52
7	4.43	6/7/2005 1:30:00 AM	65			1	USGS	

Record: 26 of 6062

ObservationTypes : Table

OBJECTID	ObservationTypeID	Variable	Units
1	60	Discharge	cubic feet per second
2	65	Gage height	feet

Record: 3 of 3

DerivedFrom : Table

DerivedFromID	ObservationID
46	1
47	2
48	3

Record: 2 of

Methods : Table

MethodID	Description	Link
1	Rating Curve corresponding to the specific USGS site	http://water.usgs.gov

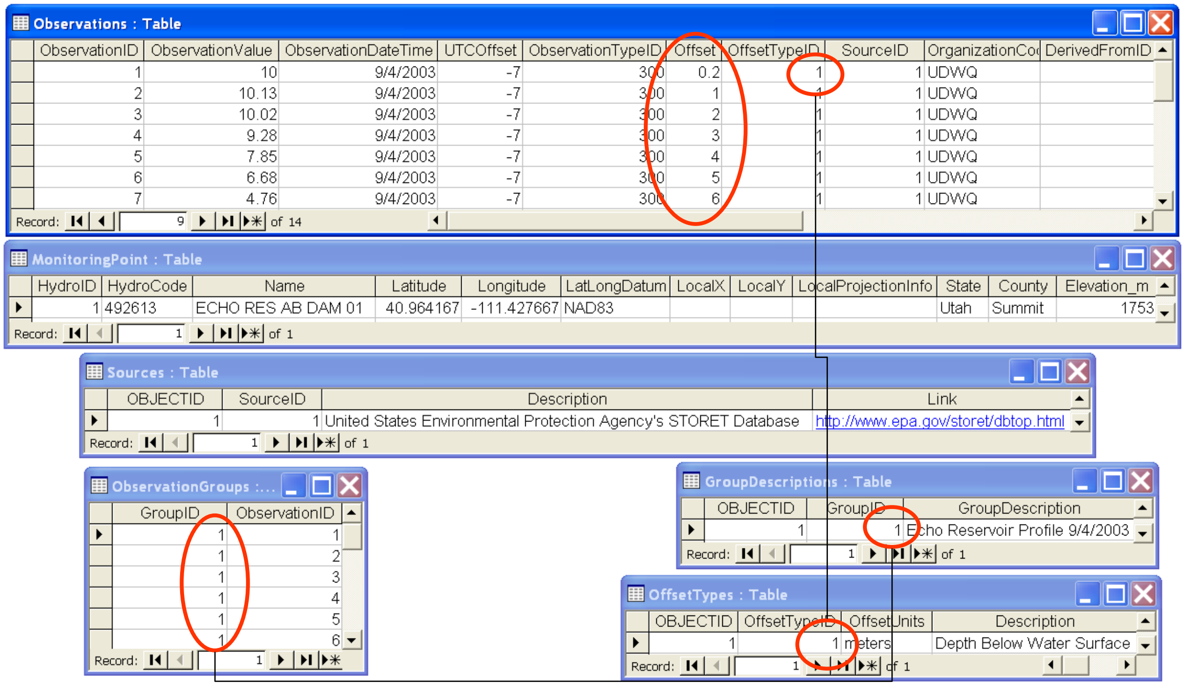
Record: 1 of 1

Figure A.1. Excerpts from tables illustrating the population of the data model with streamflow stage and discharge data.

Note that stage in feet and discharge in cubic feet per second are both in the same data table but with different observation types that reference the variable, units and other quantities associated with these observations. The link between ObservationTypeID in the Observations table and ObservationTypes table is shown. In this example, discharge measurements are presumed to be derived from stage measurements through a rating curve. The MethodID associated with each discharge record references into a method table that describes this and provides a URL that should contain metadata details for this method. The DerivedFromID in the Observations table references into the DerivedFrom table that references back to the corresponding stage in the observations table from which the discharge was derived.

Water Chemistry from a Profile in a Lake

Reservoir profile measurements provide an example of observations that should logically be grouped and observations that have an offset in relationship to the location of the sampling station. 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). The following shows an example of how these data would be stored in the proposed database structure.



Observations : Table

ObservationID	ObservationValue	ObservationDateTime	UTCOffset	ObservationTypeID	Offset	OffsetTypeID	SourceID	OrganizationCo	DerivedFromID
1	10	9/4/2003	-7	300	0.2	1	1	UDWQ	
2	10.13	9/4/2003	-7	300	1	1	1	UDWQ	
3	10.02	9/4/2003	-7	300	2	1	1	UDWQ	
4	9.28	9/4/2003	-7	300	3	1	1	UDWQ	
5	7.85	9/4/2003	-7	300	4	1	1	UDWQ	
6	6.68	9/4/2003	-7	300	5	1	1	UDWQ	
7	4.76	9/4/2003	-7	300	6	1	1	UDWQ	

Record: 9 of 14

MonitoringPoint : Table

HydroID	HydroCode	Name	Latitude	Longitude	LatLongDatum	LocalX	LocalY	LocalProjectionInfo	State	County	Elevation_m
1	492613	ECHO RES AB DAM 01	40.964167	-111.427667	NAD83				Utah	Summit	1753

Record: 1 of 1

Sources : Table

OBJECTID	SourceID	Description	Link
1	1	United States Environmental Protection Agency's STORET Database	http://www.epa.gov/storet/dbtop.html

Record: 1 of 1

ObservationGroups :...

GroupID	ObservationID
1	1
1	2
1	3
1	4
1	5
1	6

Record: 1

GroupDescriptions : Table

OBJECTID	GroupID	GroupDescription
1	1	Echo Reservoir Profile 9/4/2003

Record: 1 of 1

OffsetTypes : Table

OBJECTID	OffsetTypeID	OffsetUnits	Description
1	1	meters	Depth Below Water Surface

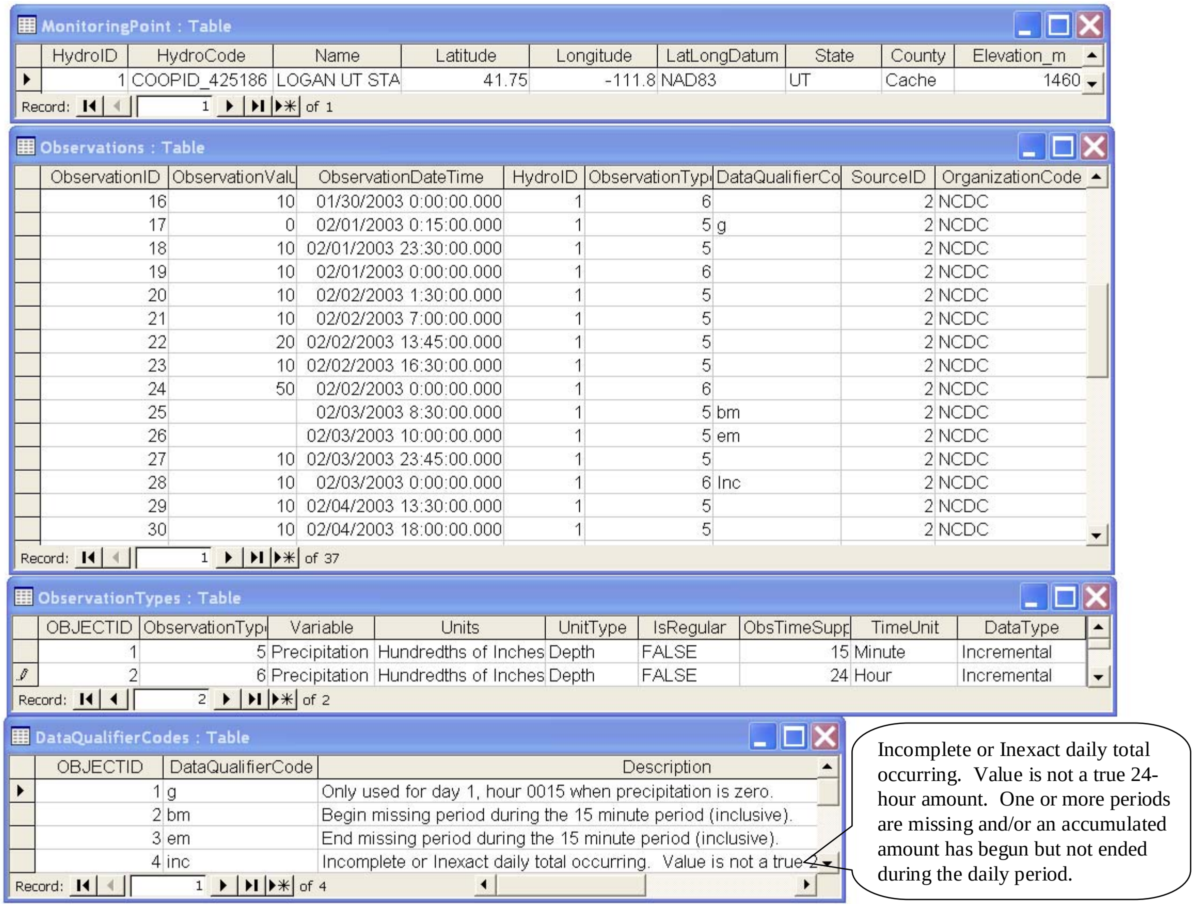
Record: 1 of 1

Figure A.2. Excerpts from tables illustrating the population of the data model with Water Chemistry data.

This example illustrates the use of the OffsetTypes table and Offset attribute to quantify the depth associated with each measurement. This example also illustrates the use of the ObservationGroups table and GroupDescriptions table to group logically related measurements. The MonitoringPoint table includes HydroID and shape information (not shown) that locates each observation geographically within a GIS, but also includes Latitude and Longitude and LocalX and LocalY coordinates to provide location information independent of the GIS system. The Sources table indicates the source of this data from the EPA STORET database with URL given.

NCDC Precipitation Data

Figure A.3 illustrates the representation of NCDC 15 min precipitation data by the Data Model. The data files include 15 min observations as well as daily totals. Separate observation types are used for the 15 min or daily total values. This data is reported at irregular intervals and only for time periods for which precipitation is non zero. This is accommodated by setting the IsRegular attribute associated with the observation type to False and specifying the ObsTimeSupport value as 15 or 24 and the TimeUnit as "Minute" or "Hour". The DataType of 'incremental' is used to indicate that these are incremental values defined over the ObsTimeSupport interval. Data qualifier codes indicate periods where the data is missing. This is necessary because of the convention that zero precipitation periods are not reported. A data qualifier code is also used to flag days where the precipitation total is incomplete due to the record being missing during part of the day.



MonitoringPoint : Table

HydroID	HydroCode	Name	Latitude	Longitude	LatLongDatum	State	County	Elevation_m
1	COOPID_425186	LOGAN UT STA	41.75	-111.8	NAD83	UT	Cache	1460

Record: 1 of 1

Observations : Table

ObservationID	ObservationValu	ObservationDateTime	HydroID	ObservationTyp	DataQualifierCo	SourceID	OrganizationCode
16	10	01/30/2003 0:00:00.000	1	6		2	NCDC
17	0	02/01/2003 0:15:00.000	1	5	g	2	NCDC
18	10	02/01/2003 23:30:00.000	1	5		2	NCDC
19	10	02/01/2003 0:00:00.000	1	6		2	NCDC
20	10	02/02/2003 1:30:00.000	1	5		2	NCDC
21	10	02/02/2003 7:00:00.000	1	5		2	NCDC
22	20	02/02/2003 13:45:00.000	1	5		2	NCDC
23	10	02/02/2003 16:30:00.000	1	5		2	NCDC
24	50	02/02/2003 0:00:00.000	1	6		2	NCDC
25		02/03/2003 8:30:00.000	1	5	bm	2	NCDC
26		02/03/2003 10:00:00.000	1	5	em	2	NCDC
27	10	02/03/2003 23:45:00.000	1	5		2	NCDC
28	10	02/03/2003 0:00:00.000	1	6	Inc	2	NCDC
29	10	02/04/2003 13:30:00.000	1	5		2	NCDC
30	10	02/04/2003 18:00:00.000	1	5		2	NCDC

Record: 1 of 37

ObservationTypes : Table

OBJECTID	ObservationTyp	Variable	Units	UnitType	IsRegular	ObsTimeSupp	TimeUnit	DataType
1	5	Precipitation	Hundredths of Inches	Depth	FALSE	15	Minute	Incremental
2	6	Precipitation	Hundredths of Inches	Depth	FALSE	24	Hour	Incremental

Record: 2 of 2

DataQualifierCodes : Table

OBJECTID	DataQualifierCode	Description
1	g	Only used for day 1, hour 0015 when precipitation is zero.
2	bm	Begin missing period during the 15 minute period (inclusive).
3	em	End missing period during the 15 minute period (inclusive).
4	inc	Incomplete or Inexact daily total occurring. Value is not a true 2

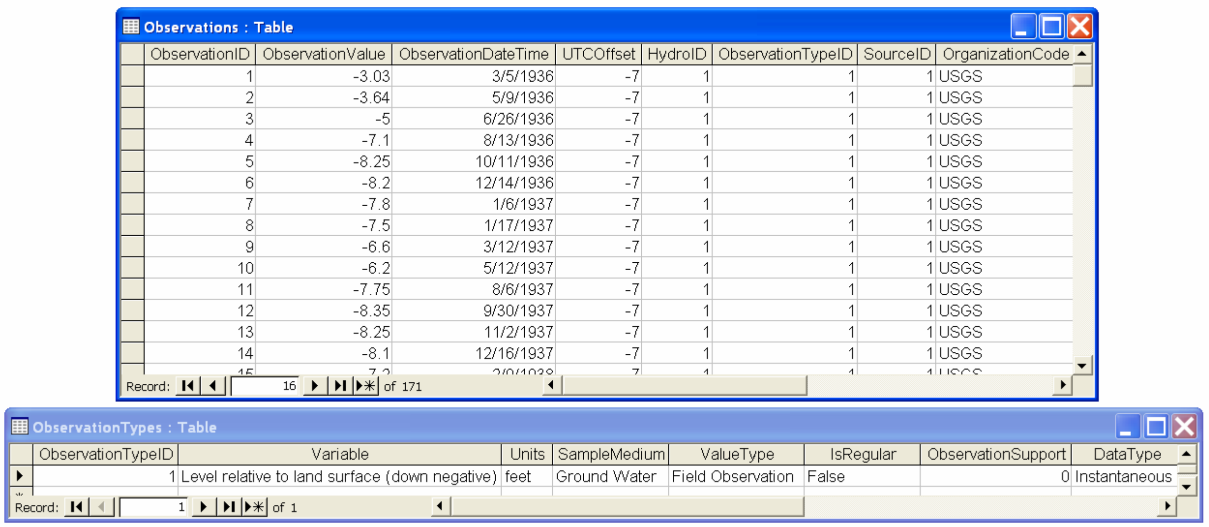
Record: 1 of 4

Incomplete or Inexact daily total occurring. Value is not a true 24-hour amount. One or more periods are missing and/or an accumulated amount has begun but not ended during the daily period.

Figure A.3. Excerpts from tables illustrating the population of the data model with NCDC Precipitation Data.

Groundwater Level

The following is an example of how groundwater level data can be stored in the proposed database structure.



Observations : Table

ObservationID	ObservationValue	ObservationDateTime	UTCOffset	HydroID	ObservationTypeID	SourceID	OrganizationCode
1	-3.03	3/5/1936	-7	1	1	1	USGS
2	-3.64	5/9/1936	-7	1	1	1	USGS
3	-5	6/26/1936	-7	1	1	1	USGS
4	-7.1	8/13/1936	-7	1	1	1	USGS
5	-8.25	10/11/1936	-7	1	1	1	USGS
6	-8.2	12/14/1936	-7	1	1	1	USGS
7	-7.8	1/6/1937	-7	1	1	1	USGS
8	-7.5	1/17/1937	-7	1	1	1	USGS
9	-6.6	3/12/1937	-7	1	1	1	USGS
10	-6.2	5/12/1937	-7	1	1	1	USGS
11	-7.75	8/6/1937	-7	1	1	1	USGS
12	-8.35	9/30/1937	-7	1	1	1	USGS
13	-8.25	11/2/1937	-7	1	1	1	USGS
14	-8.1	12/16/1937	-7	1	1	1	USGS
15	-7.2	2/9/1938	-7	1	1	1	USGS

Record: 16 of 171

ObservationTypes : Table

ObservationTypeID	Variable	Units	SampleMedium	ValueType	IsRegular	ObservationSupport	DataType
1	Level relative to land surface (down negative)	feet	Ground Water	Field Observation	False	0	Instantaneous

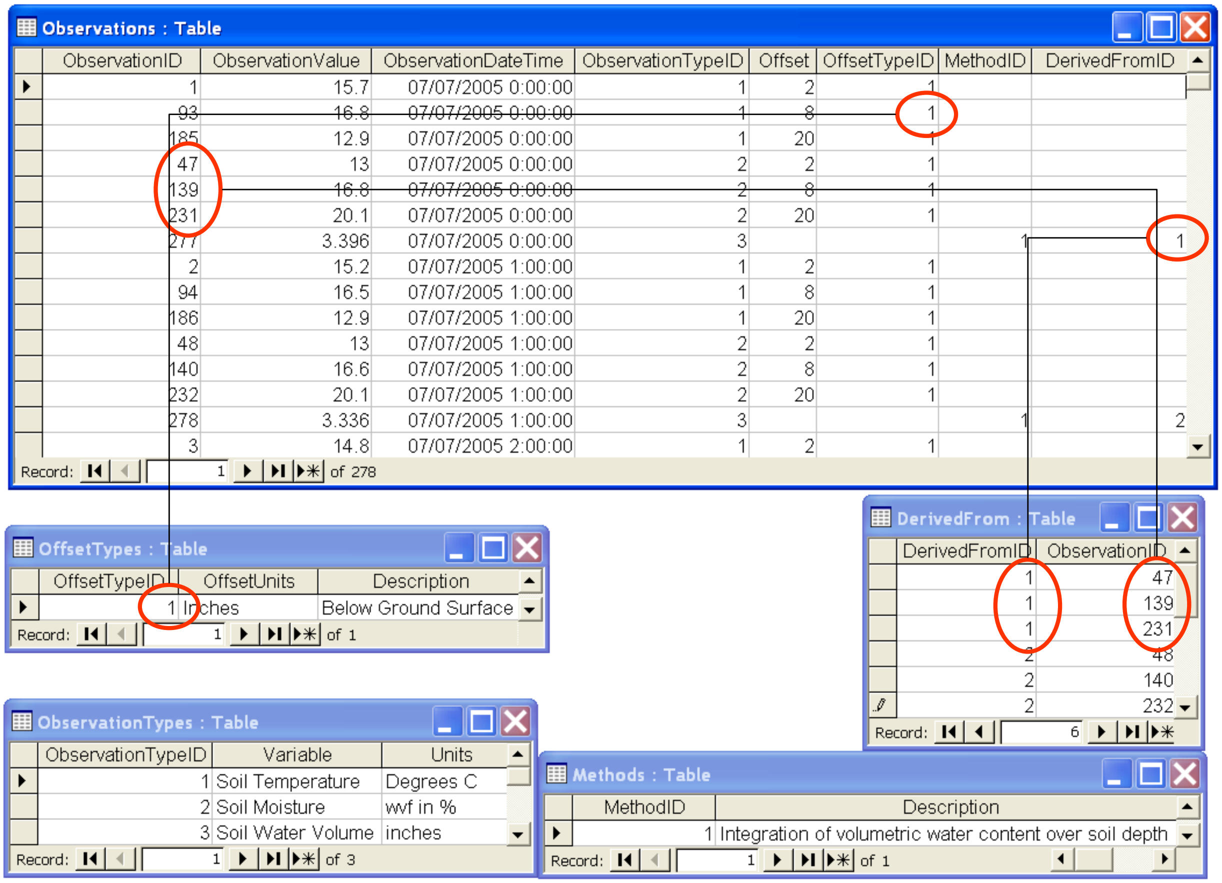
Record: 1 of 1

Figure A.4. Excerpts from tables illustrating the population of the data model with irregularly sampled groundwater data.

In this groundwater level example observations are depth relative to the ground surface reported as negative values.

Soil Moisture Sampled from a Depth

Soil moisture and soil temperature are examples of quantities that may be measured over a range of depths at a sampling location. The following (Figure A.6) is an example of how these data can be stored using the proposed database structure.



Observations : Table

ObservationID	ObservationValue	ObservationDateTime	ObservationTypeID	Offset	OffsetTypeID	MethodID	DerivedFromID
1	15.7	07/07/2005 0:00:00	1	2	1		
93	16.8	07/07/2005 0:00:00	1	8	1		
185	12.9	07/07/2005 0:00:00	1	20	1		
47	13	07/07/2005 0:00:00	2	2	1		
139	16.8	07/07/2005 0:00:00	2	8	1		
231	20.1	07/07/2005 0:00:00	2	20	1		
277	3.396	07/07/2005 0:00:00	3			1	1
2	15.2	07/07/2005 1:00:00	1	2	1		
94	16.5	07/07/2005 1:00:00	1	8	1		
186	12.9	07/07/2005 1:00:00	1	20	1		
48	13	07/07/2005 1:00:00	2	2	1		
140	16.6	07/07/2005 1:00:00	2	8	1		
232	20.1	07/07/2005 1:00:00	2	20	1		
278	3.336	07/07/2005 1:00:00	3			1	2
3	14.8	07/07/2005 2:00:00	1	2	1		

Record: 1 of 278

OffsetTypes : Table

OffsetTypeID	OffsetUnits	Description
1	Inches	Below Ground Surface

Record: 1 of 1

DerivedFrom : Table

DerivedFromID	ObservationID
1	47
1	139
1	231
2	48
2	140
2	232

Record: 6

ObservationTypes : Table

ObservationTypeID	Variable	Units
1	Soil Temperature	Degrees C
2	Soil Moisture	wvf in %
3	Soil Water Volume	inches

Record: 1 of 3

Methods : Table

MethodID	Description
1	Integration of volumetric water content over soil depth

Record: 1 of 1

Figure A.5. Excerpts from tables illustrating the population of the data model with soil moisture and temperature data collected over a profile into the soil.

In this example at each DateTime there are 3 measurements of soil moisture at depths (2, 8 and 20 inches) and 3 measurements of soil temperature at these same depths. The OffsetTypes table indicates that these measurements refer to depth below the ground. There is a single derived soil water volume obtained by integrating soil moisture over the soil profile. The methods table describes the method and the DerivedFrom table gives the groups of three soil moisture measurements that were used in deriving each soil water volume value, illustrating how this model works when groups of variables are used in obtaining a derived quantity.

Chapter 7

Remote Sensing

Ujjwal Narayan, Rahul Kanwar and Venkat Lakshmi
University of South Carolina,
Department of Geological Sciences,
Columbia SC 29208

Section I: Introduction

Hydrological modeling has been undergoing an exciting phase of transformation driven by rapid advances in remote sensing technology and computing power. In particular, NASA’s Earth Observation System (EOS) suite of satellite platforms and sensors has made available a variety of biophysical variables that have the potential to immensely advance the science and application of hydrology. These sensors use a variety of remote sensing technologies to make measurements of hydrological variables at several scales of spatial and temporal resolution. Remote sensing can be defined as the science and art of obtaining information about an object, area, or phenomenon through the analyses of data acquired by a sensor that is not in direct contact with the target of investigation [1]. In step with the availability of data obtained from satellite and/or in situ instruments, hydrologists have taken advantage of processing and storage abilities of present day computers to develop global, regional and local hydrologic models that ingest these data and allow increasingly accurate simulation and forecasting of hydrological processes. Analysis and modeling of hydrological processes has undergone a paradigm shift from a spatially-lumped approach to a spatially-distributed approach. Hydrologic phenomena of critical interest to the society such as stream flow at catchment’s outlets, contaminant transport, groundwater recharge, and weather and climate prediction are being studied using process models that are a more realistic representation of the actual physical processes that occur on the Earth rather than conceptual and statistical models that were constrained by calibration to historical in situ observations. These activities of hydrologists are of paramount importance as the need to be able to understand and predict the role of human activities in changing the hydrological processes at local and global scales is critical and urgent.

Even a brief survey of the variety of both remotely sensed and in situ data sources available to a hydrologist would illustrate the myriad of data formats, access techniques, data quality issues and temporal and spatial extents. As more and more watersheds become gauged and satellite instruments get deployed, it is very important to make data availability and usage is as stream lined as possible for potential users. The CUAHSI Hydrologic Information System (HIS) initiative “aims to provide better access to hydrologic data and provide the information technology needed to formulate and test new hydrologic science research hypothesis”. Figure 1 illustrates various components of the CUAHSI HIS and their inter-relationships. Information source or hydrologic data forms the central component of CUAHSI/HIS encompassing the various bodies of data that are needed to conduct a particular investigation [2]. Point measurements of water and energy fluxes make up the “more conventional” sources of hydrologic data. However, remote sensing is rapidly emerging technology for making hydrologic observations. Remotely sensed data are very different from point measurements in terms of their spatial and temporal coverage, sensing depths, quality issues, need for calibration and validation in some cases, data volume and data formats. Remote sensing data will provide a major component of current and future hydrologic measurements, especially so in ungauged watersheds. While the CUAHSI-HIS is in its formative stage, it is a worthwhile exercise to examine how hydrologic observations through remote sensing can fit into the information system and their benefits thereof. This paper will explore sources, data formats, data products, metadata issues of remotely sensed hydrologic data available and propose a Digital Library System (DLS) that would allow a seamless integration of remotely sensed hydrologic observations into the Hydrologic Information System.

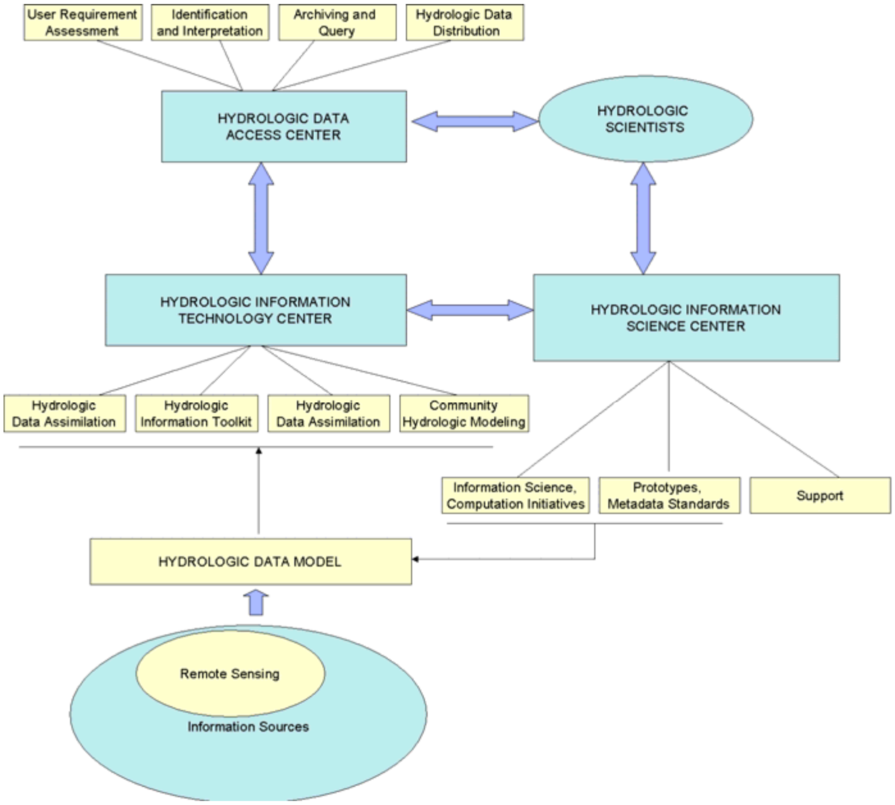

User Requirement Assessment
Identification and Interpretation
Archiving and Query
Hydrologic Data Distribution
HYDROLOGIC DATA ACCESS CENTER
HYDROLOGIC SCIENTISTS
HYDROLOGIC INFORMATION TECHNOLOGY CENTER
HYDROLOGIC INFORMATION SCIENCE CENTER
Hydrologic Data Assimilation
Hydrologic Information Toolkit
Hydrologic Data Assimilation
Community Hydrologic Modeling
Information Science, Computation Initiatives
Prototypes, Metadata Standards
Support
HYDROLOGIC DATA MODEL
Remote Sensing
Information Sources

Figure 1: Various components of the CUAHSI-HIS. Remote sensing forms an important part of Information Sources

Hydrologists need to understand how remote sensing observations should be utilized for hydrologic research and applications. On the other hand researchers involved with development of novel technologies need in situ hydrologic observations for testing and validating their hypotheses, methods and data products. Integration of remotely sensed data into the HIS initiative holds the promise to be an immense gain in capabilities for both Hydrologic and Remote Sensing communities.

The next section explores currently operational satellite platforms and the range of hydrological observations available. In section III an overview of the HDF-EOS data format is provided. HDF-EOS is the primary data format for remote sensing data in earth sciences and is probably not very well understood by hydrologist with little experience in remote sensing. In section IV we propose the building blocks of a digital library system that will host and provide access to remote sensing hydrologic data. In the design we have try to overcome deficiencies of current data providers with the goal of making remote sensing data available to the hydrology community with minimum exposure to remote sensing jargon, 'websitology', better decision making tools and most importantly provide data access through webservices rather than storage media, ftp push/pull etc. so that data streams can be directly integrated with applications using them.

Section II: Platforms and Data

The NASA's Earth Observing System (EOS) was designed to initiate a new era of integrated global observations intended to advance understanding of the entire Earth system on a global scale through a deeper understanding of the components of the Earth system, their interactions and how the Earth is changing [3]. The terrestrial biosphere forms an integral component of EOS with science objectives concerning climate change, hydrologic cycle change and changes in the terrestrial productivity. In order to fully capture the range of spatial and temporal variability in these processes and associated physical variables, EOS introduced a suite of satellites that accurately quantify this variability. Among the most current satellites launched by NASA are Terra and Aqua satellite launched in December 1999 and May 2002. Furthermore, there are a variety of satellites launched by Japan (ADEOS II), Europe (ENVISAT), India (INSAT) which provide global coverage using different sensors but sense similar variables at different overpass times. Remote sensing data is not limited to being acquired from satellite platforms solely. Lidar (LIght Detection And Ranging) is an example where aircraft mounted Lidar instruments are used to typically map the topography at high spatial resolutions. Hyperspectral remote is used to measure the spectral response of the target is measured in several fine channels leading to identification of land cover types. Most Hyperspectral remote sensing instruments are also mounted on aircrafts can be used to map the region of interest with high spatial resolutions. Ground based radars very commonly used to measure precipitation, ground based Lidars are being experimented in their use to measure water vapor profiles [4]. Figure 2 provides an overview of currently available and planned sensors and their realms of hydrologic observation. As shown in the figure, remote sensing provides measurements of most of the variables involved in the hydrologic cycle allowing the hydrologists to analyze energy and mass balance almost globally.

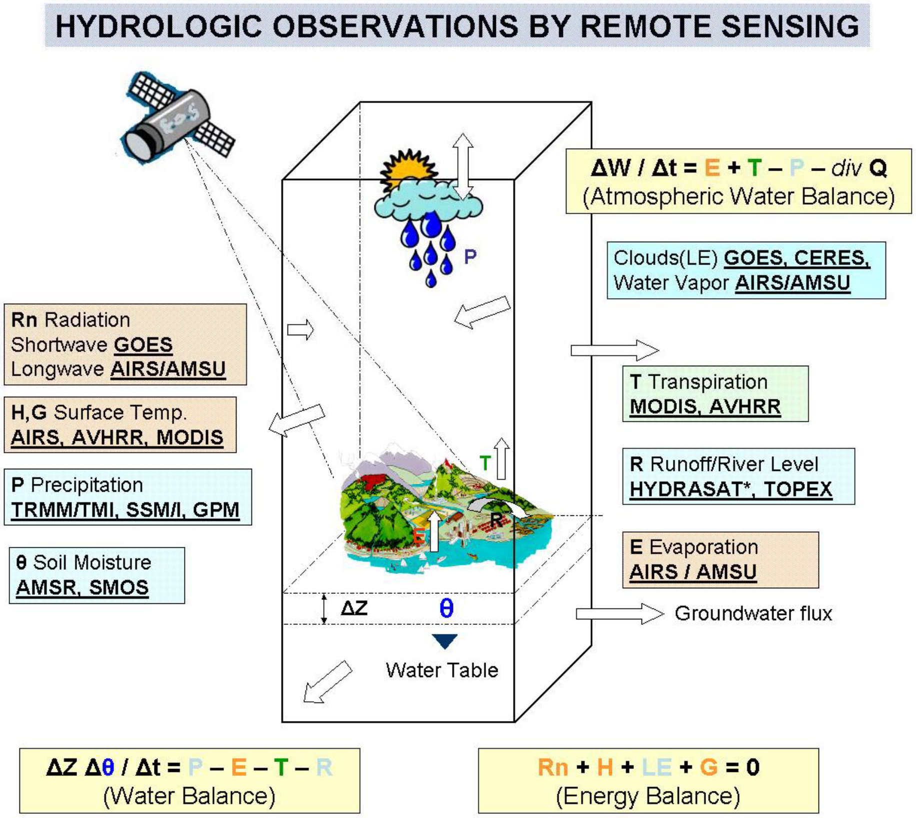


Figure 2: Remote sensing instruments associated with making observations of various hydrological variables of the hydrologic cycle. They provide the capability of closing the terrestrial and atmospheric energy and water budget.

Moderate Resolution Imaging Spectrometer (MODIS)

MODIS is aboard the Terra and Aqua satellite platforms and with a swath of 2,330 kilometers provides almost daily global coverage in 36 distinct spectral bands. MODIS greatly improves on the heritage of AVHRR and LandSat by making global measurements of vegetation, surface temperature, and land-cover at a maximum spatial resolution of 250 meters. Both Terra and Aqua have sun synchronous orbits (same local overpass time for points on the equator) with overpass times of 10:30 am for Terra and 1:30 pm for Aqua. The 10:30 am overpass is time is chosen as the best time to allow cloud free viewing and the 1:30 pm overpass time is chosen to allow full development of the atmospheric and planetary boundary layers. MODIS provides daily measurements of land surface temperature at 1 km and 5 km spatial resolutions with an accuracy of 1 K, 96-day land cover type and change at 0.05 degree resolution, 16 day and monthly averages of vegetation indices at spatial resolutions of 250 m, 500 m, 1 km and 0.05 deg, as well as 8-day leaf area indices (LAI) at a 1 km and 0.05 degree spatial resolutions.

The physics based day/night LST algorithm is used to simultaneously retrieve surface band emissivity and temperature from a pair of daytime and nighttime MODIS observations. The land cover parameter identifies 17 categories of land cover following the IGBP (International Geosphere Biosphere Programme) global vegetation database, which defines nine classes of natural vegetation, three classes of developed lands, two classes of mosaic lands, and three classes of non-vegetated lands (snow/ice, bare soil/rocks, water). (*http://modis.gsfc.nasa.gov*). MODIS vegetation indices (VI) products provide global maps of photosynthetically active vegetation vigor allowing monitoring of their spatial and temporal characteristics. The VI products contain two indices, the Normalized Difference Vegetation Index (NDVI) and a new Enhanced Vegetation Index (EVI). The NDVI provides continuance to the AVHRR that provided around 30 years of vegetation data. EVI is a MODIS-specific index and offers improved sensitivity in high biomass regions and improved vegetation monitoring by incorporating further corrections for the canopy background signal and atmospheric influences. Further details about MODIS can be found in [5].

Geostationary Operational Environmental Satellites (GOES)

The GOES series of satellites carry two remote sensing instruments, GOES imager and GOES sounder. The GOES imager has the ability to scan 3000 by 3000 km "box" centered over the United States in just 41 seconds that translates into the real time monitoring capabilities. The imager has five channels, namely the Visible, Shortwave, Moisture, IR1 and IR2 with instantaneous fields of view (IFOV, akin to spatial resolution) of 1 km, 4 km, 8 km, 4 km and 4 km respectively. It is useful for cloud identification, moisture content, heavy precipitation, and atmospheric motion studies. The GOES sounder has the ability to map temperature profiles, moisture profile, and ozone content and reflected solar radiation via its 18 thermal infrared and 1 visible band. The sounder measures these variables at 10 km – 50 km spatial resolution (depending on product) and in 40 pressure levels in terms of vertical profiling of the atmosphere. Currently, the GOES system consists of GOES-12 operating as GOES-East at 75° west longitude, and GOES-10 operating as GOES-West at 135° west longitude. Further details about GOES can be found in [6].

Advanced Microwave Scanning Radiometer (AMSR)

AMSR and AMSR-E (Advanced Microwave Scanning Radiometer for EOS) were launched aboard ADEOS-II (NASDA) and Aqua (NASA) and provide morning (10:30 am) and afternoon (1:30 pm) measurements of water vapor, cloud liquid water, precipitation, sea surface temperature, sea surface wind speed, sea ice concentration, snow water equivalent, and soil moisture. The two-satellite combination allows frequent sampling of hydrological phenomena with high temporal variability. AMSR-E was an integral part of the Aqua mission that was specifically designed to help scientists better understand the impact of climate change on the water cycle. AMSR-E provides five level 3 products on a daily, weekly and monthly basis with global coverage. Ocean products 0.25 x 0.25 degree spatial resolution measurements of sea surface temperature, cloud liquid water, wind speed and atmospheric water vapor over ocean. The Land product is available daily at a 25 km spatial resolution and provides estimates of soil moisture, surface temperature and vegetation water content. Care must be taken in interpreting the soil moisture product, as it is valid for regions with low vegetation water content only. A snow water equivalent product is also available at 25 km spatial resolution and daily, 5-day and monthly temporal scales. Global rain product provides monthly rainfall accumulation on 5 x 5 degree grids for both land and ocean. Sea ice products are available from AMSR-E at three resolutions: 6.25 km, 12.5 km, and 25 km. Further details about AMSR can be found in [7, 8].

Atmospheric Infrared Sounder (AIRS), Advanced Microwave Sounding Unit (AMSU-A)

AIRS is a part of the Aqua mission that was launched in May, 2002. The science objective of AIRS and AMSU-A was to study the global water and energy cycles, climate variability and effect of greenhouse gases on climate. AIRS has infra red channels with a spectral coverage from 3.7 to 15.4 µm and AMSU-A works in the microwave frequency range of 27 to 89 GHz. AIRS, AMSU-A and the Humidity Sounder for Brazil (HSB) on the Aqua mission together have the capability of measuring the atmospheric temperature in the troposphere with radiosonde accuracies of 1 K over 1 km-thick layers under both clear and cloudy conditions, while the accuracy of the derived moisture profiles exceeds that obtained by radiosondes. Land and ocean surface temperature, surface emissivity, cloud fraction and cloud top height are also estimated using AIRS/AMSU-A. Temperature profiles (24 levels), moisture profiles (2 layers), cloud cleared outgoing long-wave radiation and surface temperatures (air, skin) are estimated as Level 3 products at a 1 x 1 degree spatial resolution and daily, 8-day and monthly temporal resolutions. Further details about AIRS can be found in [9].

Tropical Rainfall Measuring Mission (TRMM)

TRMM is aimed at obtaining measurements of tropical rainfall and understanding how this rainfall affects the global climate. Individual sensors that make up TRMM are the Precipitation Radar (PR), Tropical Microwave Imager (TMI), Visible and Infra-red Scanner (VIRS), Cloud and Earth Radiant Energy Sensor (CERES) and the Lightning Imaging Sensor (LIS). PR provides measurements of the 3-d structure of a storm estimating parameters such has rainfall intensity profile, storm depth, rain type and the height at which snow melts into rain. Its ground resolution is 2.5 km and can detect rainfall intensity as low as 0.7 mm/hr. TMI uses passive microwave measurements to estimate rain rates across a wider swath as compared to the PR. VIRS measures upwelling radiation in five channels in the visible and infra-red frequencies and is used to delineate rainfall. CERES measures energy at the top and within the atmosphere and its data is used to derive estimates of cloud height, thickness, particle size and other properties. Further details about TRMM can be found in [10].

A more comprehensive list has been provided in Table I with information about data products, satellite platform, sensor name, channels (frequency or wavelength) of operation, spatial resolution and revisit time periods. We describe in the rest of this section some of the important operational satellite instruments aimed at observing hydrologic processes and their data products.

Snow extent	NOAA series	AVHRR	.62, 10.8 µm	1 Km	2 / day
	SPOT	HRV	0.59, 0.69, 0.89 µm	10 - 25 m	26 days (steerable)
	TERRA / AQUA	MODIS	620 - 2155 µm	250 - 1000 m	1 - 2 days
	GOES	I-M	0.55 - 0.75 µm	1 km	2 / hour
Snow Depth	NIMBUS 7	SMMR	18.37 GHz	30 km	2 / day
SWE	DMSP	SSM/I	19.3, 37 GHz	25 km	2 / day
	AQUA	AMSR	18.7, 36.5 GHz	21 km	Daily
	MOS-1	MSR	23, 31 GHz	23 - 32 km	2 / day
Snowmelt	ERS - 1,2	SAR	C band (5.3 GHz) VV	30 m	35 days
	Radarsat-1	SAR	C band (5.3 GHz) HH	8 - 25 m	3 - 16 days
	Envisar	Advanced SAR	C band (5.3 GHz)	30 - 150 m	35 days (steerable)
			HH, HV, VV, VH		
Landcover / Veg	Landsat	TM	0.52, 0.6, 0.69, 0.9	80 m	8 - 16 days
			1.75, 2.35, 12.5 µm		
	Landsat	MSS	0.55, 0.65, 0.75, 0.9 µm	80 m	8 - 16 days
	NOAA series	AVHRR	0.62, 0.91 µm	1 km	2 / day
	SPOT	HRV	0.59, 0.69, 0.89 µm	10 - 25 m	26 days (steerable)
Soil Moisture	AQUA	AMSR	6.9, 10.7 GHz	36 - 58 km	Daily
	Radarsat - 1	SAR	C-band (HH)	8 - 25 m	3 - 16 days
	Envisat	Advaned SAR	C-band (5.3 GHz)	30 - 150 m	35 days (steerable)
			HH, VV, HV, VH		
Surface Water	SPOT	HRV	0.59, 0.69, 0.89 µm	10 - 25 m	26 days (steerable)
	ERS - 1,2	SAR	C band (5.3 GHz) VV	30 m	35 days
	Envisat	Advaned SAR	C-band (5.3 GHz) HH	30 - 150 m	35 days (steerable)
	Landsat	TM	0.48, 0.56, 0.66 µm	30 m	8-16 days
	DMSP	SSM/I	19.3, 37, 85 GHz	25 km	Daily

Table I: Variables and satellite sensors with spatial resolution and temporal resolution and repeat cycle [11].

Section III: Hierarchical Data Format for Earth Observation Sciences (HDF-EOS):

Having introduced some of the currently available remote sensing platforms and data products we next discuss the HDF-EOS data format. Most of remote sensing data available from NASA's EOS satellites are being provided in the HDF-EOS format. Table II provides a listing of some of the data format, volume and providers for some remote sensing missions.

SENSOR	DATA VOLUME	DATA FORMAT	DATA PROVIDER
MODIS	600 GB / day (L1B)	HDF - EOS	GES DISC DAAC
AMSR	2.5 GB / day (L2A)	HDF - EOS	EOS-DG , NSIDC DAAC
AIRS/AMSRU	30 GB / day (L1B)	HDF-EOS	GES DISC DAAC
GOES	40 GB / day	McIDAS, GVAR	NWC SEC
AVHRR	1 GB / day	GAC, LAC, BINARY	NOAA CLASS
TRMM	14 GB / day	HDF - EOS	GES DISC DAAC
ASTER	80 GB/day	HDF - EOS	EOS-DG

Table II: Data Characteristics

It is seen that HDF-EOS is the most prevalent data format with only GOES and AVHRR data in other formats, mainly because HDF-EOS was developed after these missions became operational. Hierarchical Data Format (HDF) is currently the standard data format for all NASA

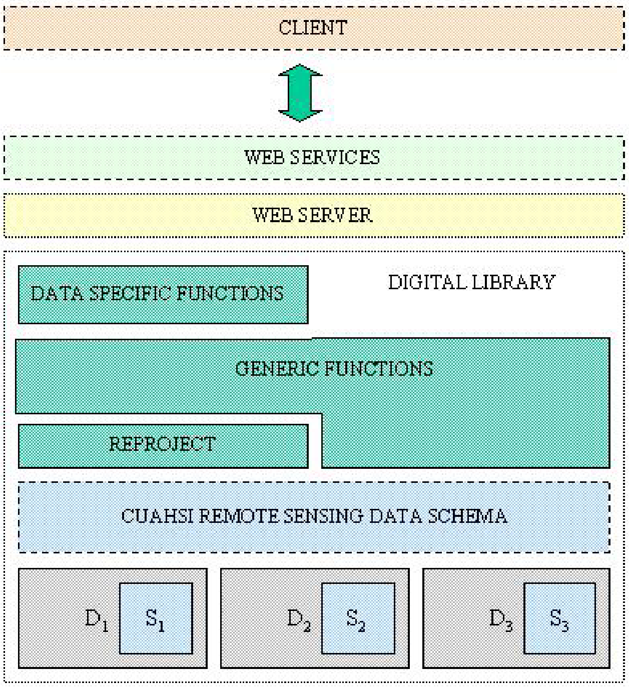
EOS data products. HDF-EOS is based on the more generic, HDF data format, and an underlying HDF library. It is a multi-object file format developed at the National Center for Supercomputing Applications (NCSA) in order to assist users in the transfer and manipulation of scientific data across diverse operating systems and computer platforms. HDF contains a variety of simpler data formats and is supported by an HDF library that contains interfaces for storing and retrieving data stored in these formats. Some of the different data formats are tables (n-dimensional arrays), raster images, color palettes, text, etc. XML tags are used to define the type, amount, data dimensions and file location of various objects giving HDF files a self-describing capability that helps users to fully understand the file's structure and contents from the information stored in the file itself. HDF-EOS is essentially HDF that supports geolocation, time stamping and defines certain metadata placement and type. HDF-EOS API defines a mandatory inclusion of ECS (EOSDIS Core System) Core metadata. Three geospatial data types are supported –

- **Point Data Types** - Irregularly spaced in time and/or space (such as gauging data).
- **Swath Data Types** - Time-ordered satellite data, representing time sequences of scan lines, profiles (such as vertical profiles), or other array data.
- **Grid Data Types** - Regularly gridded data, grid based on certain Earth/map projection.

HDF-EOS libraries allow operations such as geographical subsetting, access to time information etc. with much greater ease than would be possible if one were using simply HDF data format. HDF itself has several versions the most recent of which is called HDF5. HDF5 improves over HDF4 by providing simpler source codes, more consistent and fewer data models, and the ability to work with large data sets (> 2GB). A new version of the HDF-EOS Library, called HDF-EOS3.0, is totally based on the new HDF5 Library with an entirely different functionality. With HDF-EOS3.0 the structural Metadata/geolocation information will be contained within HDF5 objects where as with HDF-EOS2.6 the user accesses structural metadata to map geolocation data with the scientific data. Data for the TERRA and AQUA sensors are provided using the HDF4 Library (and, thus, HDF-EOS 2.6 if HDF-EOS is used).

Section IV: Digital Library Server

The structure of the Digital Library System (DLS) for remote sensing data has been represented in the block diagram (Figure 3).



Our goals in the design and implementation of the DLS are

- Data should be accessible through webservices
- All functions will be implemented as webservices
- Metadata obtained from the remote sensing data provider should be mapped onto the CUAHSI metadata schema
- Unnecessary remote sensing ‘jargon’ will be hidden from the user. While requesting data at the NASA DAAC’s for example, the primary search criteria is sensor name. Our implementation of the remote sensing digital library server will have hydrological parameters as the primary search criteria. (Give me soil moisture data as opposed to give me AMSR-E data, which not everyone knows, provides soil moisture measurements!)
- Using gridded data and a uniform map projection for all data products, subsetting operations across multiple datasets (such as all soil moisture data for regions with vegetation water content of less than 2.5 kg/m^2) will be supported. Current data providers do not implement such requests.
- The goal of the DLS is not to provide new data products. However, *Generic functions* such as subsetting, resampling, buffering, collocation, masking will be implemented.
- Data product *specific functions* will be implemented (for example, a user will be allowed to generate a composite precipitation rate map by combining estimates from AMSR-E and TMI which will lead to better temporal coverage over the study area. Another example maybe that user interested in change in MODIS EVI rather than the absolute values will be provided with the change image)
- A rich visual interface will be provided on the *client* to allow an easy and optimal selection of datasets by the hydrologist. The client will work on the CUAHSI remote sensing metadata associated with each dataset. Hence it will be

independent of the server providing the data – it maybe ArcIMS, the SRB based Digital Library System (from SDSC), Mapserver etc. The client has to allow the user to explore the digital library using metadata attributes and using tools such as – a time series plot indicating temporal coverage of data products, size and spatial resolution inter-comparison of different datasets etc.

Now we will discuss the components of the DLS in more detail. The (D_i, S_i) pairs at the bottom of Figure 3 represent data input to the DLS from various data providers. Most of the freely available remote sensing data for hydrology comes from NASA's Distributed Active Archive Centers (DAAC) and is provided in the form of gridded HDF-EOS files. As a reference, table III provides a list of various NASA DAAC's that archive and distribute remote sensing data for hydrologic sciences.

NASA Distributed Data Archive Centers

- Alaska Satellite Facility (ASF) DAAC – Synthetic Aperture Radar (SAR), Sea Ice, Polar Processes
- GSFC Earth Sciences Center (GES) DAAC – Upper Atmosphere, Atmospheric Dynamics, Global Precipitation, Global Biosphere, Ocean Dynamics, Solar Irradiance
- Global Hydrology Resource Center (GHRC) DAAC – Hydrologic Cycle, Severe Weather Interactions
- Land Processes(LP) DAAC – Land Processes
- NASA Langley Atmospheric Sciences Data Center (LARC) DAAC – Radiation Budget, Clouds
- National Snow and Ice Data Center (NSIDC) DAAC – Snow and Ice, Cryosphere and Climate
- Oak Ridge National Laboratory (ORNL) DAAC – Biogeochemical Dynamics, Ecological Data, Environmental Processes
- Physical Oceanography (PO) DAAC – Oceanic Processes, Air-Sea Interaction

Table III: NASA DAAC's that archive and distribute remote sensing data for hydrologic sciences.

While the data format is more or less uniform, there are still subtle differences between individual data products. For example, the geolocation fields of a data granule in ASTER surface temperature data product are called 'GeodeticLatitude' and 'Longitude' where as in most other data products they are named 'Latitude' and 'Longitude'. While this discrepancy seems trivial, the Digital library system will have to implement wrapper for core HDF-EOS metadata format as well as wrappers for structural and granular metadata formats for each data product so that they can be all mapped onto the same CUAHSI remote sensing metadata specification. Mapping HDF-EOS metadata to the CUAHSI metadata schema will allow development of functions that will operate on all remote sensing datasets within the DLS. These functions are represented by the '*Generic*' box in figure 3. Different datasets may have to be reprojected to the same map projection so as to allow inter comparison of datasets. This will be done on the fly depending on the nature of request made by the user. Typical examples of generic functions will be spatial and temporal subsetting, spatial and temporal resampling, simple buffer analysis (give me the precipitation data within 50 km of a particular rain gauge) etc. Scenario's such as a Hydrologist needing surface temperature estimates for areas within a watershed that have a vegetation water content of more than 2.5 kg/m^2 will be implemented, that is a data set may be subsetted based on the attributes of another dataset. Such an operation will require reprojection of datasets to the same map projection. Note that EOS DAAC does not provide such operations. Data specific functions will also have to be provided. For example, the ASTER surface temperature data product has a geolocation strategy wherein 11 x 11 arrays of latitude and longitude values get mapped to a data array of dimensions 830 x 700. So a function specific to the ASTER data product will be written to stretch the geolocation array to the data array dimensions.

Webservices are based on a producer consumer relationship between the server and the client. The webserver exposes its functionality to the world through webservices protocols such as SOAP, WSDL etc. that allows the client to make minimum assumptions about the underlying webserver and thus makes the client independent from the server implementation. This also enables the client to be developed in any operating system. This allows the integration of the datastreams directly into commercial products like ArcMap, Matlab, Excel etc. Other webservers can take advantage of provided webservices to produce new content. For example, a web application may consume soil moisture estimates obtained via a webservice provided by a DLS that archives remote sensing data and assimilate in situ measurements of soil moisture provided by another webservice to produce assimilated data product which will be available as a third webservice. The potential of a webservice being consumed by any application on any platform was the reason that our DLS implementation focuses on making data available through webservices as the highest priority task.

On the client side, we intend to provide the user with a representation of the entire contents of the library by interacting with a visual interface. For example, the user will be able to generate and view plots of time series of various data products so that a selection such as when can I get both surface temperature and precipitation data over my region of interest can be made. When storage space is a concern, simple plots of say spatial resolution versus data volume will allow a quick intercomparison of various sensors and the user and decide an optimal spatial resolution for an application. The client will also be an interface for the user to browse CUAHSI remote sensing metadata for the data products under consideration. User should be able to search metadata, inspect quality flags, cluster data granules based on metadata (give me all sensors that provide longwave radiation) and develop subsetting criteria (for what days of the year do I have precipitation data over the Neuse river basin). All webservices developed will be implemented on the client apart from being accessible programmatically.

In conclusion, remote sensing data will continue to grow in importance as a data source for hydrologic measurements. With the progress of the Internet as medium for dissemination of remote sensing data to intended users, it is important that integration of data to applications is as easy and streamlined as possible. Our DLS implementation aims to be proof of concept for such a service that would provide on demand programmatic access to remote sensing data across a variety of platforms and applications.

References

[1] Svehlak, D., Maidment, D., R, and Helly, J., “Technical Specifications for the CUAHSI Hydrologic Information System”, CUAHSI symposium, March 2005

[2] Consortium of Universities for the Advancement of Hydrological Science Inc. (CUAHSI) Hydrologic Information Systems, Prepared by the CUAHSI Hydrologic Information Systems Committee, Version October 7, 2002.

[3] “EOS Science Plan”, 1999. http://eospso.gsfc.nasa.gov/science_plan/index.php

[4] C. Senff, J. Bo¨senberg, and G. Peters, “Measurement of watervapor flux profiles in the convective boundary layer with lidar and radar-RASS”, J. Atmos. Ocean. Technol. 11, 85–93, 1994.

[5] Justice CO, Townshend JRG, Vermote EF, et al. “An overview of MODIS Land data processing and product status”, Remote Sensing of the Environment, 83 (1-2): 3-15, 2002

[6] Menzel, W. Paul, Purdom, James F.W., “Introducing GOES-I: The First of a New Generation of Geostationary Operational Environmental Satellites”, Bulletin of the American Meteorological Society 75: 757-781, 1994

[7] Njoku, E. G., T. Jackson, V. Lakshmi, T. Chan and S. Nghiem, “Soil moisture retrieval from AMSR-E”, IEEE Transactions on Geoscience and Remote Sensing, 41(2), pp215-229, 2003

[8] Wilheit, T., Kummerow, C. and R. Ferraro, “Rainfall algorithms for AMSR-E”, IEEE Transactions on Geoscience and Remote Sensing, 41(2), pp204-214

[9] Susskind, J., Barnet, C., and Blaisdell, J., “Retrieval of atmospheric and near surface parameters from AIRS/AMSU/HSB data in the presence of clouds”, IEEE Transactions on Geoscience and Remote Sensing, 41(2), pp390-409, 2003

[10] Kummerow, C. and others, “The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit”, Journal of Applied Meteorology, Vol. 39, pp 1965-1982, 2000

[11] Pietroniro A, Leconte R. “A review of Canadian remote sensing applications in hydrology, 1999–2003”. Hydrological Processes 19:285–301, 2005.

Chapter 8

Digital Watershed for the Neuse Basin

Venkatesh Merwade, Gil Strassberg, and David Maidment,
Center for Research in Water Resources,
University of Texas at Austin
Jonathon Goodall
Nicholas School for the Environment,
Duke University
Praveen Kumar and Ben Ruddell
University of Illinois at Urbana-Champaign

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) carried out a paper prototype study of the design of a Hydrologic Observatory using the Neuse watershed in North Carolina as their illustrative example (Reckhow et al., 2004). During that study a considerable amount of GIS and hydrologic observation data were compiled for the watershed by the Center for the Analysis and Prediction of River Basin Environmental Systems at Duke University (http://www.env.duke.edu/cares/).

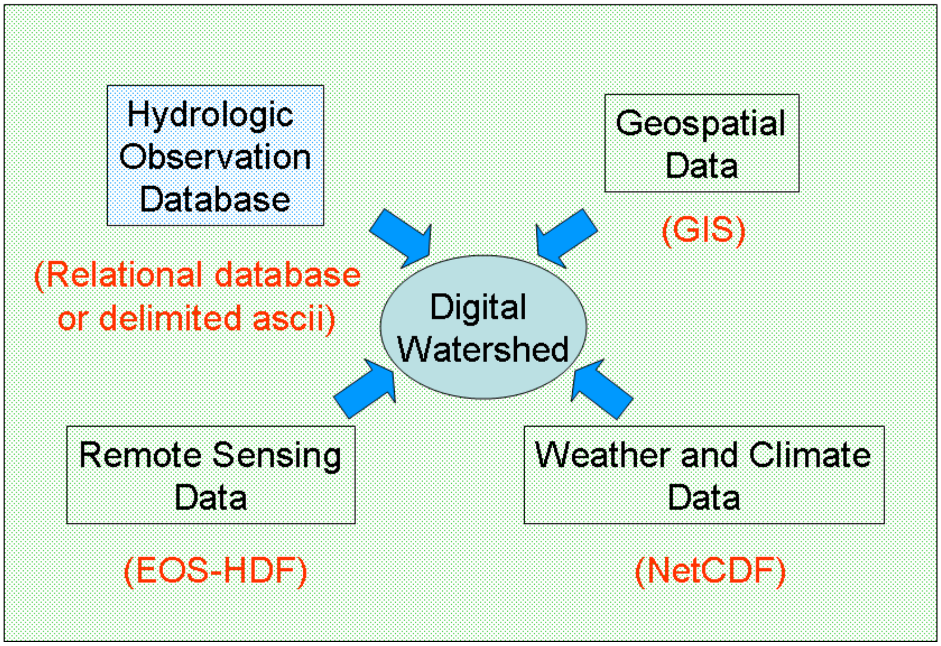
The CUAHSI Hydrologic Information System (HIS) team has added further information to this dataset, including 3D models of the hydrogeology of the Neuse coastal plain aquifer obtained from the USGS, time sequences of groundwater levels from the North Carolina Division of Water Resources interpreted to form piezometric head maps in the surficial aquifer, 3-hour and monthly land surface-atmosphere fluxes of energy and water from the North American Regional Reanalysis of climate, NEXRAD rainfall data from the National Weather Service, real-time water quality data collected by North Carolina State University, and a sequence of MODIS satellite images interpreted to show the time variation of greenness of the landscape. There is thus formed a rich and growing body of information that describes many aspects of the physical character and the hydrologic functioning of the Neuse basin.

The CUAHSI HIS team has termed the synthesis of hydrologic observations data, GIS data, weather and climate grids and remote sensing images a *Digital Watershed*. Each of these types of information comes in its own data formats, and spatial coordinates and time scales. By a process of *data fusion*, the various datasets can be transformed into a common set of geographic coordinates with a common time scale, and be synthesized into a set of compatible data formats so that they can be analyzed as a single large body of information. ArcGIS has been used as the data synthesis platform for this work.

The information in the Neuse Digital Watershed is presented in three datasets: *AtmosphericWater*, *SurfaceWater*, and *Groundwater*. These datasets are in the form of ArcGIS geodatabases that contain raster, vector and time series information presented in a way that makes all the information interoperable, that is, all the datasets are in the same geographic coordinates and time frame and they are in compatible data formats for analysis within ArcGIS. The

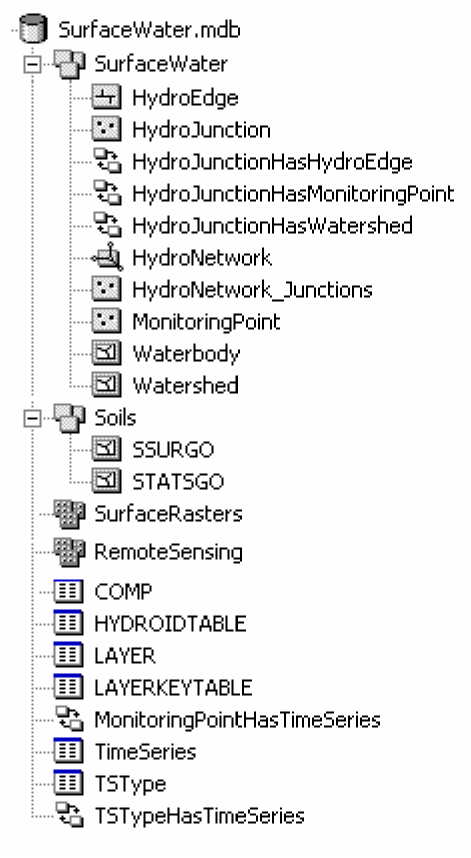
CUAHSI HIS team has also demonstrated how time series of hydrologic observation data from the Neuse Digital Watershed can be served on the internet using ArcIMS with output in the form of delimited ascii files, so the geospatial time series information is thus readily available for study in Excel and other hydrologic analysis systems.

The purpose of this paper is to describe the contents of the Neuse Digital Watershed as it stands at present. As the CUAHSI HIS project continues the Neuse Digital Watershed will be expanded to include elements such as three-dimensional representation of the stream channel and flood plain, and site specific studies being done with the Neuse watershed.



Description of Neuse SurfaceWater

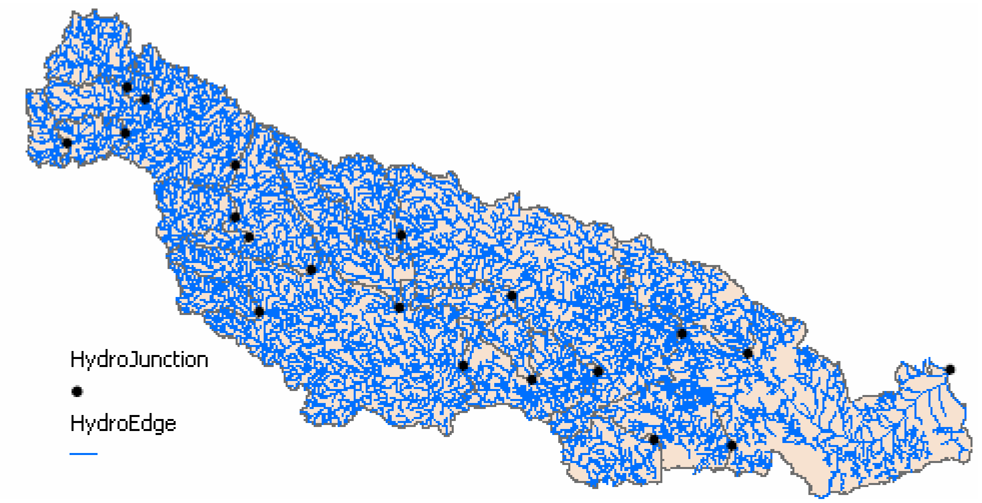
The SurfaceWater geodatabase for Neuse contains two feature datasets, two raster catalogs, time series table and tables related to soil data. A *feature dataset* is an ArcGIS folder with a defined coordinate system and geographic extent that contains a set of *feature classes*, which may include points, lines, polygons or volumes (multipatches). A *raster catalog* is a set of rasters such as for terrain, land cover or piezometric head that are indexed by a summary table. The time series information is contained in a modified form of the Arc Hydro time series format as a set of tables. The Arc Hydro time series model is explained in detail in Maidment (2002). In addition to geographic features and time series records, the SurfaceWater geodatabase contains relationships that link features to features, features to time series, etc. The structure of the SurfaceWater geodatabase is shown below:



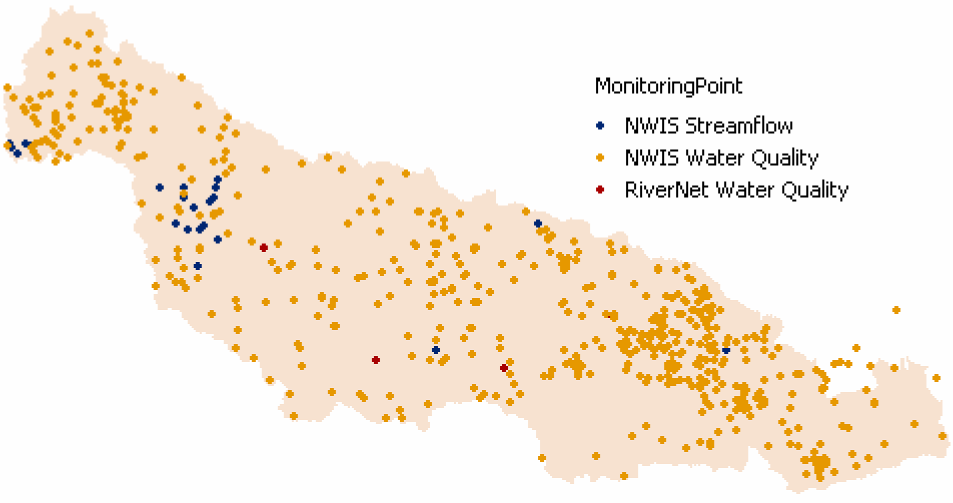
- SurfaceWater.mdb
  - SurfaceWater
    - HydroEdge
    - HydroJunction
    - HydroJunctionHasHydroEdge
    - HydroJunctionHasMonitoringPoint
    - HydroJunctionHasWatershed
    - HydroNetwork
    - HydroNetwork_Junctions
    - MonitoringPoint
    - Waterbody
    - Watershed
  - Soils
    - SSURGO
    - STATSGO
  - SurfaceRasters
  - RemoteSensing
  - COMP
  - HYDROIDTABLE
  - LAYER
  - LAYERKEYTABLE
  - MonitoringPointHasTimeSeries
  - TimeSeries
  - TSType
  - TSTypeHasTimeSeries

SurfaceWater

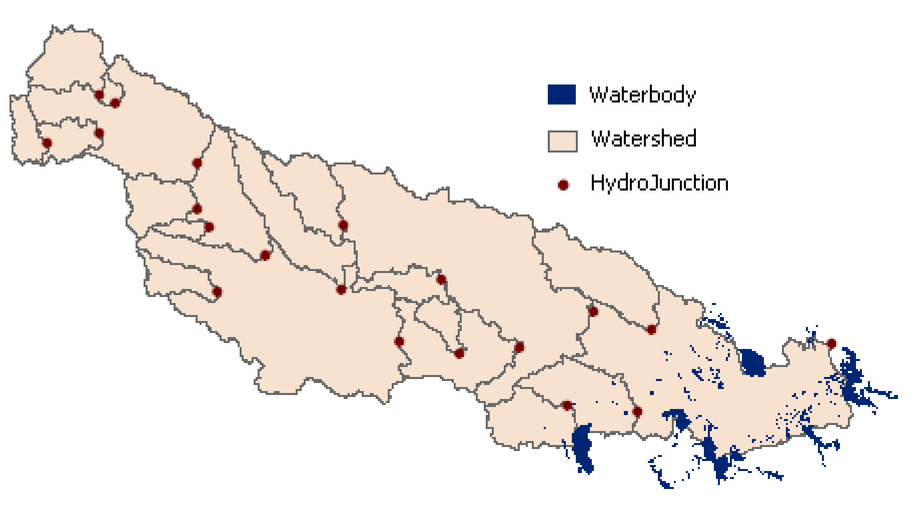
The **SurfaceWater** feature dataset contains six feature classes which can be populated using data from national and local sources. The **HydroEdge** feature class contains stream network for the Neuse basin created by using medium resolution (1:100000) flowlines from the National Hydrography Dataset (NHD). The NHD flowlines are pre-processed (removal of disjointed lines, looped features, etc) before exporting to the HydroEdge feature class. The **HydroJunction** feature class contains active NWIS stream flow gaging stations snapped onto HydroEdge features. HydroNetwork is the geometric network built from HydroEdge and HydroJunction. HydroNetwork_Junctions are the points that are created as a part of building HydroNetwork.



The **MonitoringPoint** feature class contains NWIS stream flow measurement stations (steamflow and water quality) and RiverNet points (water quality measurement points operated by North Carolina State University). Different types of points in the MonitoringPoint feature class are distinguished by assigning a different FType (feature type) attribute.

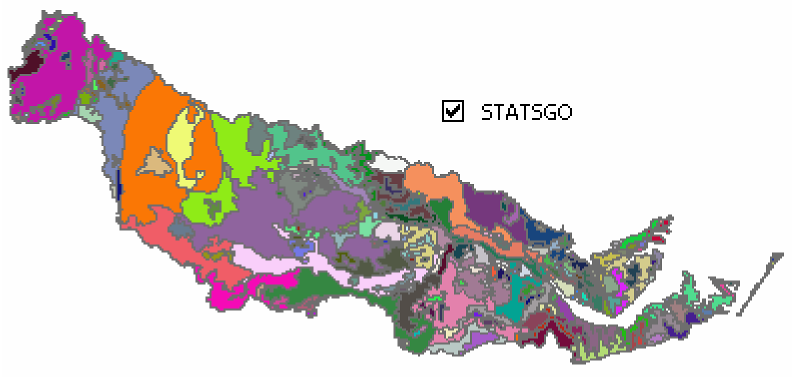


The **WaterBody** feature class contains water bodies (lakes, reservoirs, swamps/marshes) imported from NHD (NHDWaterbody). The **Watershed** feature class contains drainage areas for HydroJunction points created by using the Arc Hydro terrain processing tools.



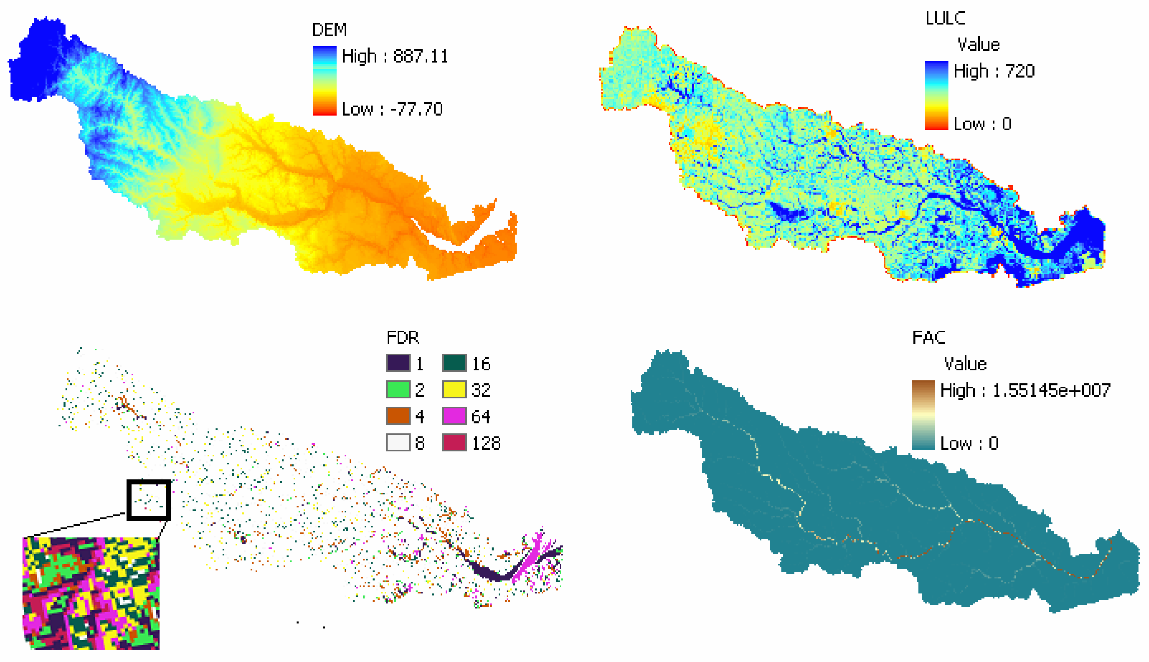
Soils

The Soils feature dataset contains two feature classes obtained from the National Resources Conservation Service (NRCS). **STATSGO** and **SSURGO** feature classes contain soils data from NRCS State Soil Geographic (STATSGO) database and Soil Survey Geographic (SSURGO) database, respectively.



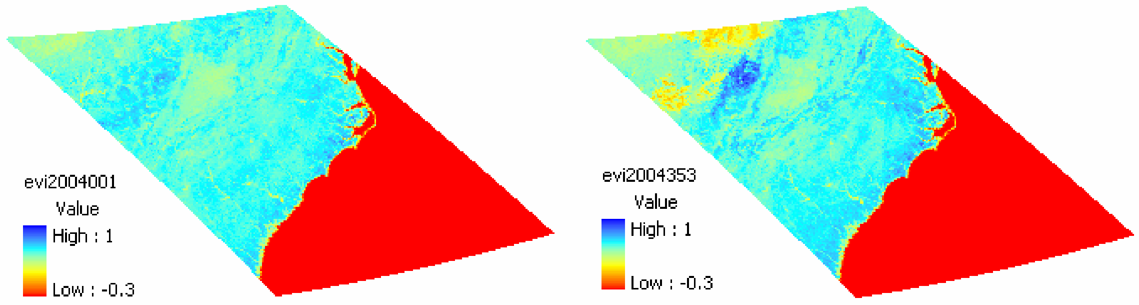
SurfaceRasters

The **SurfaceRasters** raster catalog contains four raster datasets with terrain, land use, and hydrologic descriptions. **DEM** is a 50 feet digital elevation model created by using LIDAR points obtained from the Center for the Analysis and Prediction of River Basin Environmental Systems at Duke University (http://www.env.duke.edu/cares/). **LULC** is the land use/land cover data from the Environmental Protection Agency. **FDR** and **FAC** are the flow direction grid and flow accumulation grid, respectively derived from DEM by using the Arc Hydro terrain processing tools.



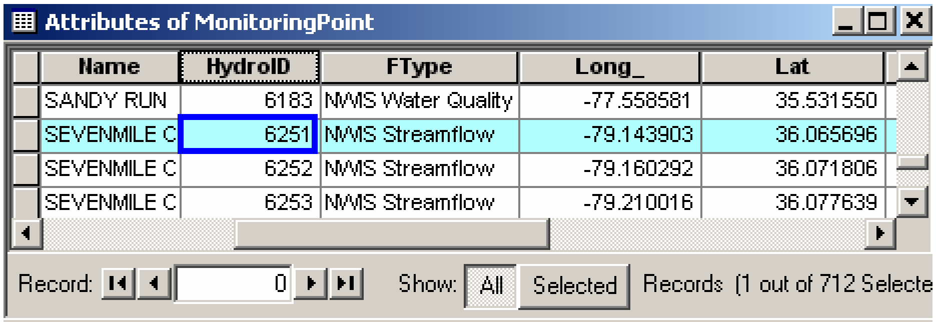
RemoteSensing

The **RemoteSensing** raster catalog contains 23 MODIS images stored as raster grids. The value for each cell in the MODIS grids represent enhanced vegetation index (EVI). Two sample grids with EVI on two different days around the Neuse basin are shown below:



TimeSeries table and Relationships in SurfaceWater

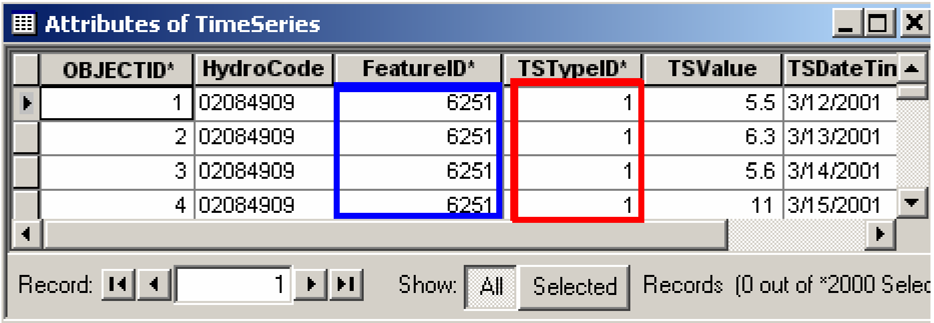
The **TimeSeries** table in the SurfaceWater geodatabase contains time series records for stream flow and water quality. The records in the TimeSeries table correspond to the geographic features in the MointoringPoint feature class. Relationships between features and time series records are established through HydroID, an ArcHydro attribute, which is unique for any geographic feature within a geodatabase. When a time series record is stored for a geographic feature, the FeatureID in the time series table is matched to the HydroID of the corresponding feature. Besides geographic features, the time series table is also related to TSType table, which stores the information about the time series data. The TimeSeries table is related to TSType table through TSTypeID.



Attributes of MonitoringPoint

Name	HydroID	FType	Long_	Lat
SANDY RUN	6183	NWIS Water Quality	-77.558581	35.531550
SEVENMILE C	6251	NWIS Streamflow	-79.143903	36.065696
SEVENMILE C	6252	NWIS Streamflow	-79.160292	36.071806
SEVENMILE C	6253	NWIS Streamflow	-79.210016	36.077639

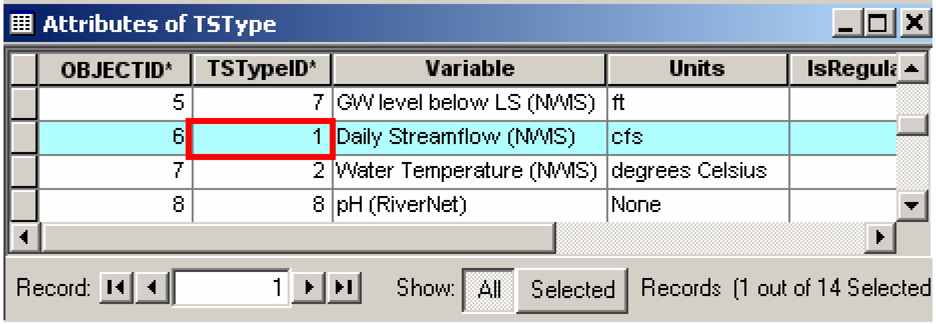
Record: 0 Show: All Selected Records (1 out of 712 Selecte



Attributes of TimeSeries

OBJECTID*	HydroCode	FeatureID*	TSTypeID*	TSValue	TSDateTin
1	02084909	6251	1	5.5	3/12/2001
2	02084909	6251	1	6.3	3/13/2001
3	02084909	6251	1	5.6	3/14/2001
4	02084909	6251	1	11	3/15/2001

Record: 1 Show: All Selected Records (0 out of *2000 Selec



Attributes of TSType

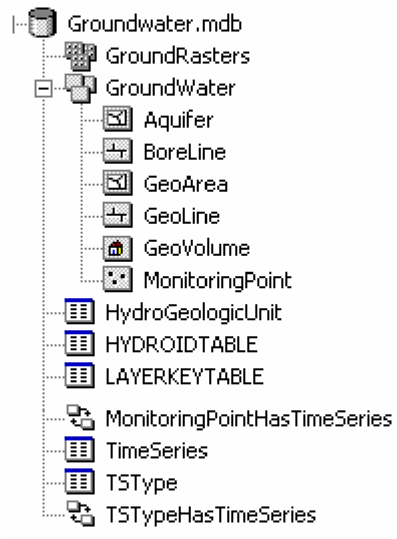
OBJECTID*	TSTypeID*	Variable	Units	IsRegula
5	7	GW level below LS (NWIS)	ft	
6	1	Daily Streamflow (NWIS)	cfs	
7	2	Water Temperature (NWIS)	degrees Celsius	
8	8	pH (RiverNet)	None	

Record: 1 Show: All Selected Records (1 out of 14 Selected

In addition to spatial-temporal relationships, there exist relationships for linking one spatial feature to the other as well. For example, HydroJunctionHasWatershed relates the HydroJunction points to Watershed polygons. This relationship is accomplished by matching the JunctionID of Watershed polygons to the HydroID of HydroJunction points. The same idea is used for relating HydroJunction and MonitoringPoint (HydroJunctionHasMonitoringPoint).

Description of Neuse Groundwater

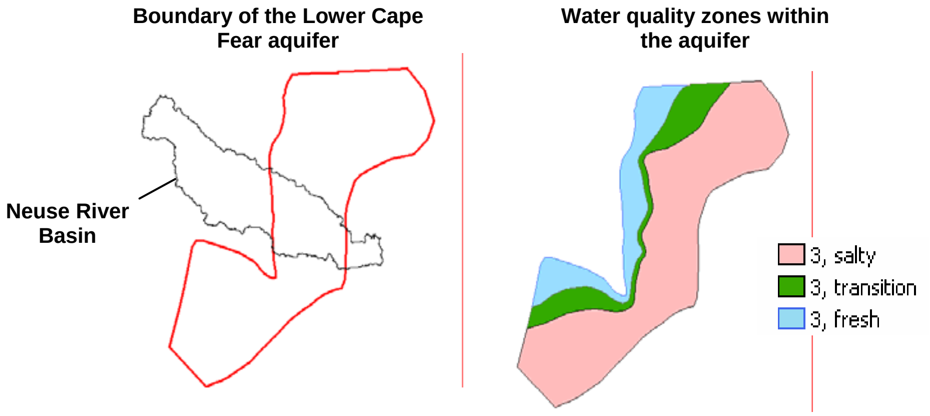
The Neuse Groundwater geodatabase follows the format of the Arc Hydro groundwater data model designed at the Center for Research in Water Resources. The geodatabase contains a description of the hydrogeology of the aquifer system which underlies the Neuse River Basin. The database contains one feature dataset (Groundwater) and one Raster Catalog (GroundRasters). In addition to spatial features, the geodatabase contains relational tables to store temporal information (Time Series), using the Arc Hydro time series table format.



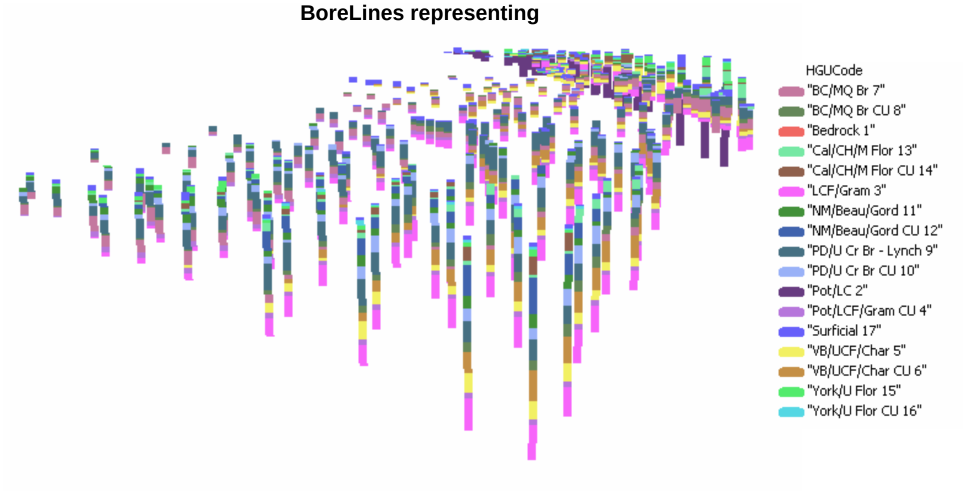
```
Groundwater.mdb
    GroundRasters
    GroundWater
        Aquifer
        BoreLine
        GeoArea
        GeoLine
        GeoVolume
        MonitoringPoint
    HydroGeologicUnit
    HYDROIDTABLE
    LAYERKEYTABLE
    MonitoringPointHasTimeSeries
    TimeSeries
    TSType
    TSTypeHasTimeSeries
```

Groundwater

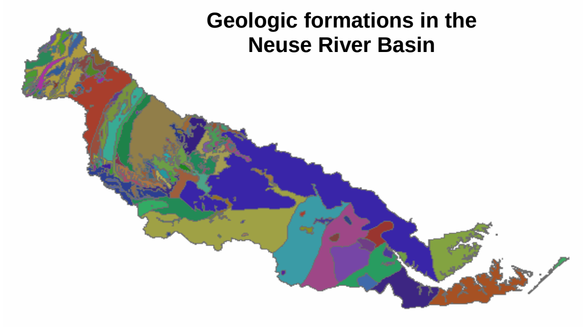
The Groundwater feature dataset contains information describing the hydrogeology of the aquifer system. The **Aquifer** feature class describes the boundary of aquifers within the study area, and water quality zones within the aquifer. The data were obtained from the Center for the Analysis and Prediction of River Basin Environmental Systems at Duke University (http://www.env.duke.edu/cares/neuse/GIS.html).

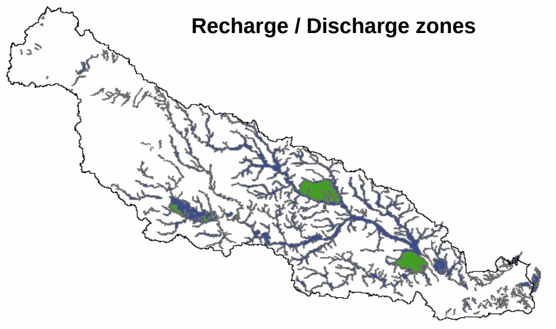

Boundary of the Lower Cape Fear aquifer
Water quality zones within the aquifer
Neuse River Basin
🟥 3, salty
🟩 3, transition
🟦 3, fresh

The **BoreLine** feature class contains 3D lines which represent the stratigraphy at boreholes. The hydrostratigraphy information is from a USGS model. Each feature in the BoreLine feature class is related to a stratigraphy well in the MonitoringPoint feature class. The 3D BoreLines can be viewed in ArcScene.

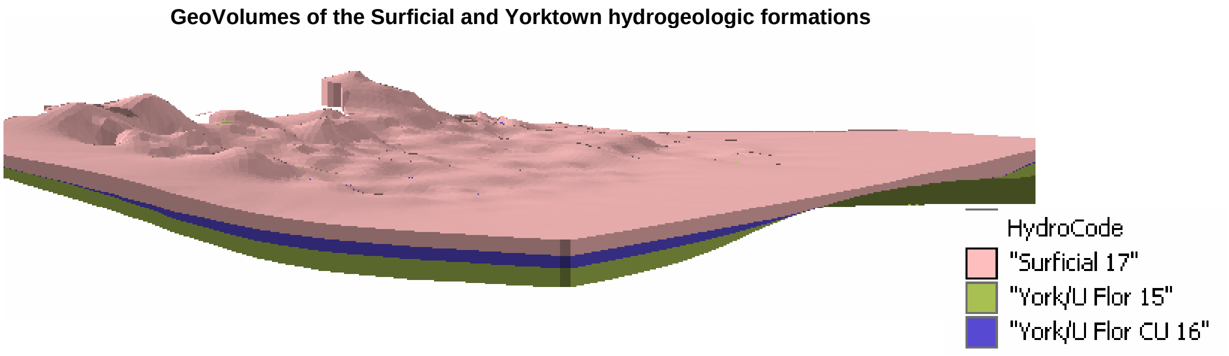

BoreLines representing

The **GeoArea** feature class describes geologic formations and recharge and discharge areas. The recharge and discharge areas are distinguished from geologic formations by a feature type (FType) subtype. The zones are also categorized by areas of recharge/discharge (value of 0 in the Elevation attribute) and areas of high ground elevation (values of 1 in the Elevation attribute). The information was obtained from the Center for the Analysis and Prediction of River Basin Environmental Systems (http://www.env.duke.edu/cares/neuse/GIS.html)

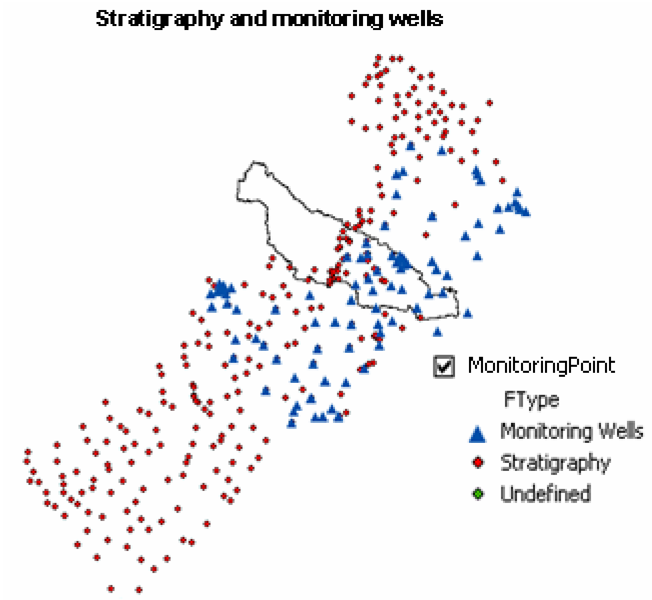

Geologic formations in the Neuse River Basin


Recharge / Discharge zones

GeoVolumes represent solid models which describe the hydrogeology of the subsurface. The volumes were created by the USGS in the Groundwater Modeling System (GMS) and were extracted from the GMS files into the geodatabase. In the geodatabase three solids are included which represent the top two aquifers (Surficial and Yorktown) and the confining unit between them. The solids can be viewed in ArcMap (2D) and ArcScene (3D).

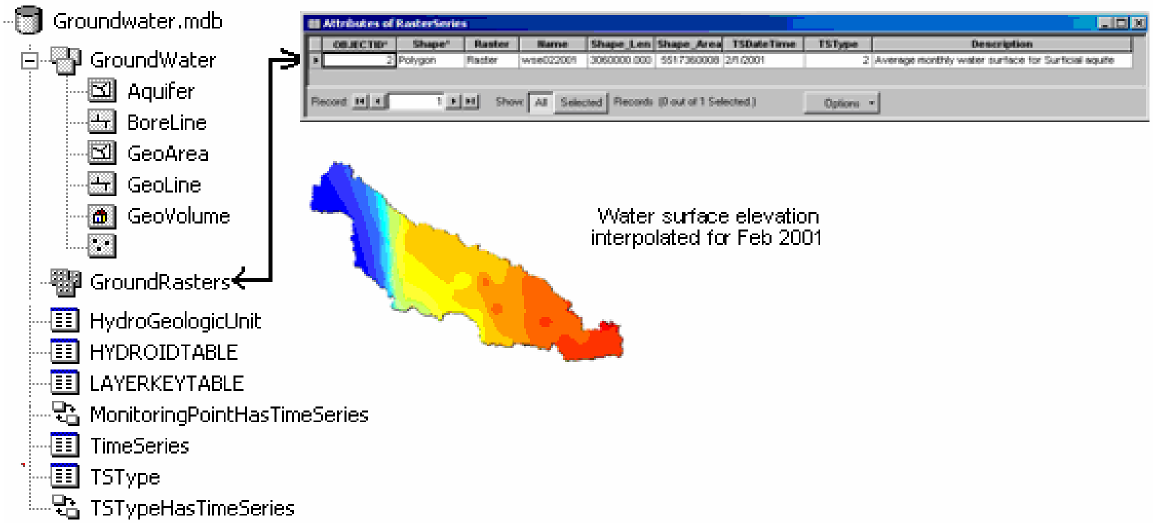

GeoVolumes of the Surficial and Yorktown hydrogeologic formations
HydroCode
"Surficial 17"
"York/U Flor 15"
"York/U Flor CU 16"

The **MonitoringPoint** feature class stores wells as two dimensional points. Two types of wells are stored in the MonitoringPoint feature class, stratigraphy and monitoring, which are distinguished by assigning a different feature type (FType) attribute. Stratigraphy wells were created from a the USGS GMS model and the monitoring wells were obtained from the North Carolina Division of Water Resources website http://www.ncwater.org/Data_and_Modeling/Ground_Water_Databases/wellaccess.php



GroundRasters

The **GroundRasters** raster catalog stores raster grids indexed by time. In this example dataset, a set of 11 raster grids represent the average monthly water table elevation of the Surficial aquifer for each month in 2001. It should be cautioned that in many cases, this interpolation was done from a sparse set of wells in some locations and likely needs to be refined or replaced by piezometric head surfaces computed from a groundwater flow model. The USGS is undertaking a time-varying groundwater flow model (Modflow) for the Coastal Plain aquifer which is expected to be completed in about two years.



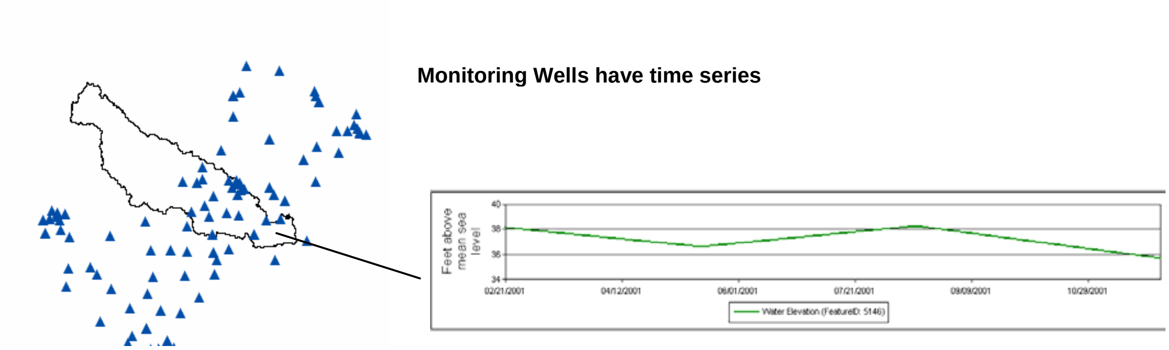
Attributes of RasterSeries

OBJECTID*	Shape*	Raster	Name	Shape_Len	Shape_Area	TSDateTime	TSType	Description
2	Polygon	Raster	wse022001	3060000.000	5517360008	2/1/2001	2	Average monthly water surface for Surficial aquife

Record 1 Show: All Selected Records (0 out of 1 Selected.) Options

TimeSeries

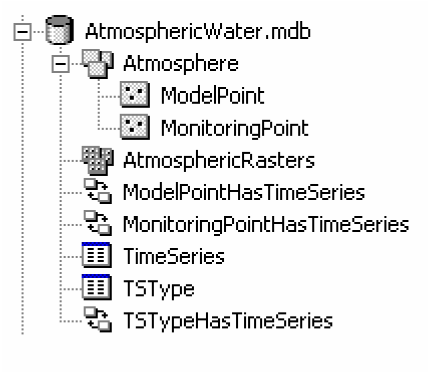
In addition to raster grids indexed by time, the temporal information is represented by time series records stored in the **TimeSeries** table. The time series table contains water elevations for the Surficial aquifer (feet above mean sea level) for the year 2001. Water elevation measurements are related to the monitoring wells in the MonitoringPoint feature class through HydroID. The time series were obtained from the North Carolina Division of Water Resources website http://www.ncwater.org/Data_and_Modeling/Ground_Water_Databases/wellaccess.php



Similar to the SurfaceWater geodatabase, the TimeSeries table is related to MonitoringPoint and TSType table through FeatureID and TSTypeID, respectively.

Neuse AtmosphericWater

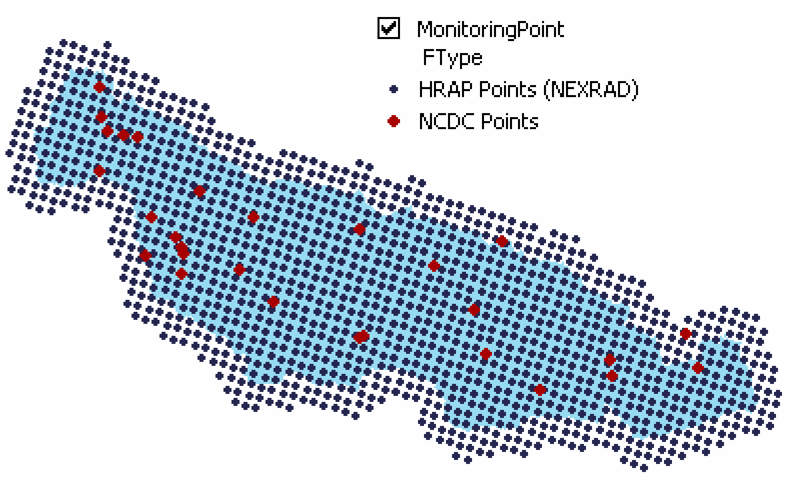
NeuseAtmosphericWater contains one feature dataset, one raster catalog and one time series table as shown below.



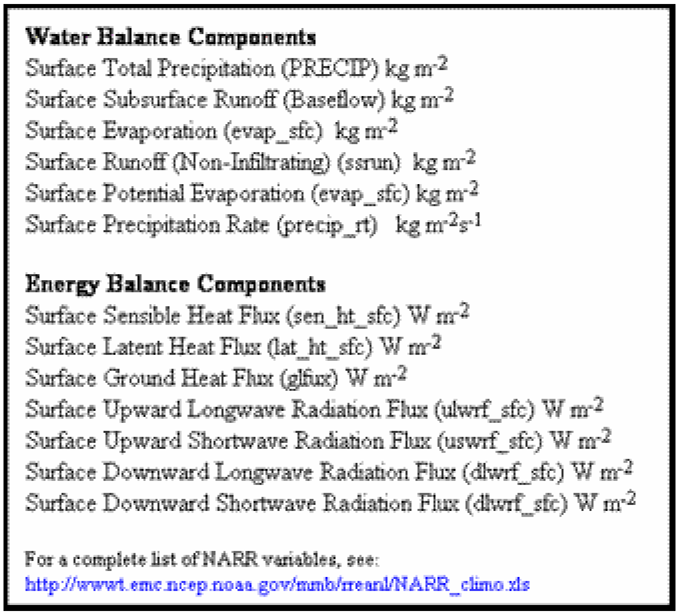
- AtmosphericWater.mdb
  - Atmosphere
    - ModelPoint
    - MonitoringPoint
  - AtmosphericRasters
  - ModelPointHasTimeSeries
  - MonitoringPointHasTimeSeries
  - TimeSeries
  - TSType
  - TSTypeHasTimeSeries

Atmosphere

The Atmosphere feature dataset contains two feature classes: MonitoringPoint and ModelPoint. The **MonitoringPoint** feature class can store the geospatial features marking the location of any atmospheric observation station. In the Neuse AtmosphericWater.mdb, the MonitoringPoint feature class contains NCDC rainfall measurement stations and HRAP points (centroids of HRAP polygons) surrounding the Neuse River Basin. Different types of features in the MonitoringPoint feature class are distinguished by assigning a different feature type (FType) attribute.



The **ModelPoint** feature class contains grid points for the North American Regional Reanalysis (NARR) model developed and maintained by the National Centers for Environmental Prediction (NCEP) http://wwwt.emc.ncep.noaa.gov/mmb/rreanl/ . Although NARR is a continental scale model (32-km grid cells), the output provides information on the energy and water fluxes important for closing the water and energy budgets. The list below gives the variables computed by NARR and imported into Atmospheric.mdb.

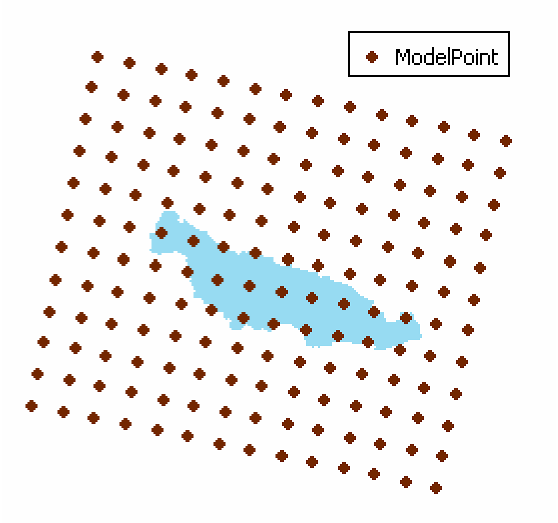


Water Balance Components
Surface Total Precipitation (PRECIP) kg m^{-2}
Surface Subsurface Runoff (Baseflow) kg m^{-2}
Surface Evaporation (evap_sfc) kg m^{-2}
Surface Runoff (Non-Infiltrating) (ssrun) kg m^{-2}
Surface Potential Evaporation (evap_sfc) kg m^{-2}
Surface Precipitation Rate (precip_rt) kg m^{-2}s^{-1}

Energy Balance Components
Surface Sensible Heat Flux (sen_ht_sfc) W m^{-2}
Surface Latent Heat Flux (lat_ht_sfc) W m^{-2}
Surface Ground Heat Flux (glfux) W m^{-2}
Surface Upward Longwave Radiation Flux (ulwrf_sfc) W m^{-2}
Surface Upward Shortwave Radiation Flux (uswrf_sfc) W m^{-2}
Surface Downward Longwave Radiation Flux (dlwrf_sfc) W m^{-2}
Surface Downward Shortwave Radiation Flux (dlwrf_sfc) W m^{-2}

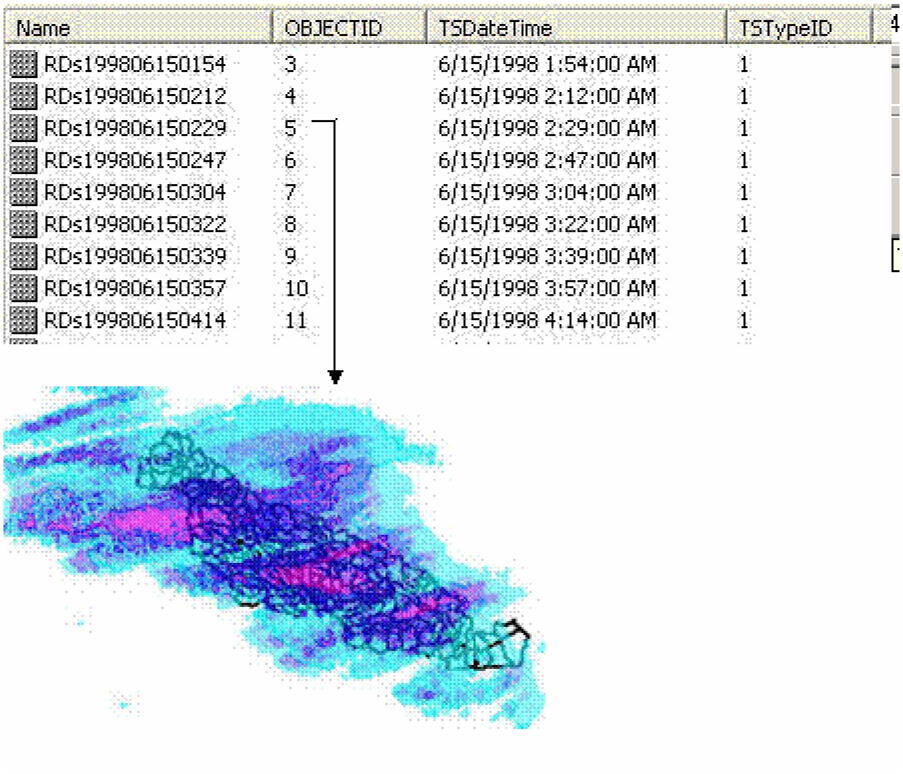
For a complete list of NARR variables, see:
http://wwwt.emc.ncep.noaa.gov/mmb/rreanl/NARR_climo.xls

NARR data is available on 3hr and monthly averaged time steps, but only the monthly averages were imported into the geodatabase.



AtmoshpericRasters

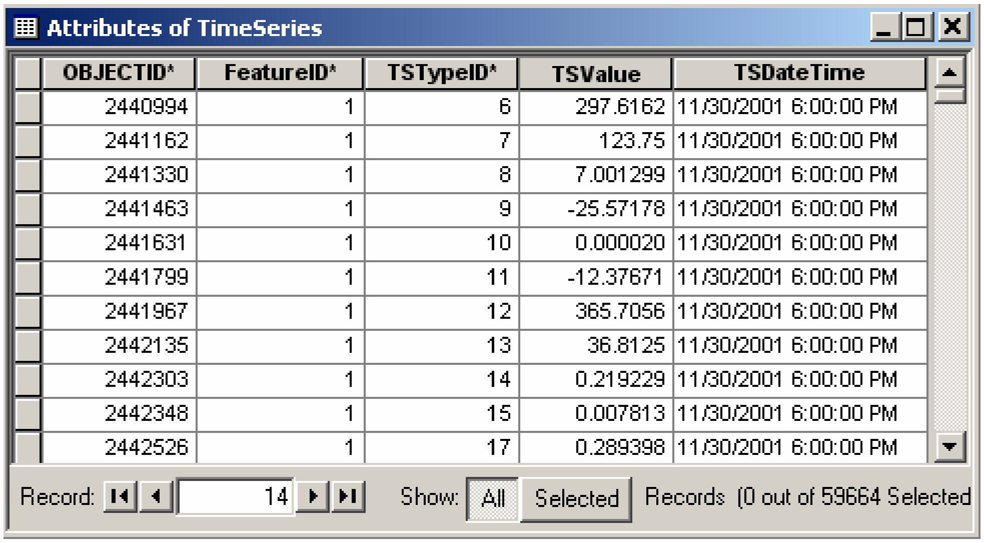
The **AtmosphericRasters** raster catalog contains NEXRAD rasters obtained from the NCDC Java viewer (http://www.ncdc.noaa.gov/oa/radar/jnx/). It can also be used to store rasters generated from interpolation of the NARR points or the NCDC rainfall gages. Each raster within the raster catalog is indexed by a time (TSDateTime) and a time series type (TSType). This structure (a raster indexed by time and a time series type) is called a RasterSeries.



Name	OBJECTID	TSDateTime	TSTypeID
▨ RDs199806150154	3	6/15/1998 1:54:00 AM	1
▨ RDs199806150212	4	6/15/1998 2:12:00 AM	1
▨ RDs199806150229	5	6/15/1998 2:29:00 AM	1
▨ RDs199806150247	6	6/15/1998 2:47:00 AM	1
▨ RDs199806150304	7	6/15/1998 3:04:00 AM	1
▨ RDs199806150322	8	6/15/1998 3:22:00 AM	1
▨ RDs199806150339	9	6/15/1998 3:39:00 AM	1
▨ RDs199806150357	10	6/15/1998 3:57:00 AM	1
▨ RDs199806150414	11	6/15/1998 4:14:00 AM	1

TimeSeries

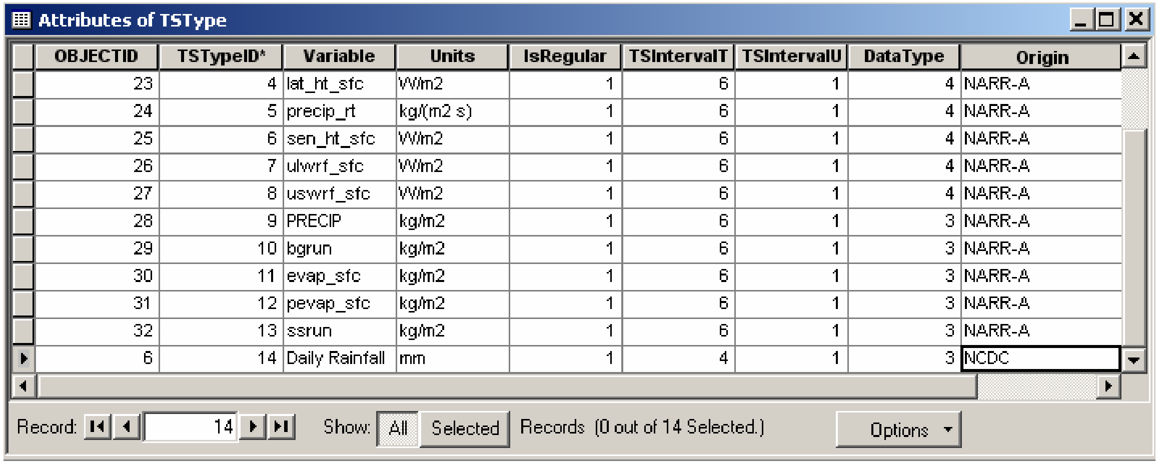
The **TimeSeries** table for Neuse AtmosphericWater, shown below, contains time series records for precipitation and the NARR variables.



Attributes of TimeSeries

OBJECTID*	FeatureID*	TSTypeID*	TSValue	TSDateTime
2440994	1	6	297.6162	11/30/2001 6:00:00 PM
2441162	1	7	123.75	11/30/2001 6:00:00 PM
2441330	1	8	7.001299	11/30/2001 6:00:00 PM
2441463	1	9	-25.57178	11/30/2001 6:00:00 PM
2441631	1	10	0.000020	11/30/2001 6:00:00 PM
2441799	1	11	-12.37671	11/30/2001 6:00:00 PM
2441967	1	12	365.7056	11/30/2001 6:00:00 PM
2442135	1	13	36.8125	11/30/2001 6:00:00 PM
2442303	1	14	0.219229	11/30/2001 6:00:00 PM
2442348	1	15	0.007813	11/30/2001 6:00:00 PM
2442526	1	17	0.289398	11/30/2001 6:00:00 PM

The details about each variable can be found in the TSType table.



OBJECTID	TSTypeID*	Variable	Units	IsRegular	TSIntervalT	TSIntervalU	DataType	Origin
23	4	lat_ht_sfc	W/m2	1	6	1	4	NARR-A
24	5	precip_rt	kg/(m2 s)	1	6	1	4	NARR-A
25	6	sen_ht_sfc	W/m2	1	6	1	4	NARR-A
26	7	ulwrf_sfc	W/m2	1	6	1	4	NARR-A
27	8	uswrf_sfc	W/m2	1	6	1	4	NARR-A
28	9	PRECIP	kg/m2	1	6	1	3	NARR-A
29	10	bgrun	kg/m2	1	6	1	3	NARR-A
30	11	evap_sfc	kg/m2	1	6	1	3	NARR-A
31	12	pevap_sfc	kg/m2	1	6	1	3	NARR-A
32	13	ssrun	kg/m2	1	6	1	3	NARR-A
6	14	Daily Rainfall	mm	1	4	1	3	NCDC

The time series records are related to geographic features (MonitoirngPoint and ModelPoint) through FeatureID and to TSType table through TSTypeID.

Summary

A brief description of different datasets used for creating a digital watershed for the Neuse basin is presented. The raw data for populating the digital watershed can come from several sources, and may vary from one study area to the other depending on project needs. The digital watershed concept is still under development, and the prototype that is presented in this document will evolve over time to describe the surface, subsurface and atmospheric components of hydrologic cycle in a manner such that each component can be linked to the other using a coupler table and relationships among different objects.

Reference

Reckhow, K., et al., (2004) Designing hydrologic observatories: a paper prototype of the Neuse watershed, CUAHSI Technical Report No. 6, Consortium of Universities for the Advancement of Hydrologic Science, Inc, 84 pp.December.

Chapter 9

Hydrologic Flux, Flow and Storage

By Jonathan Goodall
Nicholas School for the Environment
Duke University

David R. Maidment and Gil Strassberg
Center for Research in Water Resources
University of Texas at Austin

Abstract

Constructing a water, mass or energy balance of a hydrologic region requires accounting for the horizontal flow of water through the landscape in streams, rivers and aquifers, for the vertical fluxes of water between the atmosphere, land surface, soils and groundwater, and for changes in storage within any of these systems. A data model for representing time-varying fluxes, flows and storage in continuous and discrete spatial domains is presented. A hydrologic flux coupler is described which identifies the fluxes and flows that have to be considered when doing a water balance of a particular feature such as a watershed. The methodology is illustrated with an application for water balance computation on the Neuse River basin in North Carolina.

Introduction

CUAHSI is developing a program of hydrologic observatories for which a paper prototype study of the Neuse watershed has been completed to illustrate how such a hydrologic observatory could be designed (Reckhow et al., 2004). The Neuse Observatory study draws inspiration from the Water Science and Technology Board (2001) who state "What is needed for understanding water resources is a more holistic conceptual framework that encompasses regional scale hydrologic systems, land-atmosphere interactions, and the biogeochemical cycles that control contaminant transport".

According to Reckhow et al. (2004, p.1), "the measurement approach at hydrologic observatories will meet general requirements: (1) quantitative assessment of the fluxes and stores of water, sediment, and nutrients, (2**)** temporally and spatially integrated measurements of these fluxes and stores, and (3) acquisition of measurements in spatially stratified manner that allows for predictive understanding at the river basin scale". Reckhow et al. (2004, p.2) state further "four basic properties of a catchment repeatedly emerged as important. These properties are: (1) mass in each "store", (2) residence time within stores, (3) fluxes between stores, and (4) flowpaths among stores."

Since the CUAHSI Hydrologic Information System provides the information framework for hydrologic observatory data it is important to the success of the overall CUAHSI mission that that this framework should be supportive of the general considerations just stated concerning hydrologic observatory design. This paper takes the general requirements for hydrologic observatory design as a point of departure and proposes a space-time model for hydrologic fluxes, flows and storage on watersheds. An example application is presented for a monthly water balance for 2001 integrating atmospheric water, surface water and groundwater in the Neuse basin.

Hydrovolumes

Suppose we take the whole of the Neuse basin, as shown in Figure 1, and extrude the watershed boundary vertically upward into the atmosphere sufficient to encompass all atmospheric phenomena likely to be of interest on the basin, and similarly extrude the watershed boundary downwards into the geological strata sufficient to encompass all hydrogeologic phenomena affecting the hydrology of the basin. If the resulting vertical surfaces are enclosed by horizontal planes at the top and bottom, a volume in space representing the Neuse basin has been isolated from its surroundings, and can thus be subjected to analysis. In fluid mechanics, this is called a *control volume*. For the present purposes, it is termed a *hydrovolume*, defined as “a volume in space through which water, energy and mass flow, are stored internally, and transformed”.

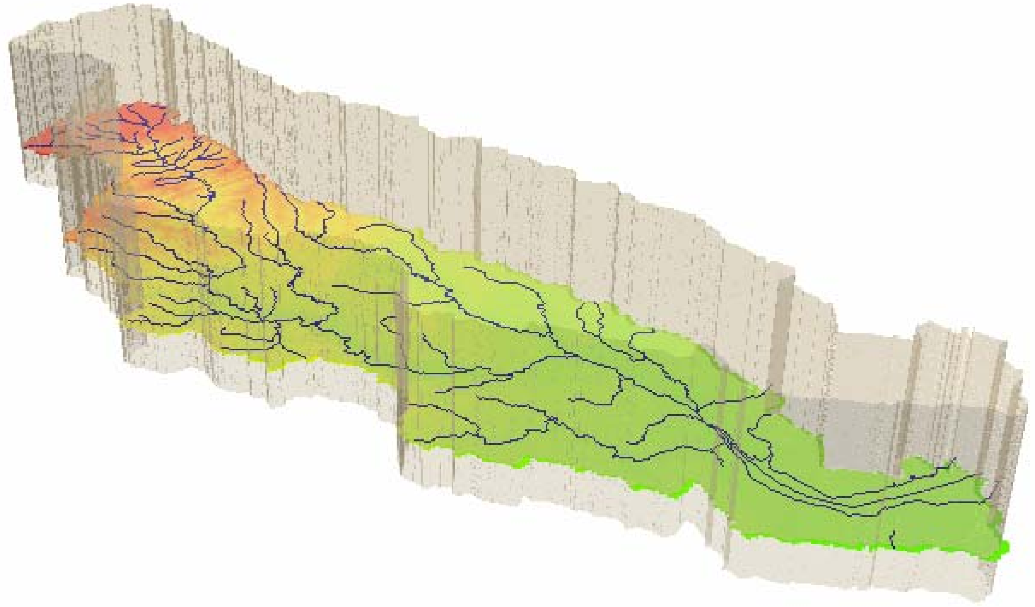


Figure 1. A hydrovolume of the Neuse basin

The exercise just performed could similarly be carried out at the watershed scale for any watershed within the Neuse basin. For purposes of illustration, the basin has been divided into 20 watersheds, by using selected USGS streamgaging station as outlet points, as shown in Figure 2. This arrangement is arbitrary and uses only a portion of the USGS gaging stations in the basin, sufficient to create a subdivision of the basin into reasonable number of watersheds of similar size.

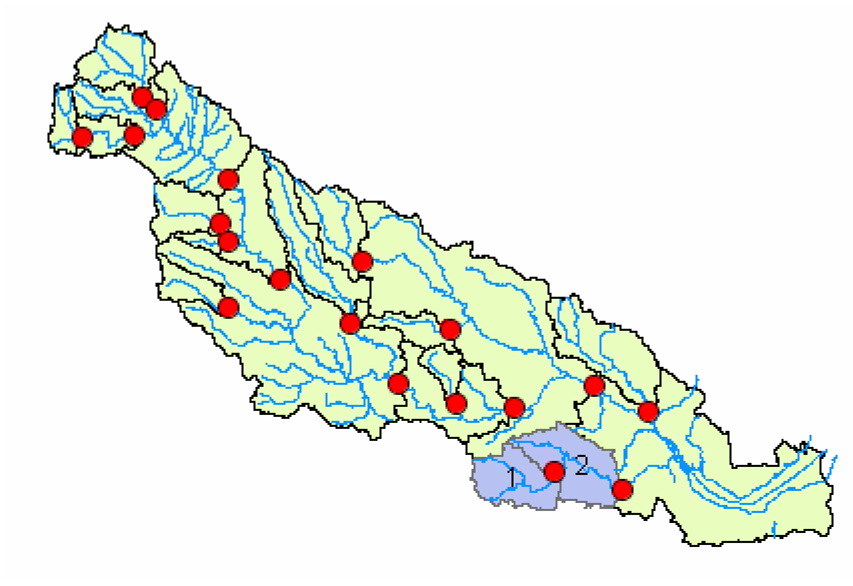


Figure 2. The Neuse basin divided into watersheds.

In particular, two watersheds are highlighted in Figure 2: (1) the watershed draining to the USGS gage 02092500 on the Trent River near Trenton NC; and (2), the watershed draining to the USGS gage 02092554 downstream on the Trent River at Pollocksville NC, as shown in Figure 3.

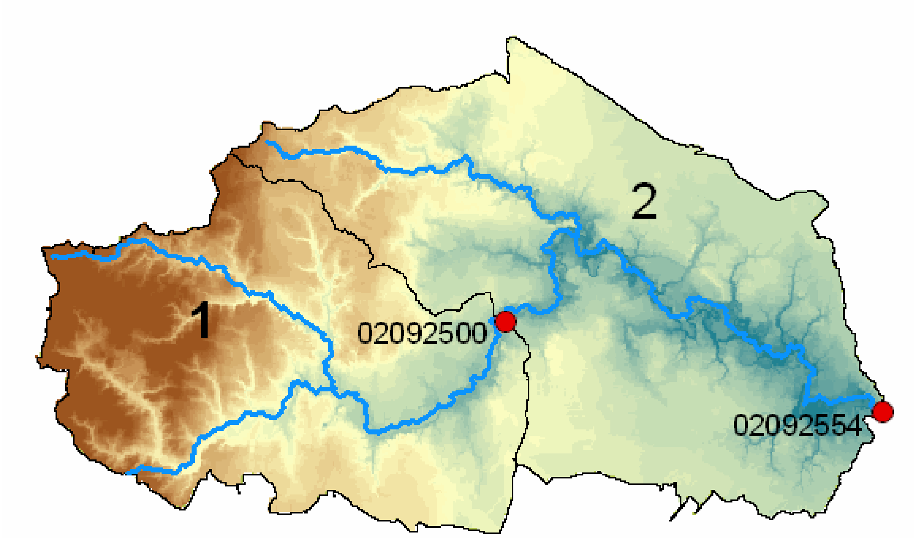


Figure 3. Two watersheds draining to USGS gaging stations in the Neuse basin

Watershed 1 has only one stream outlet and no stream inlets, while watershed 2 has one stream inlet and one outlet. Inspection of Figure 2 shows that there are other watersheds with as many as three stream inlets and one outlet. Suffice it to say, that if a set of stream gages is selected, the resulting drainage analysis subdivides the basin into a set of discrete watersheds, where all streamflow passing through the boundary between one watershed and another are measured at a gaging station.

If Hydrovolumes are drawn around these watersheds, the result is shown in Figure 4. Within a hydrovolume, one can define a *geovolume* as “the portion of a hydrovolume containing solid earth materials”.

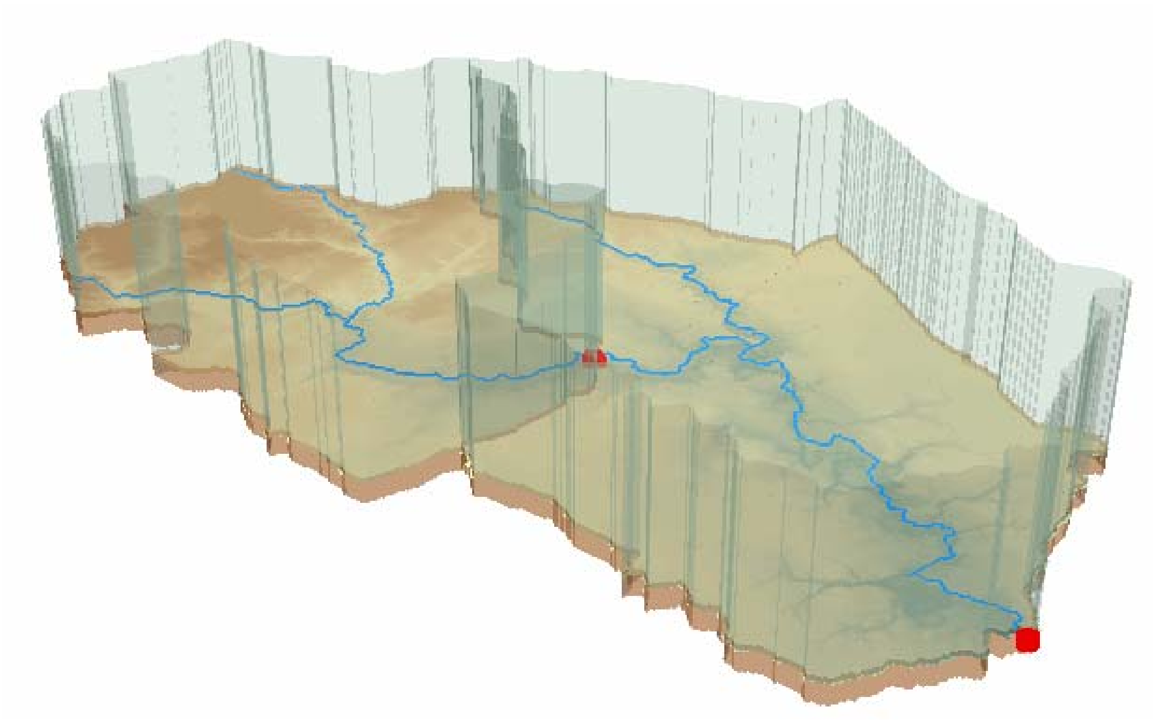


Figure 4. Geovolumes and hydrovolumes in the Neuse basin.

The process of spatial subdivision of the Neuse basin hydrovolume can be applied repeatedly to create smaller and smaller *watershed hydrovolumes*; each of these can be layered vertically to produce hydrovolumes representing atmospheric layers, soil layers, and hydrogeologic units; *channel hydrovolumes* representing stream and river reaches can be created as separate hydrovolumes within a watershed hydrovolume; *estuary hydrovolumes* can be differentiated from the streams draining into them, and so on. Any spatial subdivision of a hydrovolume is a hydrovolume.

Figure 5 shows a three-dimensional channel hydrovolume created for part of the Trent River using the River Channel Morphology Model developed by Merwade (2004), which mathematically relates the size and shape of the river channel cross-section to its planform sinuosity and hydraulic geometry parameters. The three-dimensional character of the channel is represented geospatially by a wire mesh of Crossections transverse to the flow and ProfileLines in the direction of flow. Hydraulic modeling can be used to estimate the flow velocity and depth throughout the length of this channel reach for a range of discharge values. A small hydrovolume for one channel segment is also shown in Figure 5.

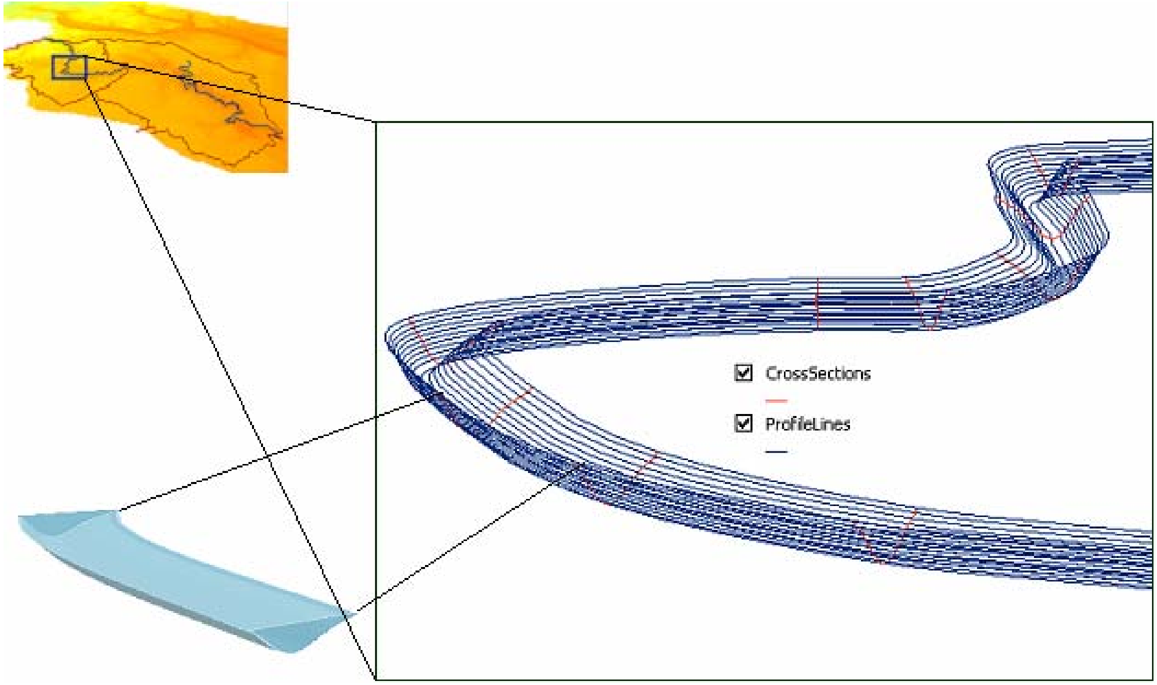


Figure 5. Three-dimensional channel hydrovolumes created for the Trent River.

The CUAHSI Hydrologic Information System is creating the tools needed to define hydrovolumes, and geovolumes as three dimensional geospatial features in a watershed system.

Flux, Flow and Storage

The CUAHSI conceptual model of a hydrologic observatory calls for "quantitative assessment of the fluxes and stores of water, sediment, and nutrients". This calls for some formal definition of the terms flux and store. In hydrology, the volumetric flow rate of water is usually symbolized by Q, measured for streams in cubic feet per second. If a surface is of area A, and the flow of water passing through that surface is Q, then the *flux* is the "flow per unit of surface area", or q = Q/A. For groundwater flow, the Darcy flux, q, is the conventional way of describing groundwater flow as the discharge rate per unit of cross-sectional area of porous medium.

If mass is considered instead of water volume, the mass flow rate is the amount of mass passing through a surface in a given interval of time, and the *mass flux* is the mass flow rate divided by the surface area. For example, the National Atmospheric Deposition Program quantifies the rate of deposition onto the land surface of chemicals in rainfall in units of kg/ha-year.

When considering land surface – atmospheric interactions, the fluxes of water and energy are intimately linked, so an *energy flux* can be defined as the rate at which energy passes through a surface, usually measured in Watts/m^2, where 1 Watt = 1 Joule/sec. For example, the average net radiation absorbed by the earth's surface over the globe and over the year is 105 W/m^2.

Strictly speaking, what has so far been defined as a flux is really an *area flux* since it is defined by flow per unit area. There are also *line fluxes* defined by flow per unit length, such as a channel loss rate in cfs/mile of stream channel. Line fluxes will not be considered further in this paper.

A *store* is a location where a quantity can be accumulated. For example, fish bioaccumlate mercury in their muscular tissues, so fish tissue can be referred to as a store for mercury. Within a water body, mercury can also be dissolved in the water column, can attach to colloidal particles in the water, can be contained in aquatic plants, and can be adsorbed onto bed sediments. Each of these is a store for mercury, so a hydrovolume containing a water body and its bed sediments could have many stores defined within it.

Suppose we define the term *storage* of a quantity (i.e. water, mass or energy) within a hydrovolume as the "total amount of that quantity contained in all stores within a hydrovolume". There is thus a fundamental distinction in terms of unit dimensions among flux, flow and storage, as shown in Table 1.

	Water	**Mass**	**Energy**
Flow	$[L^3/T]$	$[M/T]$	$[E/T]$
Flux	$[L/T]$	$[M/L^2T]$	$[E/L^2T]$
Storage	$[L^3]$	$[M]$	$[E]$

Table 1. Dimensions of flux, flow and storage for water, mass and energy.

Space and Time

The CUAHSI conceptual model of a hydrologic observatory also calls for "temporally and spatially integrated measurements of these fluxes and stores". This implies the existence of a space-time reference frame in three dimensions that is capable of describing fluxes and flows in a *continuous* spatial domain such as the atmosphere, or in a *discrete* spatial domain such as a river basin with its associated streams, rivers, water bodies, watersheds, soil and hydrogeologic units, and gaging stations.

Continuous Space-Time Domain

A *continuous* space-time domain has the characteristics:

- It is spatially extensive with properties that change smoothly throughout the spatial domain;
- It may vary in one-, two-, or three- space dimensions;
- Its properties change smoothly through time and they are defined at regular intervals within the time horizon.

For example, the North American Regional Reanalysis (NARR) of climate has produced a space-time grid of weather and climate variables on a 32km grid in space over North America and 3 hour time intervals from 1979 to 2003. These data were calculated at the National Centers for

Environmental Prediction by rerunning their Eta weather forecasting model in 3 hour time steps using as input the entire weather observation record from 1979 to 2003 http://wwwt.emc.ncep.noaa.gov/mmb/rreanl/.

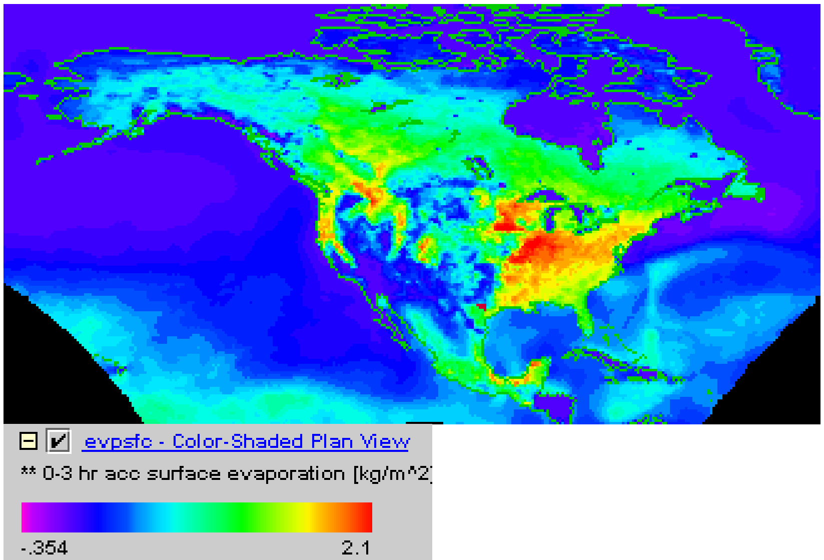

⊟ ☑ evpsfc - Color-Shaded Plan View
** 0-3 hr acc surface evaporation [kg/m^2]
-.354
2.1

Figure 6. Surface evaporation over North America from the North American Regional Reanalysis of climate visualized with Unidata's Integrated Data Viewer.

Figure 6 shows a map surface evaporation from the NARR for one 3-hour time interval visualized with Unidata's Integrated Data Viewer tool. The NARR also contains daily and monthly summaries of its variables, which include and land surface properties such as soil moisture levels, runoff, and subsurface recharge. Weather and climate information can be visualized using tools from Unidata, a data center supported by the NSF Geosciences Directorate (in much the same way as is CUAHSI), whose mission is to supply real-time atmospheric science information to academic institutions. This is a two-dimensional, time varying space-time field.

The Land Surface – Atmosphere model used in the Eta numerical weather prediction model is called NOAH, as shown in Figure 7. NOAH calculates for each grid cell and time step the values of dozens of flux and state variables, including precipitation, evaporation, potential evaporation, soil moisture level for several soil layers, surface runoff, and subsurface recharge to groundwater. These data are used in this paper as a representative climate model, but the same fluxes could be generated from a mesoscale climate model fitted just to a hydrologic observatory region, whose atmospheric boundary conditions are set by reference to the NARR data, just as the NARR is itself operating within a global numerical weather prediction model.

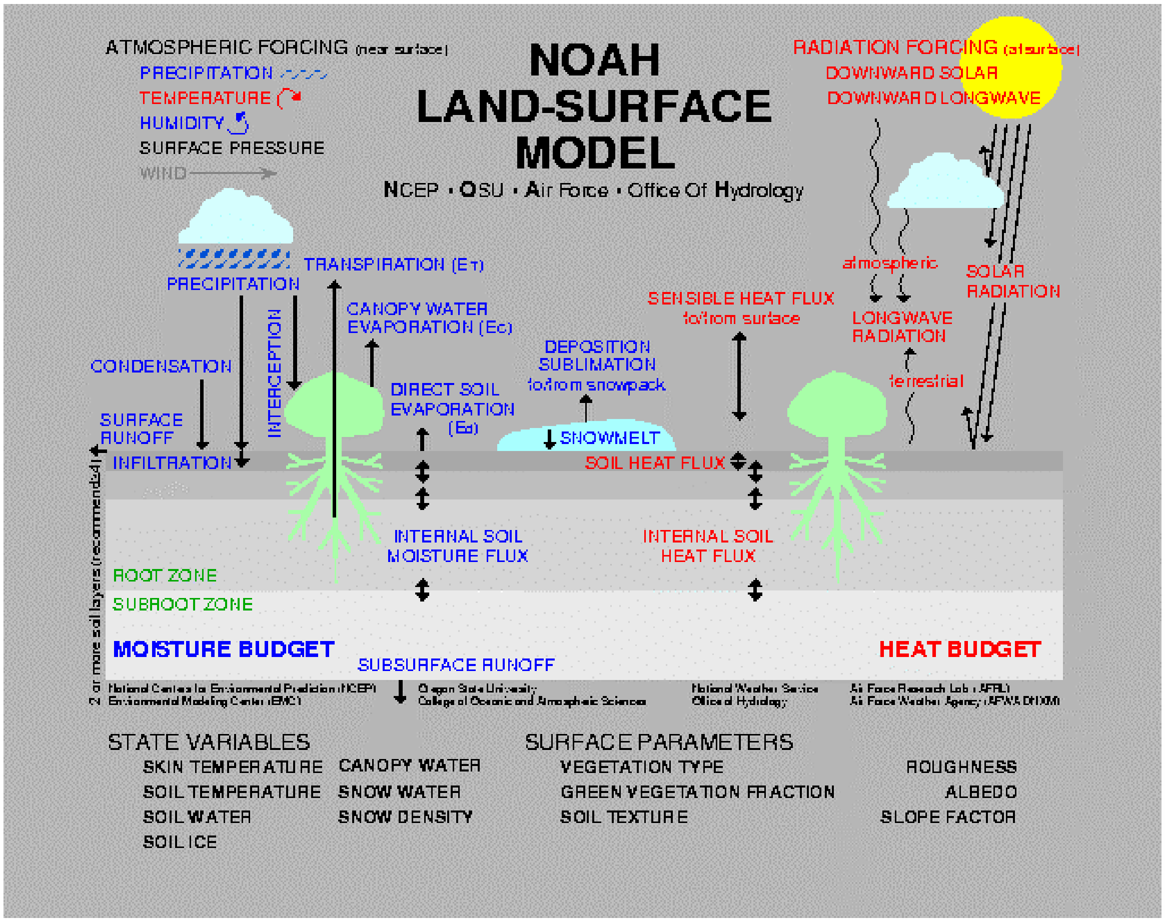
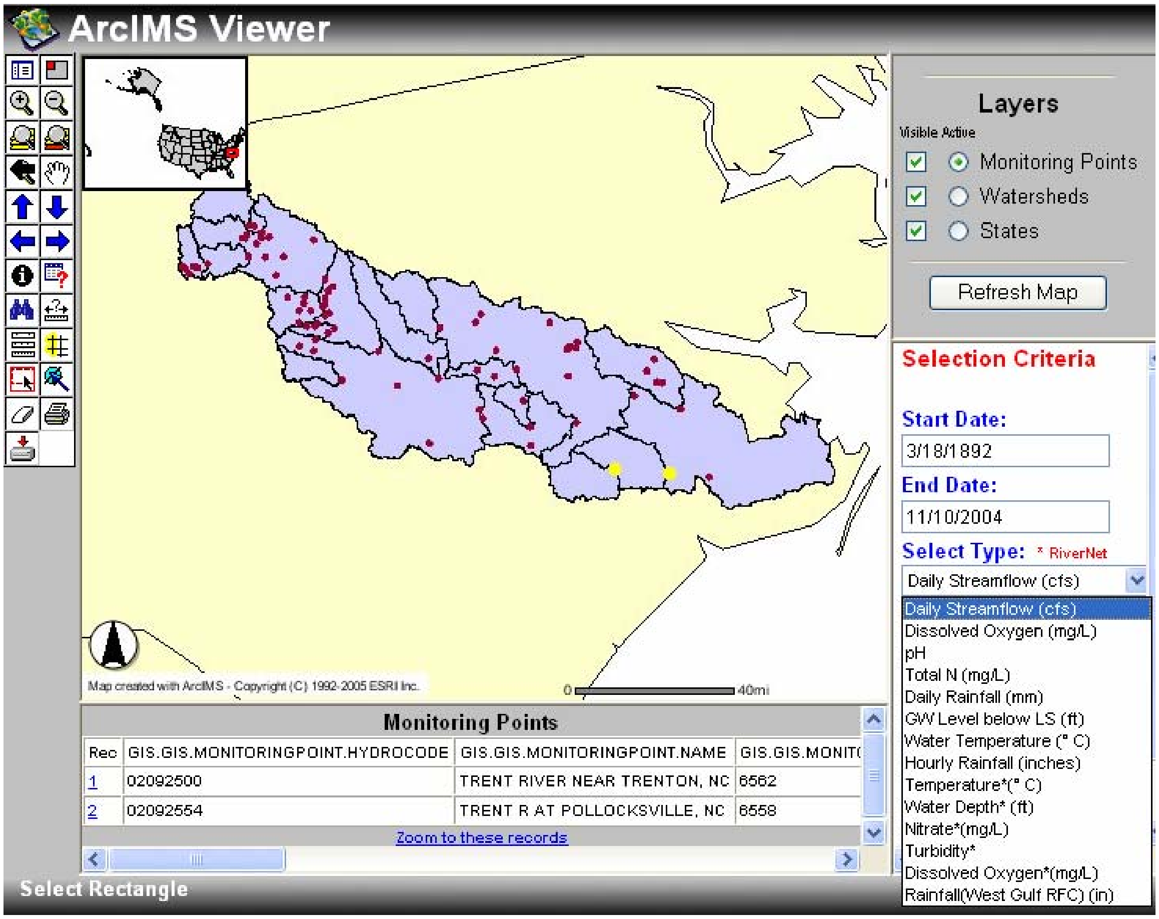

NOAH LAND-SURFACE MODEL
NCEP • OSU • Air Force • Office Of Hydrology
ATMOSPHERIC FORCING (near surface)
PRECIPITATION
TEMPERATURE
HUMIDITY
SURFACE PRESSURE
WIND
RADIATION FORCING (at surface)
DOWNWARD SOLAR
DOWNWARD LONGWAVE
TRANSPIRATION (Et)
PRECIPITATION
CANOPY WATER EVAPORATION (Ec)
INTERCEPTION
CONDENSATION
SURFACE RUNOFF
INFILTRATION
DIRECT SOIL EVAPORATION (Ed)
DEPOSITION SUBLIMATION to/from snowpack
SNOWMELT
SENSIBLE HEAT FLUX to/from surface
atmospheric
LONGWAVE RADIATION
terrestrial
SOLAR RADIATION
SOIL HEAT FLUX
INTERNAL SOIL MOISTURE FLUX
INTERNAL SOIL HEAT FLUX
ROOT ZONE
SUBROOT ZONE
MOISTURE BUDGET
SUBSURFACE RUNOFF
HEAT BUDGET
2 or more soil layers (recommended: 4)
National Centers for Environmental Prediction (NCEP) Environmental Modeling Center (EMC)
Oregon State University College of Oceanic and Atmospheric Sciences
National Weather Service Office of Hydrology
Air Force Research Lab (AFRL) Air Force Weather Agency (AFWA DNXM)
STATE VARIABLES
SKIN TEMPERATURE
SOIL TEMPERATURE
SOIL WATER
SOIL ICE
CANOPY WATER
SNOW WATER
SNOW DENSITY
SURFACE PARAMETERS
VEGETATION TYPE
GREEN VEGETATION FRACTION
SOIL TEXTURE
ROUGHNESS
ALBEDO
SLOPE FACTOR

Figure 7. The NOAH land – atmosphere model used in the North American Regional Reanalysis of climate.

Discrete Space-Time Domain

A *discrete* space-time domain has the characteristics:

- It may be represented in space by point, line, area or volume features;
- Its properties may be recorded regularly or intermittently in time;
- The domain has a boundary that represents the maximum extent in space and time of its representation.



Layers

Visible Active

☑ ● Monitoring Points
☑ ○ Watersheds
☑ ○ States

Selection Criteria

Start Date: 3/18/1892

End Date: 11/10/2004

Select Type: * RiverNet

Daily Streamflow (cfs)

Monitoring Points

Rec	GIS.GIS.MONITORINGPOINT.HYDROCODE	GIS.GIS.MONITORINGPOINT.NAME	GIS.GIS.MONIT(
1	02092500	TRENT RIVER NEAR TRENTON, NC	6562
2	02092554	TRENT R AT POLLOCKSVILLE, NC	6558

Zoom to these records

Figure 8. Hydrologic observation data presented on a discrete-space time domain for the Neuse basin.

For example, time series of hydrologic observations for the Neuse basin in North Carolina exist within the Neuse basin boundaries, each time series is linked to points in space where the observations were made, and the time range of the observations within the current Neuse Hydrologic Observations Database is from 1892 to 2004. Figure 8 shows an ArcIMS viewer developed for the CUAHSI Hydrologic Information System that permits downloading of data on streamflow, water quality, precipitation, temperature, and groundwater levels. The streamflow, precipitation and air temperature data are available regularly in time, while the water quality and groundwater level data are recorded at irregular points in time.

Data cube

Regardless of whether a hydrologic region is represented on a continuous or discrete space-time domain, data describing that region can be depicted using a *data cube*, whose axes represent the triplet {space, time, variables}. A particular observed data value, D, is located as a function of where it was observed, L, its time of observation, T, and what kind of variable it is, V, thus forming D(L, T, V), as shown in Figure 9.

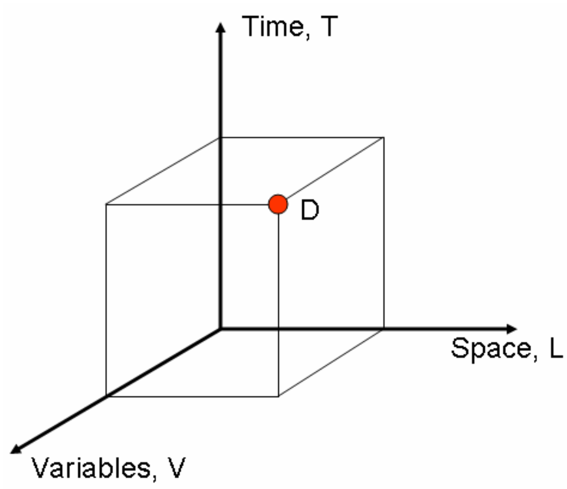


Figure 9. The data cube. A measured value D is indexed by its spatial location, L, its time of measurement, T, and what kind of variable it is, V.

NetCDF as a Continuous Space-Time Data Model

NetCDF is a data model developed at Unidata for the purpose of distributing atmospheric science data to academic institutions in the United States. The concept of netCDF is that it represents sampled values of an n-dimensional function space. Suppose we have a set of variables {X, Y} where the set {X} are independent variables whose values define *coordinate dimensions,* or points where information is represented, and the set {Y} are *variable dimensions* whose values are defined at those coordinate points. Typical examples of the set {X} are latitude, longitude, elevation, and time; typical examples of the set {Y} are temperature, humidity, wind speed, and water vapor pressure. In some cases where netCDF is used to represent atmospheric model information, the elevation dimension is replaced by pressure level, indicating the pressure level in the atmosphere of an atmospheric box whose conditions are being summarized.

When represented on the data cube, the coordinate dimensions {X} cover the space-time or L-T plane, and the variable dimensions {Y} are the variables on the V axis perpendicular to the L-T plane, as shown in Figure 10. A particular data value, D, might represent the relative humidity variable observed or calculated at a particular latitude, longitude, elevation and time.

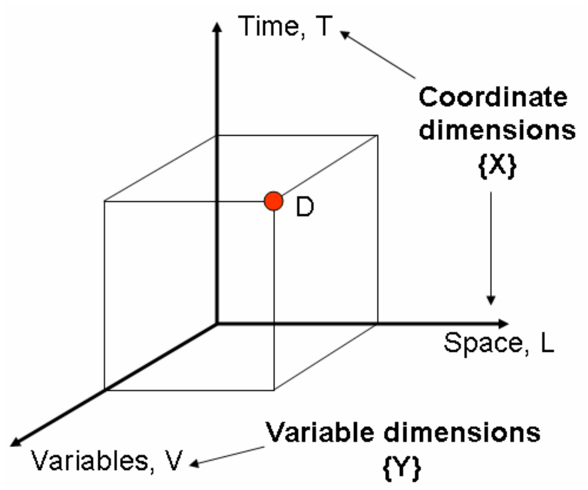


Figure 10. Representation of the data cube in netCDF.

Unidata has been in operation since 1983, and its netCDF format has proven to be widely popular in the atmospheric and ocean sciences for representing continuous fluid properties. It is also used by hydrodynamic modelers who want to record the results of their calculations on finite element or finite difference grids. NetCDF can also be used to track fluid properties along a flow path, such as when a balloon is released from the land surface and rises through the atmosphere to record atmospheric properties.

The merit of netCDF as a hydrology data model is that it can represent fluid fluxes and properties in continuous space-time domains, and it is a public domain format for which a significant body of application tools already exists. If CUAHSI were to support and use netCDF, it would supply a common data format for integration of hydrology with atmospheric and ocean sciences. The shortcomings of netCDF for hydrologic usage are that it does not describe discrete spatial domains such as watersheds, stream segments; it is not intended for describing a collection of time series on various time scales, such as those recorded by monitoring devices; and it is in a binary format that must be accessed through an application programming interface. There is an XML (eXtended Markup Language) version of netCDF which expresses netCDF data as text files in XML format that may be useful as a data exchange format for transforming netCDF files to other formats.

Arc Hydro Time Series as a Discrete Space-Time Data Model

Arc Hydro is a customization of the ArcGIS geographic information system for application in water resources, developed by a consortium of GIS developer and users (Maidment, 2002). ArcGIS is a built on a *geodatabase*, which is a relational database adapted for storing geographic objects.

In Arc Hydro, all points, lines, areas and volumes are *hydrofeatures*, that are described by a HydroID and a HydroCode. The *HydroID* is a unique long integer identifier assigned by Arc Hydro tools that is used for internal labeling and for building relationships between data tables in the geodatabase. The *HydroCode* is a text field that contains the permanent public identifier of a hydrofeature, if one exists. For example, Figure 3 shows three feature classes for the Neuse basin (Watersheds as areas, Streams as lines, and Gages as points) overlaid on a digital elevation model of the land surface terrain, which is an ArcGIS raster. The HydroCode for the gages is their USGS site number (e.g. 02092500), which identifies observational data stored within the National Water Information System at those gages. The Watersheds in Figure 3 have a numerical label 1 or 2, which has a completely different form than a USGS site number. The use of HydroID as a unique labeling system for all hydrofeatures avoids the confusion that results if each feature class is labeled in its own way.

In Arc Hydro, any hydrofeature can be related to any number of time series. The Arc Hydro time series data model as applied to hydrologic observations at monitoring points (as in Figure 8) is explained in some detail in a companion paper (Maidment, 2005) and that explanation will not be repeated here. The point relevant to the present discussion is that by using the Arc Hydro method, any point, line, area or volume feature can be related to any number of time series describing hydrologic fluxes, flows and storages that are associated with that feature. In Arc Hydro, the data value is called a TSValue, and the three axes of the data cube indexing that value are named FeatureID for space, TSTypeID for the variable type and TSDateTime for the time index, as shown in Figure 11. The FeatureID of the time series is equal to the HydroID of the feature it describes.

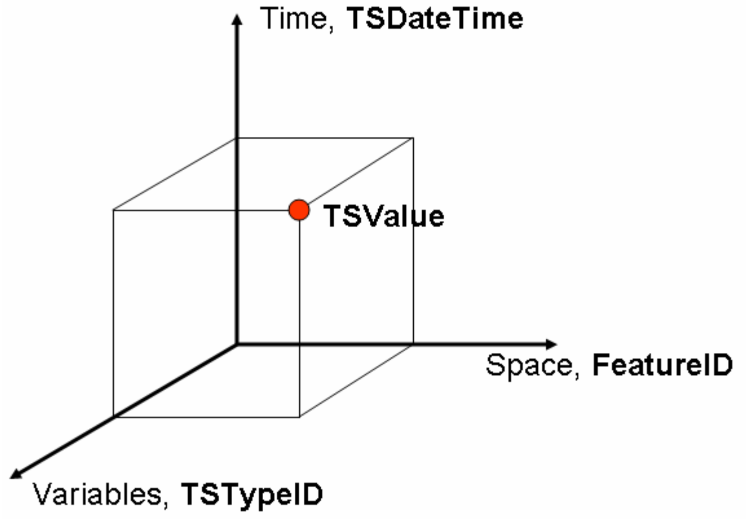


Figure 11. Representation of the data cube in Arc Hydro.

Because Arc Hydro time series are linked to the spatial feature they describe, they have associated with them a *shape*, which is the set of geographic coordinates defining how and where they are represented in space. Likewise, they have a *type*, which refers both to the nature of the variable they represent and also to the character of its representation through time. Thus, these can be thought of as *geospatial time series*, as illustrated in Figure 12.

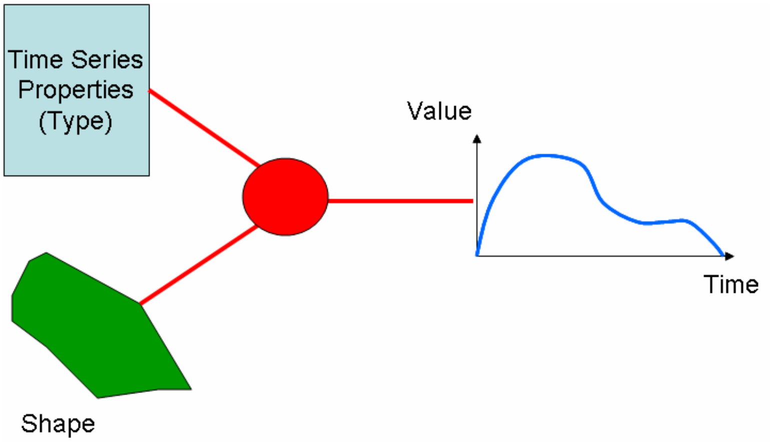


Figure 12. Geospatial time series

Although Arc Hydro time series were developed within the context of a commercial system, ArcGIS, it turns out that the time series part of Arc Hydro can be extracted from the GIS and implemented independently, as a delimited ascii .csv file, and in Excel. The CUAHSI HIS team has also shown that this time series model can be implemented in PostgreSQL, which is an open source, public domain relational database.

Linking Continuous and Discrete Space-Time Information

The continuous and discrete space-time data models just described live in quite different universes. NetCDF is a binary file format that was developed for operation on Linux and Unix operating systems. Arc Hydro time series are represented as tables in a relational database, and are normally used in the Access relational database which is part of Microsoft Office under the Windows operating system. How can data from these two systems be connected and merged?

ArcGIS allows for the inclusion of two-dimensional *rasters* or *grids*. An ArcGIS grid has square cells of a single fixed size, is defined on a rectangular domain, and describes a single variable in its cell values. A set of rasters can be stored in a *raster catalog*, and indexed by their date and time if they represent time varying information, to form a *raster series*. Rasters can be laid over spatial features such as watersheds and the average value of the raster within the boundary of each feature calculated. Two-dimensional fields in netCDF format, like the surface evaporation fields from the NARR shown in Figure 5, can be converted to geospatial time series linked to points located at the center of each NARR cell. Within ArcGIS, these geospatial time series can be spatially interpolated into a raster series, laid over the watersheds and the corresponding geospatial time series of watershed properties can be determined. This is how energy and water fluxes from the NARR were acquired for the water balance on the Neuse basin described later in this paper. NetCDF is presently being incorporated into ArcGIS as a native data format, and tools to read, write, display and operate on netCDF files will exist in the next release of ArcGIS due out at the end of 2005 or early in 2006.

Computing a Water Balance

The drainage area of a river basin can be divided into watershed hydrovolumes by delineating the watershed that flows to each streamgage in the basin, as shown in Figures 2 and 3. Upstream watersheds are bounded entirely by drainage divides and have only the gage at their outlet by which they communicate with adjacent modeling units. Downstream watersheds may have two or more gages on their boundary transmitting flow into and out of the modeling unit, as shown in Figure 3. For any one of these units, a simple water balance can be written as:

$$\frac{dS}{dt} = Q_{in} - Q_{out} + (P - E - R)A \qquad (1)$$

Where S is the storage of water in the watershed, Q_{in} is the flow coming into the unit as measured by gages on its upstream boundary, Q_{out} is the outflow at the gage at the downstream end, P is precipitation, E is evaporation, R is groundwater recharge, and A is the area of the watershed. This water balance links *vertical fluxes* of water between the atmosphere, land surface and subsurface with *horizontal flows* of water through the stream channel system.

Equation (1) is a straightforward equation but its automated application within a Hydrologic Information System is not simple. First, it requires knowledge of the spatial distribution of precipitation and evaporation over the drainage area in order to be able to get the appropriate values for the watershed as a whole; second, the dimensions of the inflow and outflow data, usually cfs, are inconsistent with typical units for P, E and Rof in/day or mm/day, respectively; third, the drainage area is needed, and that involves yet another set of units, say km^2 or $miles^2$. Q_{in} and Q_{out} are flows associated with stream gages represented as point hydrologic features in the landscape, while P, E and R are fluxes associated with the watershed as an areal hydrologic feature, so the multiplication by the watershed area is necessary to make the computations dimensionally consistent. The rate of storage change, dS/dt, and its integral through time, cumulative storage, S, are time series associated with the watershed as an areal hydrologic feature. In general, all of these hydrologic fluxes, flows and storages can be represented as geospatial time series, that is time series which have associated with them some point, line, area or volume feature, as shown in Figure 10.

Equation (1) can be partitioned by considering the term ($Q_{in} - Q_{out}$) as the *net inflow* of water to the watershed hydrovolume through the channel system. Similarly (P – E – R) is the *net influx* of water from the atmosphere to the land surface, and if this flux is multiplied by the watershed area, A, and the appropriate unit conversions done, the result (P – E – R)A, is the net inflow of water from the atmosphere to the land surface. Thus, Equation (1) can be rewritten as

$$\frac{dS}{dt} = \sum (NetInflows) \qquad (2)$$

Now, suppose this water balance does not close because of data uncertainties and errors, the degree of non-closure of the water balance can be estimated by a *residual flow,* Q_r, defined as

$$Q_r = \frac{dS}{dt} - \sum (NetInflows) \qquad (3)$$

Of course, this requires some means of independently estimating dS/dt, such as by converting changes in water surface elevation in a reservoir to changes in storage, or by mapping piezometric head fields in a groundwater system through time and looking at their time variation.

The computation of heat energy balances can be done in a similar fashion to that just described for a water balance with the additional degree of complication that energy comes in many forms (short wave and long wave radiation, sensible heat flux, latent heat flux, ground heat flux). The computation of mass balance for chemical and biological constituents of water is significantly more complex again because these constituents can exist in many stores, and they can also be created or destroyed within a hydrovolume. However, the basic principles are still the same.

Reckhow et al. (2004, p.42) describe a model for mass balance of Radon in the Neuse estuary which includes: "(1) benthic advective-diffusive exchange; (2) in situ production and loss; (3) horizontal water column advection; (4) air sea – exchange", whose result is the time variation of Radon concentration in the estuary. If this is multiplied by the volume of water in the estuary it gives the time variation of the mass of Radon stored there. They state "this approach assesses all flux terms and estimates the groundwater contribution by difference", in other words, the unknown residual flow of radon in groundwater to the Neuse estuary is estimated as a term Q_r in Equation (3) where the Net Inflow terms have also to include the difference between radon production and loss within the estuary waters.

Although the water balancing methodology described here does not address these additional layers of complexity to be confronted in mass balancing of chemical or biological constituents, it does establish a space-time context for constructing such mass balances that might be elaborated by additional developments later.

Hydrologic Flux Coupler

The *hydrologic flux coupler* is an application developed in ArcGIS to allow water, energy or mass balances to be constructed for hydrovolumes defined in atmospheric, surface and subsurface water. Each hydrovolume is treated as a separate discrete space entity with its own fluxes and flows, linked to the hydrovolume using a *coupling table*. It should be understood that it is not the purpose of the flux coupler to define the magnitudes of all the fluxes and flows, but rather to take estimates for fluxes and flows developed by measurements and modeling, and bring these together to compute a water, energy or mass balance in a hydrologic landscape divided into hydrovolumes.

Example Application to the Neuse basin

An example application of the hydrologic flux coupler to the Neuse Basin is now presented. For purposes of explanation, this example is confined to the two watershed hydrovolumes depicted in Figure 4. The purpose is to compute a monthly surface water balance for the year 2001 using as inputs USGS streamflow data at the gages, and vertical fluxes for precipitation, evaporation and groundwater recharge obtained from the North American Regional Reanalysis of climate (NARR). The example is carried out in a series of steps.

Step 1: Establish and label the hydrovolumes

A terrain analysis is conducted using a digital elevation model of the land surface, and the locations of stream gages on the river network. In this case, the Arc Hydro toolset is used to delineate the watershed draining to each gage, but that task could have been undertaken with other toolsets also. Two feature classes are used in this example, watersheds and gages. Each of the watersheds and gages is uniquely labeled with its HydroID using the Arc Hydro tool Assign HydroID. In this case, the two watersheds have HydroID's of 9623 and 9614, respectively, and the gages at their outlets have HydroID's of 9748 and 9749. HydroID's are arbitrary integers that may not have any inherent meaning – their only purpose is to serve as identifying numbers.

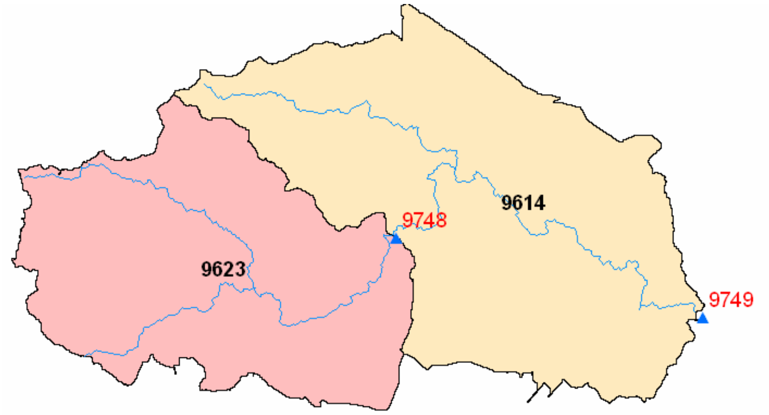


Figure 13. Watershed hydrovolumes and related streamgage features

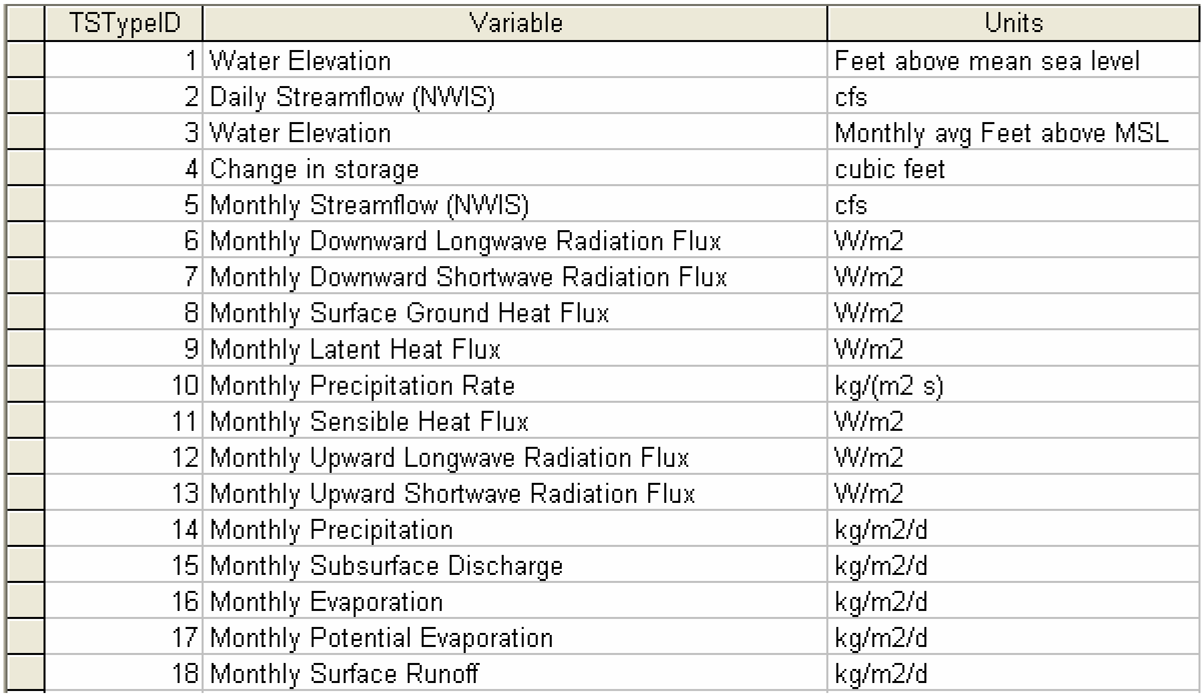
Step 2. Assemble the flux and flow data

North American Regional Reanalysis (NARR): The NARR contains the full water and energy balance for North America over a 25 year period (1979-2003). It is computed using the most current regional weather prediction model (ETA) and assimilated weather observations from the 25 year period. The grid spacing of the model is 32 km and the temporal scale is 3hours. The model results are in GRIB binary format at the NARR website http://www.emc.ncep.noaa.gov/mmb/rreanl/ and were converted to netCDF format for this example using data converters supplied by Unidata. Some data manipulation was required to convert the resulting files into monthly data calculated over each watershed, as discussed earlier in the paper.

USGS streamflow measurements: Streamflow measurements were downloaded from the USGS NWIS website (http://waterdata.usgs.gov/nwis/) – the measurements come as daily average streamflow in cubic feet per second, and these were averaged for each month to give mean monthly discharges. The hydrologic flux coupler has tools to automatically accomplish this time up-scaling.

3. Convert the fluxes and flows to Arc Hydro format

This involves reading the time series in an MS Access database to form time series tables. In this example, a total of 18 different geospatial time series were created as shown in Figure 14. Not all of these are needed to do a water balance – in fact, a complete set of energy fluxes is also in the table so that an energy balance for each watershed could also be calculated if desired.



TSTypeID	Variable	Units
1	Water Elevation	Feet above mean sea level
2	Daily Streamflow (NWIS)	cfs
3	Water Elevation	Monthly avg Feet above MSL
4	Change in storage	cubic feet
5	Monthly Streamflow (NWIS)	cfs
6	Monthly Downward Longwave Radiation Flux	W/m2
7	Monthly Downward Shortwave Radiation Flux	W/m2
8	Monthly Surface Ground Heat Flux	W/m2
9	Monthly Latent Heat Flux	W/m2
10	Monthly Precipitation Rate	kg/(m2 s)
11	Monthly Sensible Heat Flux	W/m2
12	Monthly Upward Longwave Radiation Flux	W/m2
13	Monthly Upward Shortwave Radiation Flux	W/m2
14	Monthly Precipitation	kg/m2/d
15	Monthly Subsurface Discharge	kg/m2/d
16	Monthly Evaporation	kg/m2/d
17	Monthly Potential Evaporation	kg/m2/d
18	Monthly Surface Runoff	kg/m2/d

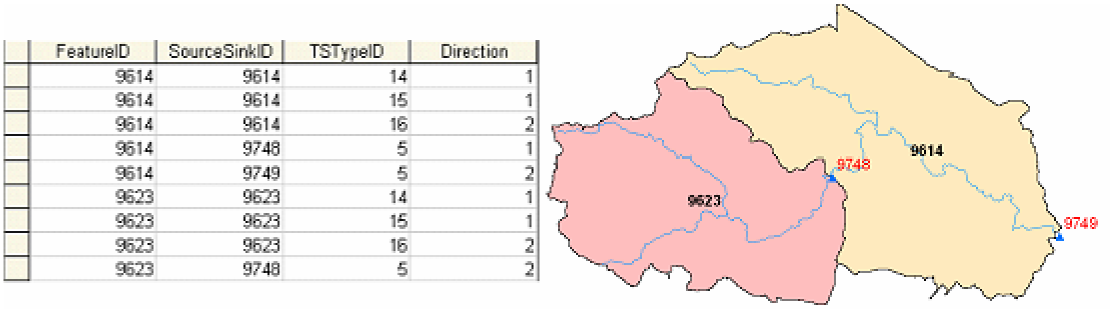
Figure 14. Geospatial time series types created for the water balance application.

Step 4. Establish the flux coupling table

The flux coupling table is shown in Figure 15. The FeatureID is the HydroID of the Hydrovolume being studied, the SourceSinkID is the HydroID of a feature that has time series on it needed for the computation, the TSTypeID specifies the type of time series, and the Direction specifies whether IN (Direction = 1) or OUT (Direction = 2).

Watershed hydrovolume 9623 has three vertical fluxes in kg/m^2-day associated with it: Types 14 (Precipitation), 15 (Subsurface discharge), 16 (Evaporation) – these are areally associated with the watershed itself (i.e. the SourceSinkID and the FeatureID are the same) to get the watershed area in order to convert these fluxes to flows. This hydrovolume also has one flow of Type 5 (Monthly streamflow in cfs) associated with it, whose SourceSinkID is 9748 which means that the flow series are attached to the stream gage at the watershed outlet.

A similar set of coupling records exists for the downstream watershed hydrovolume 9614 – except that it has two streamgage flows at SourceSinkID's of 9748 and 9749, respectively. Notice how the Direction for gage 9748 is 2 (OUT) when it is associated with hydrovolume 9623 and 1 (IN) when it is associated with hydrovolume 9614. The same flow can is used as an outflow for one hydrovolume and an inflow to the next.

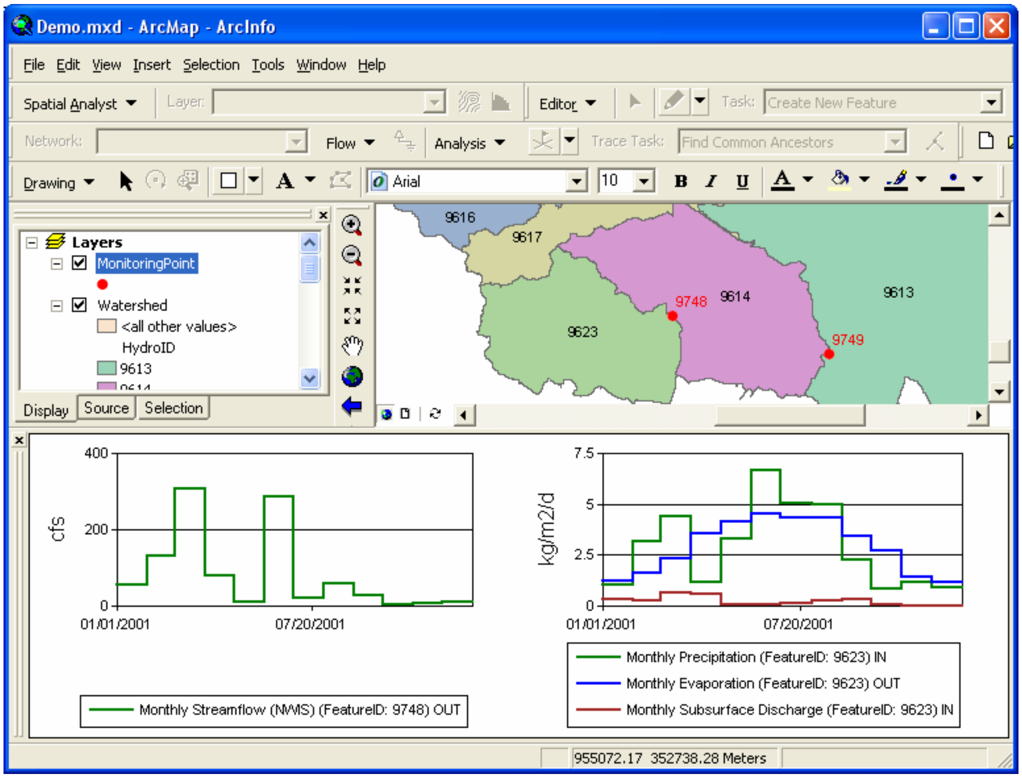


FeatureID	SourceSinkID	TSTypeID	Direction
9614	9614	14	1
9614	9614	15	1
9614	9614	16	2
9614	9748	5	1
9614	9749	5	2
9623	9623	14	1
9623	9623	15	1
9623	9623	16	2
9623	9748	5	2

Figure 15. The hydrologic flux coupling table and its related features

5. Plot the fluxes and flows

The hydrologic flux coupler creates plots in Arc Map of geospatial time series. Figure 16 shows two plots, the one on the left being the average monthly streamflow in cfs, and the one on the right the various flux components. Graphs can be “dragged” from one chart space to the other and their units and dimensions will automatically be converted to be conformal with the target chart space. The flux and flow data can be exported to another application by right clicking on the graph and selecting the desired option, such as .txt, or Excel.



Demo.mxd - ArcMap - ArcInfo
File Edit View Insert Selection Tools Window Help
Spatial Analyst
Layer:
Editor
Task: Create New Feature
Network:
Flow
Analysis
Trace Task: Find Common Ancestors
Drawing
Arial
10
B I U
Layers
MonitoringPoint
Watershed
<all other values>
HydroID
9613
9614
Display Source Selection
9616
9617
9748
9614
9613
9623
9749
400
200
0
cfs
01/01/2001
07/20/2001
Monthly Streamflow (NWIS) (FeatureID: 9748) OUT
7.5
5
2.5
0
kg/m2/d
01/01/2001
07/20/2001
Monthly Precipitation (FeatureID: 9623) IN
Monthly Evaporation (FeatureID: 9623) OUT
Monthly Subsurface Discharge (FeatureID: 9623) IN
955072.17 352738.28 Meters

Figure 16. Flows and fluxes for hydrovolume 9623

6. Calculate the Net Inflow and Net Influx

For each hydrovolume, the net inflow is calculated as the difference between the streamflow into and out of the hydrovolume ($Q_{in} – Q_{out}$) in cfs, and the net influx is computed as P – E – R, shown in Figure 17 now in units of in/hr (any one of a number of possible alternative units could have been chosen and the conversion to that unit system is made automatically).

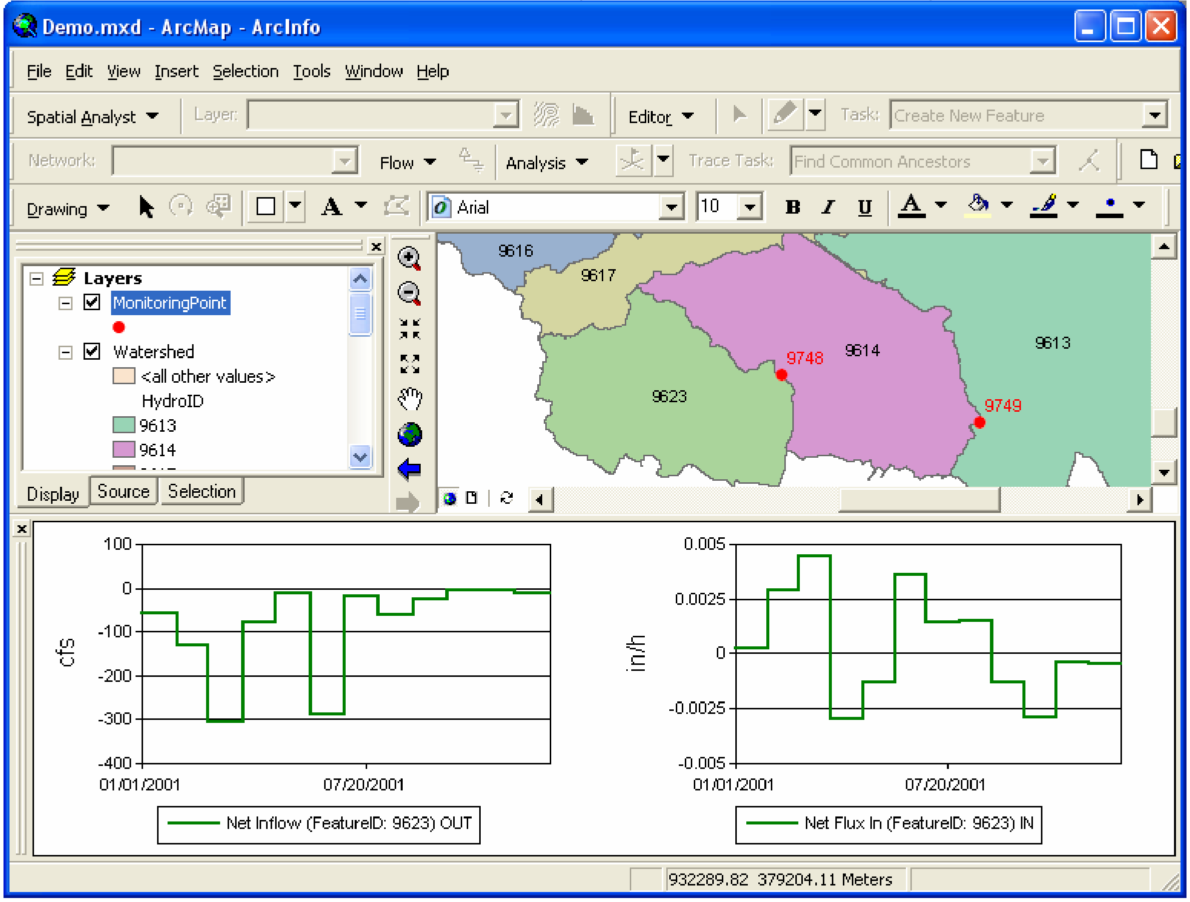


Figure 17. The net inflow and net influx of water to hydrovolume 9623.

7. Calculate the total net inflow and integrate storage through time

The total net inflow is calculated as ($Q_{in} – Q_{out}$) + (P – E – R)A where A is the watershed area – this yields a continuous change in storage in cfs, according to Equation (2) and this can be integrated through time to plot a profile of the storage of water on the watershed as shown in Figure 18. It is apparent that the water balance does not close as the storage shows a persistent downward trend through the year, 2001.

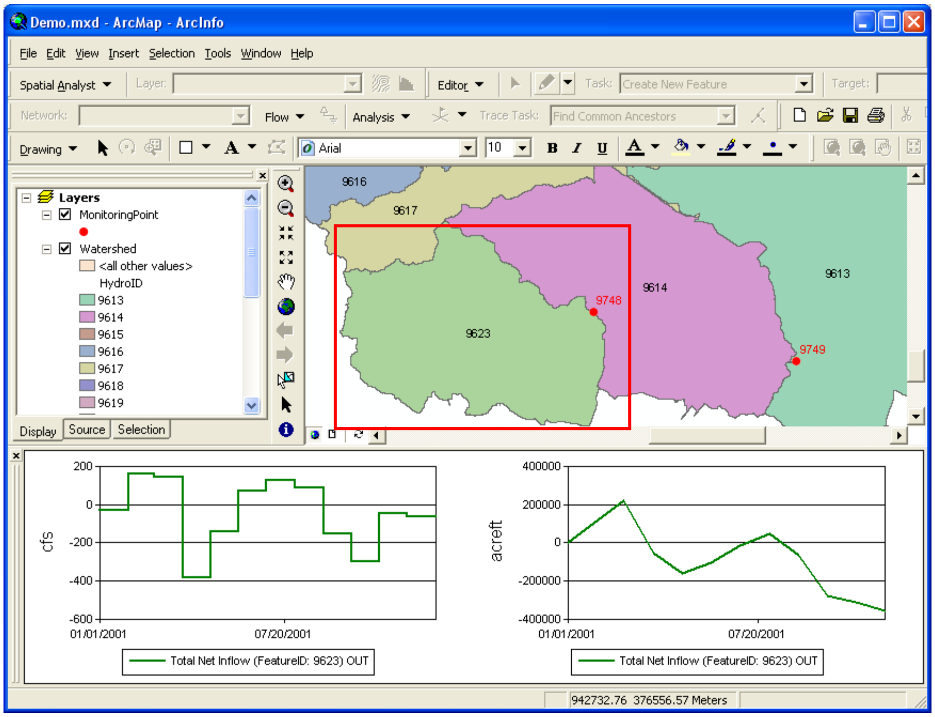


Figure 18. The total net inflow and its accumulation of storage through time for hydrovolume 9623.

8. Improve the water balance

The fluxes used in this example are all drawn from the NARR. Suppose one wished to improve the water balance with better data. Some options are:

Precipitation – Use gage-adjusted Nexrad radar precipitation instead of NARR modeled precipitation. The Neuse digital watershed presently contains one year of daily Nexrad data for 2004 and a more extensive historical archive of Nexrad data will be created. These data are acquired from the NWS West Gulf River Forecast Center.

Evaporation – Use evaporation data from atmospheric flux towers in or near the Neuse basin to adjust the evaporation fields from the NARR to more realistically estimate evaporation fields over the basin.

Subsurface Recharge – estimate from change in piezometric head elevation in the surficial aquifer. Figure 19 shows groundwater level measurements obtained form the North Carolina Division of Water Resources groundwater database (http://www.ncwater.org/Data_and_Modeling/Ground_Water_Databases/)
The measurements are in feet below the land surface at the well location, and have irregular time intervals between measurements. These data can be assembled and interpolated for each month, and the change in groundwater storage in the surficial aquifer estimated from the change in

piezometric head elevation and knowledge of the aquifer properties. The USGS is presently constructing a Modflow model for this aquifer and when completed, that model could be used to improve the surface water balance through better estimates of the recharge and discharge between surface, soil and groundwater.

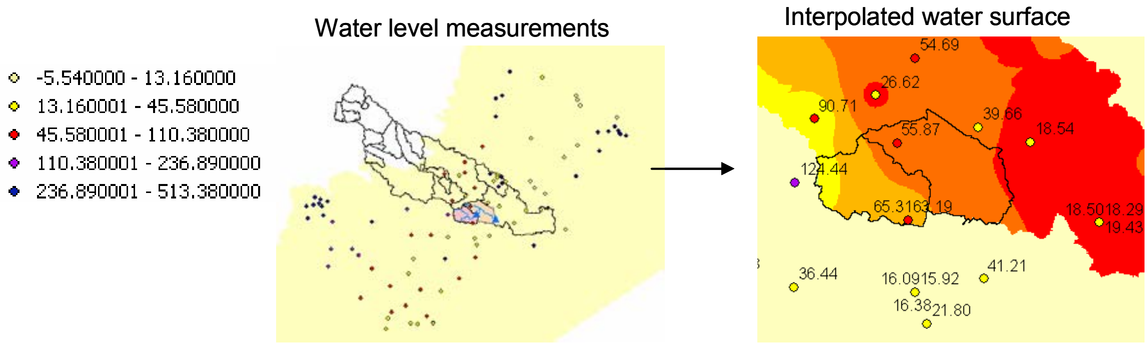


Figure 19. Average Water level measurements for February 2001 and water surface interpolated for that date

Conclusions

The methodology described in this paper shows how water balances involving hydrologic fluxes, flows and storage can be computed for watersheds within the Neuse river basin. The same methodology can be used for energy balances on the watershed surface, and with some extension, could also be applied to mass balancing of chemical and biological constituents. These balances are defined on hydrovolumes which are volumes in space drawn around the watersheds through which water, energy and mass flow, are stored internally, and transformed.

The methods described here depend on coupling fluxes and flows in both continuous and discrete space-time, accomplished for the continuous fields by using the netCDF data file format, and for the discrete space-time fields by using the Arc Hydro geospatial time series approach. A hydrologic flux coupler links flux, flow and storage and enables automated computations and unit conversions in constructing the water, mass and energy balances.

References

Maidment, D.R., (2002), “Arc Hydro: GIS for water resources”, ESRI Press, Redlands CA.

Maidment, D.R., (2005), “A data model for hydrologic observations”, Paper presented to the CUAHSI Hydrologic Information Systems Symposium, University of Texas at Austin, March.

Merwade, V. (2004). "Geospatial Desecription of River Channels in Three Dimensions". PhD dissertation submitted to the Graduate School, University of Texas at Austin. http://www.crwr.utexas.edu/gis/gishydro04/Modeling/Data/dissertation_merwade.pdf

Reckhow, K., et al., (2004), “Designing hydrologic observatories: a paper prototype of the Neuse watershed”, CUAHSI Technical Report No. 6, Consortium of Universities for the Advancement of Hydrologic Science, Inc, 84 pp., December.

Water Science and Technology Board (2001), “Envisioning the agenda for water resources research in the twenty-first century”, National Research Council, National Academy Press, Washington DC, 70p.

Chapter 10

Data Driven Discovery

Praveen Kumar[1], Peter Bajcsy[2], Vikas Mehra[1], Pratyush Sinha[1], Benjamin L. Ruddell[1], Amanda B. White[1], David Tcheng[2], David Clutter[2], and Wei-Wen Feng[2]
[1]Department of Civil and Environmental Engineering, [2]National Center of Supercomputing Applications,
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 [e-mail:kumar1@uiuc.edu]

I. Introduction

The advancement of hydrologic science using the data collected at hydrologic observatories is critically predicated on our ability to analyze very large volumes of data. The observatories aim to synthesize available historical data along with new measurement from a dense array of sensors. Furthermore, satellites data provides additional spatial coverage of critical variables and are expected to be an important part of an observatory investigation. The Data Driven Discovery component of the CUAHSI-HIS project aims to provide a set of tools for handling a variety of data sets and the capability to analyze them in a systematic framework. Because of the size of the data volume a paradigm shift is needed in our thinking to achieve these goals. This is primarily because we may continue to use traditional tools of scientific inquiry, such as statistical analysis or data assimilation, over these large datasets. There are several limitations of these methods. These techniques do not work very well for heterogeneous datasets resulting in small fragments of the entire volume being used. This limits our ability for formulating and testing hypothesis. In addition, our scientific vision is stymied due to the use of fragmented and limited datasets, and our ability to handle only "few variables" at a time. This limits the nature of hypothesis that are proposed and tested. The value of data is typically predicated on the ability to extract higher level information: information useful for decision support, for exploration, and for better understanding of the phenomena generating the data. Our traditional physics based and data driven approaches of scientific inquiry breakdown as the volume and dimensionality of data increases, thereby reducing the value of observed data.

The premise of our development effort is that scientific inquiry methods developed for small datasets or “few variable" problems may not be effective for large datasets or “many variable" problems. During the last several years, data mining or automatic knowledge discovery in databases (KDD) tools capable of identifying implicit knowledge in databases have become available and these tools address some of the limitations identified above. Their use in commercial settings has lead to very successful applications. However, their specialized use for various scientific problems is limited, but initial work is underway. Data mining application to scientific data will enable us to develop hypothesis about relationships of variables from observed data. These new hypothesis combined with the existing understanding of the physical processes we already have can result in an improved understanding and novel formulations of physical laws and an improved predictive capability.

Our development approach can be classified into three categories.

1. Developing support for integrating spatio-temporal data in GIS geodata model framework.
2. Developing data mining support for remote sensing data products.
3. Developing data mining algorithms for time-series observational data.

Our development effort will use the D2K (http://alg.ncsa.uiuc.edu/do/tools/d2k) application environment for data mining. D2K is a rapid, flexible data mining and machine learning system that integrates analytical data mining methods for prediction, discovery, and deviation detection, with data and information visualization tools. It offers a visual programming environment that allows users to connect programming modules together to build data mining applications and supplies a core set of modules, application templates, and a standard API for software component development. All D2K components are written in Java for maximum flexibility and portability. Major features that D2K provides to an application developer include:

1) Visual Programming System Employing a Scalable Framework
2) Robust Computational Infrastructure
    a. Enables processor intensive applications
    b. Supports distributed computing
    c. Enables data intensive applications
    d. Provides low overhead for module execution
3) Flexible and Extensible Architecture
    a. Provides plug and play subsystem architectures and standard APIs
    b. Promotes code reuse and sharing
    c. Expedites custom software developments
    d. Relieves distributed computing burden
4) Rapid Application Development (RAD) Environment
5) Integrated Environment for Models and Visualization

6) *D2K Module Development*: NCSA's Automated Learning Group (ALG) has developed hundreds of modules that address every part of the data mining process. Some data mining algorithms implemented include Naive Bayesian, Decision Trees, and apriori, as well as visualizations for the results of each of these approaches. In addition, ALG has developed modules for cleaning and transforming data sets and a number of visualization modules for deviation detection problems. Modules have also been created for specific projects and collaborations.

ALG NCSA is continuing development of modules with the short-term goal of enhancing the cleaning and transformation modules, improving the data mining algorithms and continuing development of feature subset selection modules. Long-term, ALG plans to continue development of modules for predictive modeling, image analysis and textual analysis, particularly toward enabling them for distributed and parallel computing. This type of work expedites the process of applying the latest research developments to be used on real-world applications.

7) D2K-driven Applications: D2K can be used as a stand-alone application for developing data mining applications or developers can take advantage of the D2K infrastructure and D2K modules to build D2K-driven applications such as the ALG application I2K-Image to Knowledge. These applications employ D2K functionality in the background, using modules dynamically to construct applications. They present their own specialized user interfaces specific to the tasks being performed. Advantages of coupling with D2K to build highly functional data mining applications such as these include reduced development time through module reuse and sharing, and access to D2K distributed computing and parallel processing capabilities.

Below we summarize the development effort in each of these three categories.

II. Modelshed: A tool for integrating spatio-temporal data in GIS geodata model framework

The ArcHydro data model for water resources has been successfully established as a standard for the modeling and communication of hydrologic datasets, and is being adopted by many branches of government, industry, and the academy. However, it is still difficult to process large gridded datasets from numerical simulations and remote sensors, and to meaningfully relate that data to other objects in an ArcHydro-modeled database. The Modelshed geodata model is presented as a generalized GIS data model for the organization and modeling of diverse geospatial data. Modelshed is a geodata model for diverse environmental science and hydrologic applications, capable of representing four-dimensional (4D) model domains, vertical layering, environmental fluxes, dynamic spatial features, statistical timeseries data, and relationships between heterogeneous model domains (Fig. 1). Modelshed extends the capabilities of the ArcHydro data model, and is fully compatible with that model's structures and software tools. It is based on the ESRI ArcObjects™ and geodatabase technologies, and therefore stores its data objects with geospatial location and projection information compatible with OpenGIS spatial metadata standards. Modelshed-modeled data may be read and spatially integrated by GIS applications, and its data may be accessed by industry-standard database software such as Microsoft Access™, Oracle™, and IBM DB2™. With the added flexibility, Modelshed is able to model a diverse variety of environmental systems, and connect those systems with the hydrologic structures modeled in ArcHydro (Fig. 2). It provides data structures to facilitate the geospatial analysis of time-indexed raster datasets and the integration of raster data with the vector structures of the data model. The study of relationships within this data model is simple and powerful, based on queries of indexed data tables in a relational database. The entire suite of modelshed tools is available at: http://cee.uiuc.edu/research/hydrology/hydroinf_Intro.html

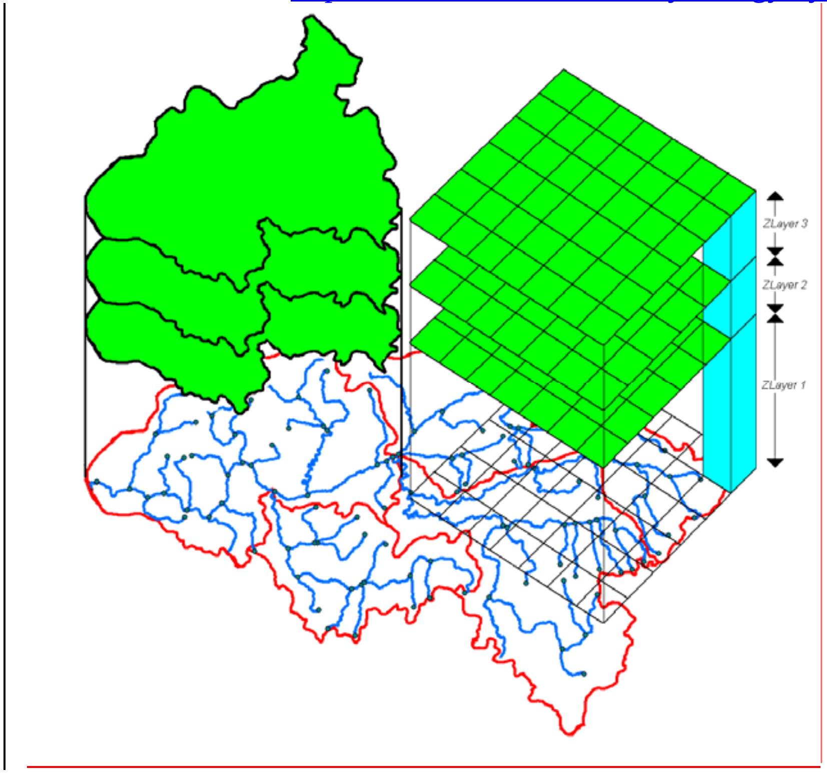


Figure 3: A visualization of the Modelshed geodata model framework, including grid and watershed-based Modelshed Type domains, vertical ZLayer indexing, and the associated hydrographic data in an ArcHydro-compatible model (from Ruddell and Kumar, 2005).

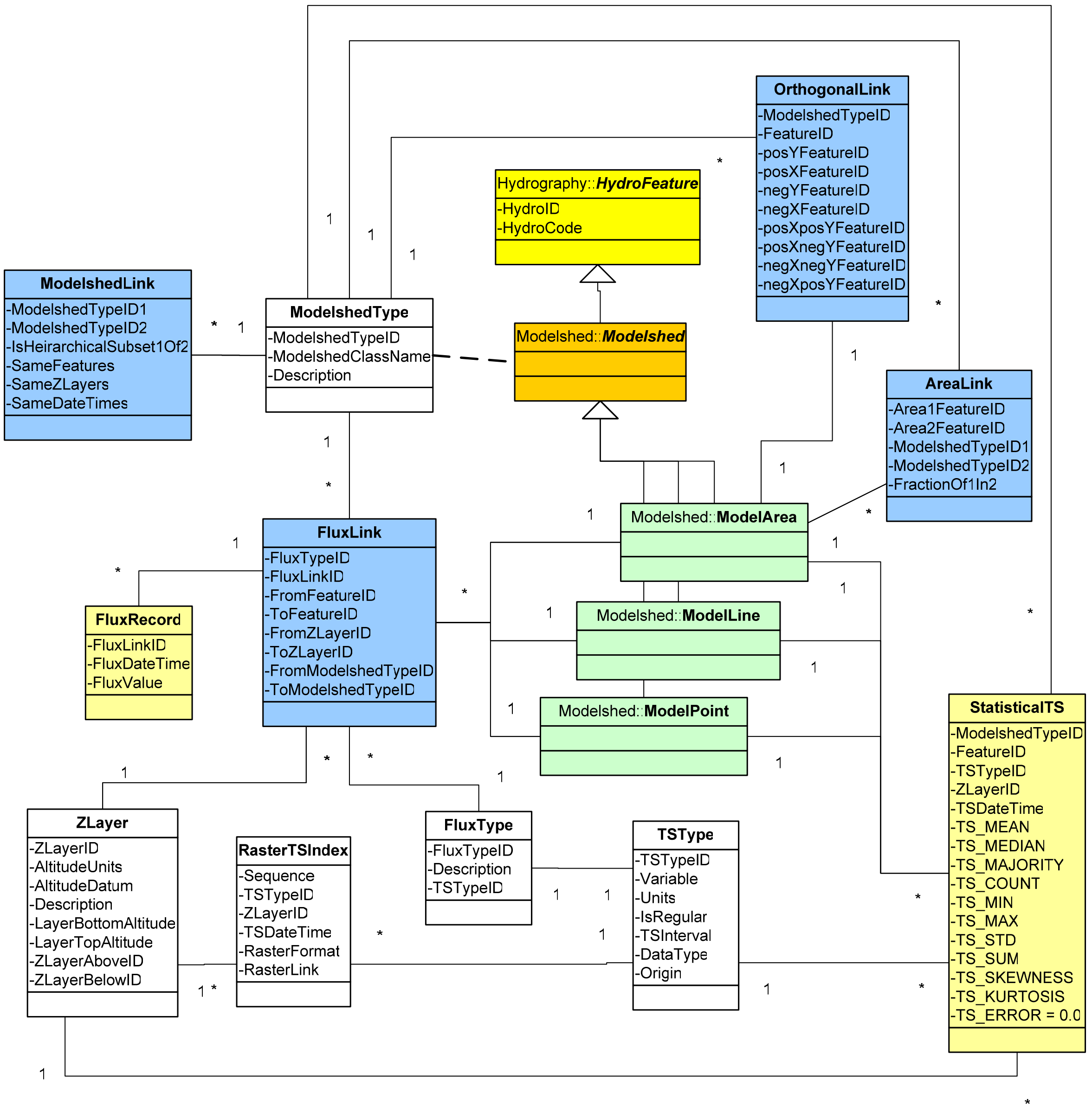
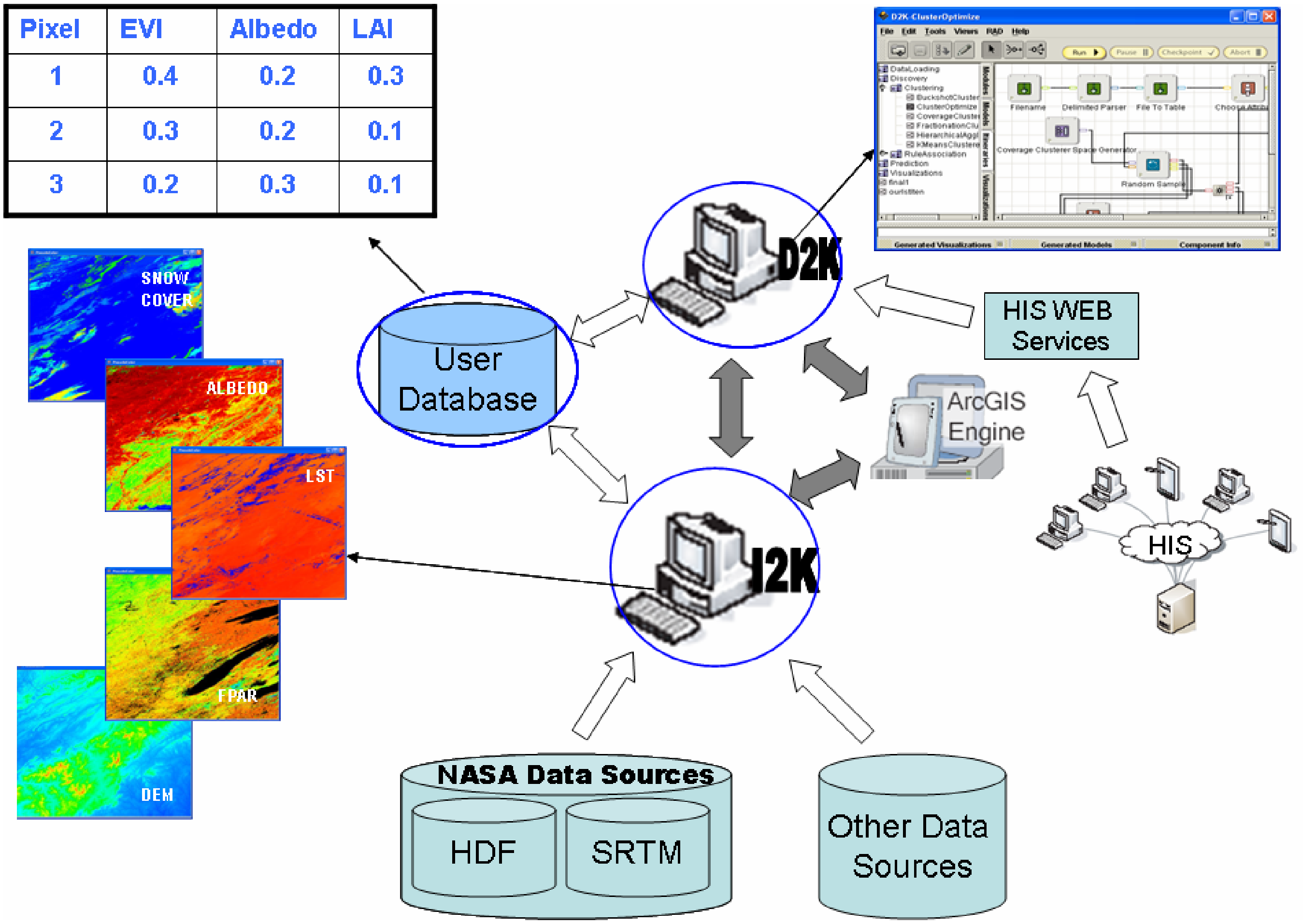


Figure 4: Modelshed framework UML and data dictionary. The Modelshed geodata model combines spatial features inheriting from the ArcHydro HydroFeature with a network of metadata classes.



Pixel	EVI	Albedo	LAI
1	0.4	0.2	0.3
2	0.3	0.2	0.1
3	0.2	0.3	0.1

Fig. 3. Illustration of overall system architecture for data ingestion, preprocessing, integration, visualization and data analysis using various data mining algorithms. I2K reads all data sets from different data sources and visualize (snow cover, Albedo). It calls GIS functions using ArcGIS engine interface to perform feature extraction tasks (slope, aspects). All measured and derived variable are ingested in to database after preprocessing (spatial and temporal adjustment, removing bad pixels using QA/QC). D2K is used to analyze this database and results are visualized in I2K.

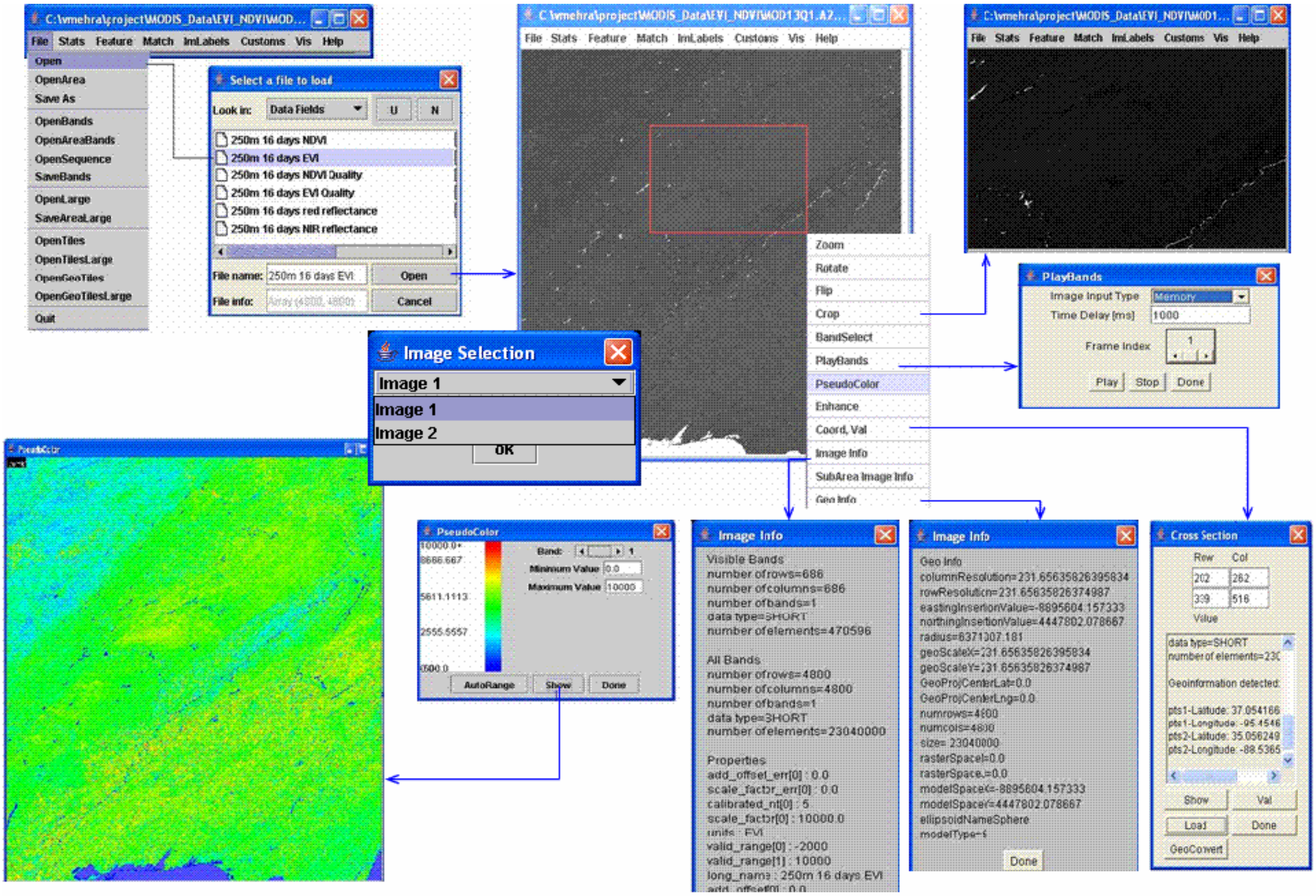
III. Data mining support for remote sensing data products

To support various data formats, a common interface is designed to visualize, preprocess and analyze the data. Some of these datasets include hierarchical data formats (HDF), digital elevation model (DEM), and geographical information system (GIS) supported vector files. The overall system architecture has been divided into four parts (Fig. 1). These components are explained below:

1) *Read Different data format using I2K:* I2K is an image analysis tool, designed to automate processing of huge datasets and is capable of analyzing multi-dimensional and multivariate image data. When analyzing multiple geographic datasets over similar geographic areas, it is necessary to preprocess and integrate heterogeneous datasets. I2K is a key component for preprocessing, visualizing and integrating the diverse datasets. I2K uses HDF libraries to load HDF data, and links to ArcGIS Engine functionalities

to operate on GIS readable data formats. Fig. 3 shows the visualization of different scientific data sets: Snow cover, Albedo, LST (Land Surface Temperature), FPAR (fraction of Photosynthetic active radiation) and DEM.

Fig. 4 shows Graphical User Interface (GUI) associated with the visualization of HDF data in I2K. An HDF file may contain more than one scientific data set. User can select the scientific dataset (SDS) for display. Once the image is loaded, user can zoom, crop and play all spectral bands. Geographical information and image related information associated with the data sets can also be viewed by selecting GeoInfo and ImageInfo options respectively in the menu bar.



```
Visible Bands
number ofrows=686
number ofcolumns=686
number ofbands=1
data type=SHORT
number ofelements=470596

All Bands
number ofrows=4800
number ofcolumns=4800
number ofbands=1
data type=SHORT
number ofelements=23040000

Properties
add_offset_err[0] : 0.0
scale_factor_err[0] : 0.0
calibrated_nt[0] : 5
scale_factor[0] : 10000.0
units : EVI
valid_range[0] : -2000
valid_range[1] : 10000
long_name : 250m 16 days EVI
add_offset[0] : 0.0
```

```
Geo Info
columnResolution=231.65635826395834
rowResolution=231.65635826374987
eastingInsertionValue=-8895604.157333
northingInsertionValue=4447802.078667
radius=6371007.181
geoScaleX=231.65635826395834
geoScaleY=231.65635826374987
GeoProjCenterLat=0.0
GeoProjCenterLng=0.0
numrows=4800
numcols=4800
size= 23040000
rasterSpaceI=0.0
rasterSpaceJ=0.0
modelSpaceX=-8895604.157333
modelSpaceY=4447802.078667
ellipsoidNameSphere
modelType=6
```

```
data type=SHORT
number of elements=230

Geoinformation detected:

pts1-Latitude: 37.054166
pts1-Longitude: -95.4546
pts2-Latitude: 35.056249
pts2-Longitude: -88.5365
```

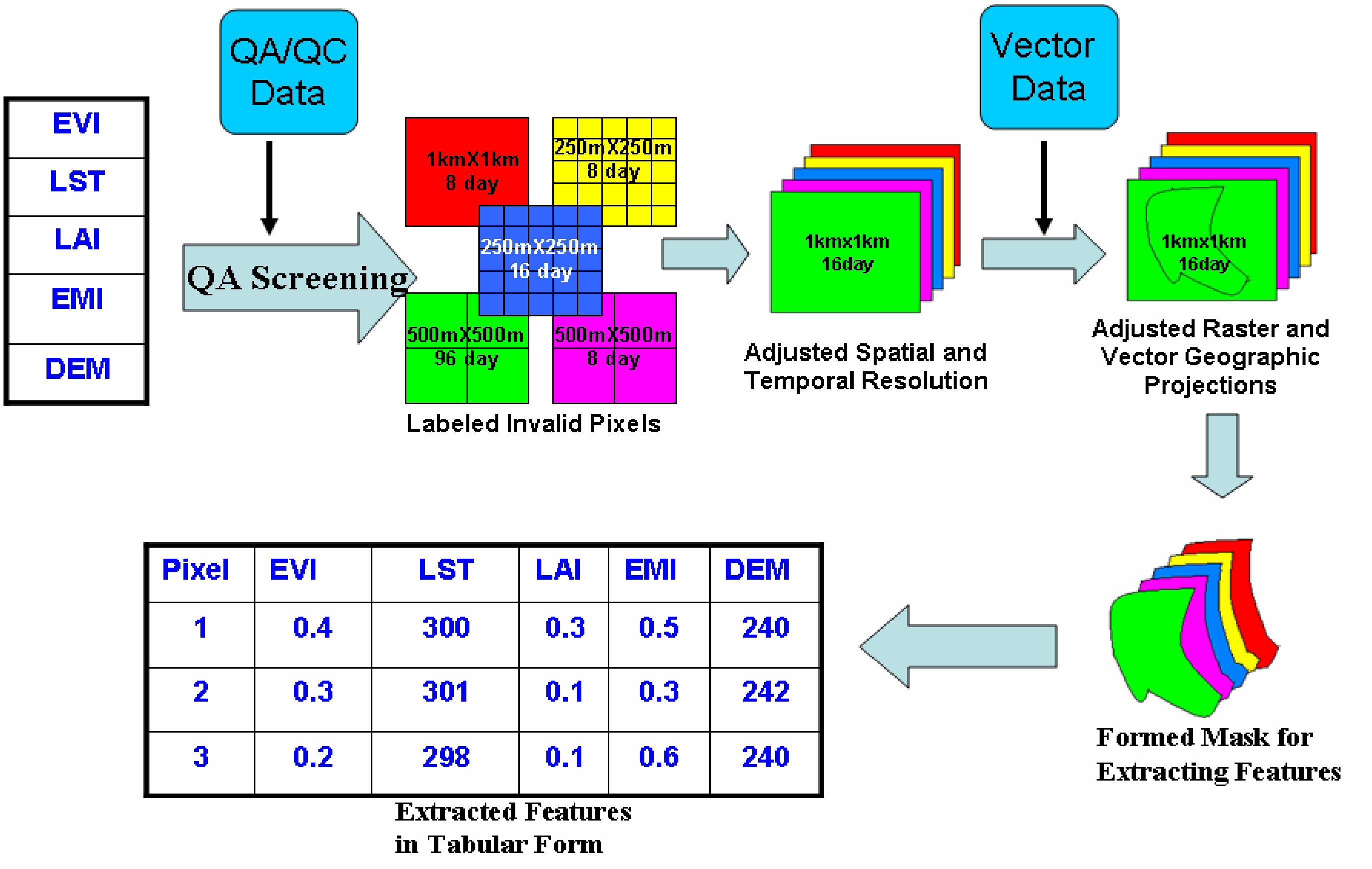
Fig. 4. Generic tool for loading different datasets. Interactive visualization environment (zoom, crop, geographical information, play all spectral bands of data) for integrating data mining and visualization processes.

2) *ArcGIS Engine*: It is a complete library of GIS components which can be embedded into custom applications. I2K links to these libraries for features extractions e.g. calculate flow accumulation grid from DEM, calculate slope and aspects from DEM. These derived variables are used for analysis along with the measured data sets.

3) *Create Relational Database:* Creating a user Database (Fig. 5) is a data preprocessing and integration step. Different scientific datasets like Enhanced Vegetation Index (EVI), Albedo, Leaf Area

Index (LAI), Emissivity and Sea Surface Temperature (SST) are at different spatial and temporal resolution. Also there is quality assurance and quality control (QA/QC) data associated with each scientific variable. QA/QC data provide information about the quality of data for each pixel inside a scientific dataset.

To create an analysis database, we need to choose a unique spatial and temporal resolution. This is done by upscaling or downscaling the data. The unique spatial and temporal resolution is supplied by user as an input before creating the database. User may be interested in analyzing the data for a particular region only (Fig. 3). In that case he can create a mask by selecting the area that he wants to analyze. QA/QC data is used to remove bad pixel values e.g. no data values or bad pixel data received by satellite due to clouds. This option is again provided by user. After all the above processing is done, integrated scientific and derived data sets are written into a database (Fig. 5).



Pixel	EVI	LST	LAI	EMI	DEM
1	0.4	300	0.3	0.5	240
2	0.3	301	0.1	0.3	242
3	0.2	298	0.1	0.6	240

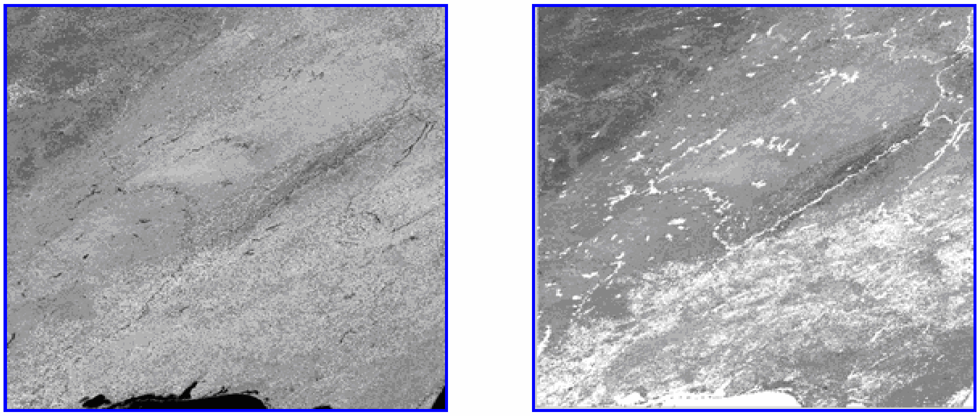
Fig. 5. Remote sensing data product analysis workflow. Database Table includes scientific data and derived variables (slope, aspects) after performing multiple preprocessing operations (use QA/QC data to remove bad pixel or no data values, spatial and temporal sampling adjustments, masking data sets, and error checking) and data integration.

4) *Use D2K for data mining:* This task plays the central role to enable automatic knowledge discovery through data mining. D2K uses database created in the above step as an input. It has modules for variety of algorithms like multiple regression, Naïve Bayes, Decision Tree, and Neural Network to find various characteristic of data sets. Scientific question which we aim to answer are: (1) identify the dependence of the dynamically evolving variables on each other and their temporal scales of variability and identify the roles of climate variability as a determinant of the variability in the dynamically observed quantities (2) identify how land-surface characteristics (elevation, slope, aspects, soil properties etc) further modulate the dynamical evolution of vegetation.

Overall procedure can be summarized as follows (Fig. 3):

STEP 1. Read all data sets using I2K
STEP 2. Visualize each scientific data set
STEP 3. Use native I2K functions along with ArcGIS Engine links to perform various feature extraction tasks.
STEP 4. Use QA/QC to remove bad quality pixel
STEP 5. Perform upscaling or downscaling of SDS to get unique spatial and temporal resolution
STEP 6. Mask the data set
STEP 7. Data Integration by writing all SDS and derived variables from SDS into database
STEP 8. Use D2K to run data mining algorithms on database created in above step.
STEP 9. Visualize results in I2K.
STEP 10. Run Modelshed tools on the processed data.

Case Study for Blue Ridge Region



Original Image

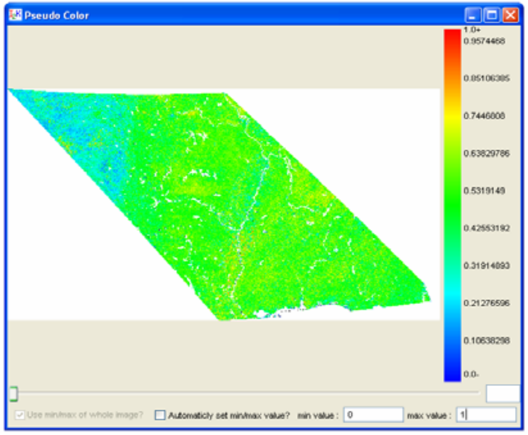
Image after applying QA/QC Mask

Fig. 6 Difference in original Image and Image after applying quality assurance/quality control mask. White region show pixels removed due to their bad quality.

In this section we use a case study for Blue Ridge Region to illustrate the potential of the techniques described in this paper. The purpose of this section is to provide an intuition as to how our system works based on the techniques described previously.

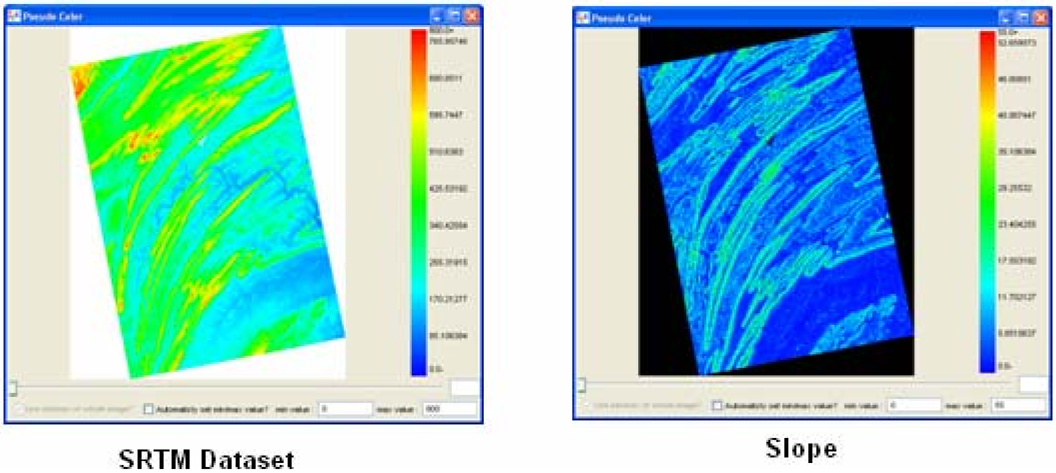
Preprocessing of remotely-sensed heterogeneous datasets (HDF Format): Let us suppose we have a Modis EVI dataset whose resolution is 250m, temporal scale is 16 day average, and projection is sinusoidal. Suppose our goal is to change spatial resolution to 500m, temporal resolution to 32 days average and projection to Albers Equal Area Conic. For analysis on remote sensing datasets, it is important to consider quality of the pixel. In our analysis, we are considering only the best quality pixels as identified by QA/QC mask available with our data products. The first step in preprocessing the datasets is to apply quality assurance and quality control mask and apply land water mask which will remove all water pixels from EVI datasets. Fig. 6 shows the comparison of original EVI dataset and processed quality EVI dataset. White regions show the pixels that were removed. This mask is applied to all EVI datasets selected for preprocessing.

The second step is to change the temporal resolution of EVI datasets. User can choose from any of the four methods for upscaling the temporal resolution. The methods are described as follows:
1. Assign no data as the average value of pixel, if pixel has no data for any one of the EVI dataset.
2. Neglect pixels with no data values and take average of others.
3. Replace all no data values pixels with 0 and take average.
4. Maximum of all values over time for the pixel from all EVI datasets is assigned (Maximum Value Composite).



☑ Use min/max of whole image? □ Automaticly set min/max value? min value : 0 max value : 1

Fig. 7 Reprojected and Resampled Image of Modis EVI dataset



SRTM Dataset

Slope

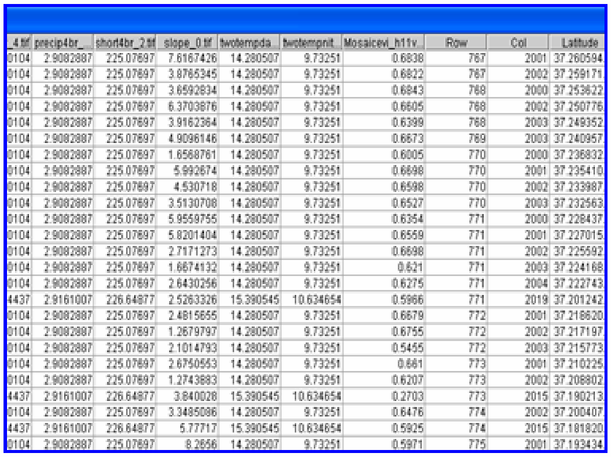
Fig. 8. Image shows SRTM dataset in albers equal area projection. Slope is one of feature extracted from SRTM dataset using I2K and arcengine.

Third step is to change the projection and resample the EVI dataset. This step calls arc engine functions to reproject and resample the datasets, and save in the disk in GEOTIFF format. Fig. 7 shows the reprojected (Albers Equal Area Conic) image of EVI datasets obtained after above processing. Similarly, other Modis datasets such as LAI, FPAR, LST, and Snow Cover can be processed and brought together at same spatial and temporal resolution.

Feature Extraction from SRTM and DEM Datasets: Feature extractions such as slope, aspect, contour, flow direction, and flow accumulation are calculated by calling functions in arc engine libraries or by native functions inmplemented in I2K for this purpose, and are saved on disk in (32 bit) GEOTIFF format. Each of this dataset can be reprojected to other projection using Arc Engine functions. Similarly, user can select the option for up sampling or down sampling of each datasets before extracting features from elevation. Fig. 8 shows the NASA SRTM dataset and one of the extracted features (slope).

Arc GRID format to TIFF: Any dataset which is in arc grid format can be reprojected, resampled and saved on disk in (32 bit) GEOTIFF format using I2K environment.

At this stage, preprocessing and integration of different datasets from various sources is complete. Using above mentioned tools most datasets can be brought together at same spatial and temporal scale that can be further used for analysis.



_4.tif	precip4br_	short4br_2.tif	slope_0.tif	twotempda	twotempnit	Mosaicevi_h11v	Row	Col	Latitude
0104	2.9082887	225.07697	7.6167426	14.280507	9.73251	0.6838	767	2001	37.260594
0104	2.9082887	225.07697	3.8765345	14.280507	9.73251	0.6822	767	2002	37.259171
0104	2.9082887	225.07697	3.6592834	14.280507	9.73251	0.6843	768	2000	37.253622
0104	2.9082887	225.07697	6.3703876	14.280507	9.73251	0.6605	768	2002	37.250776
0104	2.9082887	225.07697	3.9162364	14.280507	9.73251	0.6399	768	2003	37.249352
0104	2.9082887	225.07697	4.9096146	14.280507	9.73251	0.6673	769	2003	37.240957
0104	2.9082887	225.07697	1.6568761	14.280507	9.73251	0.6005	770	2000	37.236832
0104	2.9082887	225.07697	5.992674	14.280507	9.73251	0.6698	770	2001	37.235410
0104	2.9082887	225.07697	4.530718	14.280507	9.73251	0.6598	770	2002	37.233987
0104	2.9082887	225.07697	3.5130708	14.280507	9.73251	0.6527	770	2003	37.232563
0104	2.9082887	225.07697	5.9559755	14.280507	9.73251	0.6354	771	2000	37.228437
0104	2.9082887	225.07697	5.8201404	14.280507	9.73251	0.6559	771	2001	37.227015
0104	2.9082887	225.07697	2.7171273	14.280507	9.73251	0.6698	771	2002	37.225592
0104	2.9082887	225.07697	1.6674132	14.280507	9.73251	0.621	771	2003	37.224168
0104	2.9082887	225.07697	2.6430256	14.280507	9.73251	0.6275	771	2004	37.222743
4437	2.9161007	226.64877	2.5263326	15.390545	10.634654	0.5966	771	2019	37.201242
0104	2.9082887	225.07697	2.4815655	14.280507	9.73251	0.6679	772	2001	37.218620
0104	2.9082887	225.07697	1.2679797	14.280507	9.73251	0.6755	772	2002	37.217197
0104	2.9082887	225.07697	2.1014793	14.280507	9.73251	0.5455	772	2003	37.215773
0104	2.9082887	225.07697	2.6750553	14.280507	9.73251	0.661	773	2001	37.210225
0104	2.9082887	225.07697	1.2743883	14.280507	9.73251	0.6207	773	2002	37.208802
4437	2.9161007	226.64877	3.840028	15.390545	10.634654	0.2703	773	2015	37.190213
0104	2.9082887	225.07697	3.3485086	14.280507	9.73251	0.6476	774	2002	37.200407
4437	2.9161007	226.64877	5.77717	15.390545	10.634654	0.5925	774	2015	37.181820
0104	2.9082887	225.07697	8.2656	14.280507	9.73251	0.5971	775	2001	37.193434

Fig. 9. First Image shows the eco region mask applied over 19 variables and second image shows the table created for selected Blue Ridge region. Table includes latitude and longitude of each pixel lies with in the boundary.

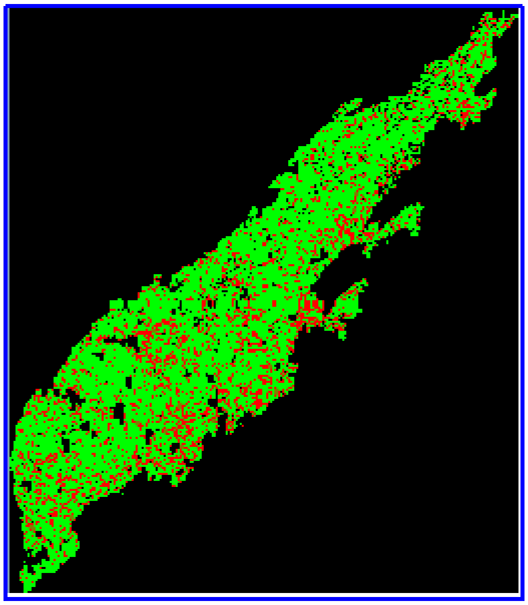


Fig. 10. Image shows the results obtained by applying decision tree model of D2k on above created table for Blue Ridge region.

Decision Tree Analysis using D2K Itinerary: In our experiment, we have taken 19 datasets at a time to run data mining algorithm. These datasets are EVI, LULC (Land Use Land Cover), slope, aspect, flow direction, flow accumulation, LST, Climate variables such as precipitation, short wave, long wave, temperature day, temperature night, and soil properties such as pH, AWC (available water capacity), percentage clay, silt and sand. All these datasets are loaded in the I2K environment for analysis. Different Tiles of HDF datasets are mosaicked together to get one mosaicked image. Mosaicking is necessary because Blue Ridge region spans more than one Modis tile. After geographically aligning all datasets, ecoregion mask is applied. User can select the boundary for Blue Ridge region and create table of all the pixels lying within the selected boundary. Each row in this table contains pixel value for all 19 variables along with latitude and longitude of the pixel (Fig. 9). It neglects all those pixels which have no data value in any one of the datasets.

D2K itinerary (Decision Tree model) is applied on the above table while taking EVI as a dependent variable and all other datasets as independent variable. Relevance of whole tree and relevance of the input variables (slope, aspect, land cover) to the output variable (EVI) at every depth of tree is calculated. Results from decision tree are mapped back spatially for visualization at each level of tree (Fig. 10).

IV. DATA INGESTION AND ANALYSIS USING WEB SERVICES

We have also developed a web service interface in the D2K environment to access other federal data sources as provided by the hydrologic information science portal. Web service is a standardized way of transferring the data on Internet using XML messaging system. XML is used to encode all communications to a Web service. A client can invoke a Web service by sending an XML message, and then wait for a corresponding XML response. Because all communication is in XML, Web services are not tied to any one operating system or programming language—Java can talk with .Net; Windows applications can talk with UNIX applications etc. One of the advantages of web service is that the data can be requested and received within an application. Received data can be then processed inside the application. Thus there is no need of separately downloading data and then doing the processing.

Case Study of CUAHSI NWIS Web Services: NWIS Web Data for USA provides access to water-resources data collected in the USA. To get the data, a user goes through the tedious task of going to the USGS NWIS website (http://waterdata.usgs.gov/nwis/) and filling site selection criterion, timestamp etc. He then clicks the submit button to save the data in a file locally. User can then process this data on his desktop. A web service implementation of this service was done under CUAHSI (Consortium of University for Advancement of Hydrologic Science, Inc.) effort. The WSDL URL for the web service is http://water.sdsc.edu/hydrologictimeseries/nwis.asmx?WSDL. The web service is written on .Net platform. It supports a bunch of functions through which desired data can be ingested directly into the application. If the application supports a GUI, user can vary his parameters and simultaneously visualize results.

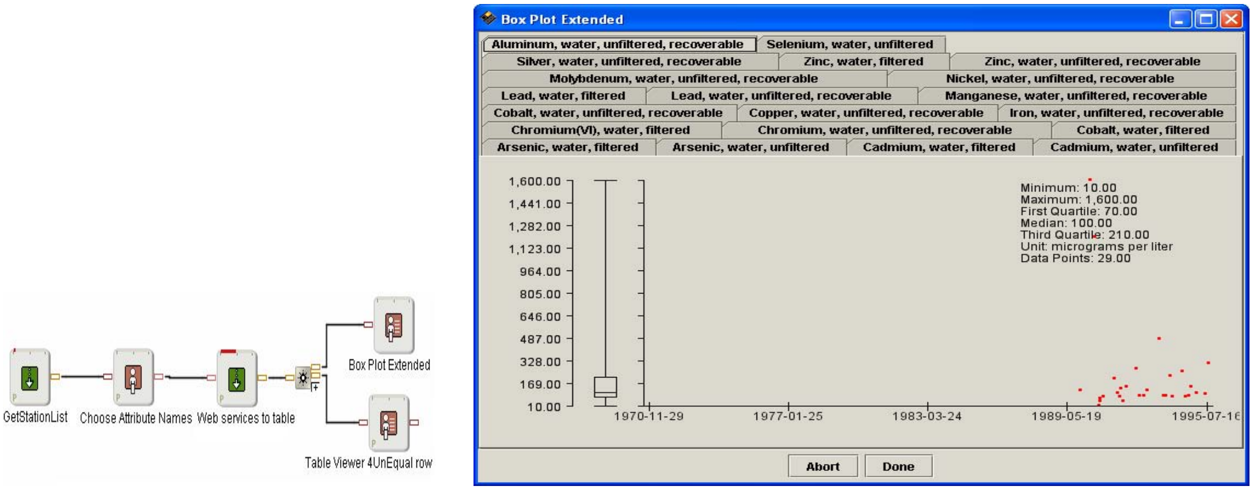

GetStationList
Choose Attribute Names
Web services to table
Box Plot Extended
Table Viewer 4UnEqual row
Box Plot Extended
Aluminum, water, unfiltered, recoverable
Selenium, water, unfiltered
Silver, water, unfiltered, recoverable
Zinc, water, filtered
Zinc, water, unfiltered, recoverable
Molybdenum, water, unfiltered, recoverable
Nickel, water, unfiltered, recoverable
Lead, water, filtered
Lead, water, unfiltered, recoverable
Manganese, water, unfiltered, recoverable
Cobalt, water, unfiltered, recoverable
Copper, water, unfiltered, recoverable
Iron, water, unfiltered, recoverable
Chromium(VI), water, filtered
Chromium, water, unfiltered, recoverable
Cobalt, water, filtered
Arsenic, water, filtered
Arsenic, water, unfiltered
Cadmium, water, filtered
Cadmium, water, unfiltered
1,600.00
1,441.00
1,282.00
1,123.00
964.00
805.00
646.00
487.00
328.00
169.00
10.00
1970-11-29
1977-01-25
1983-03-24
1989-05-19
1995-07-16
Minimum: 10.00
Maximum: 1,600.00
First Quartile: 70.00
Median: 100.00
Third Quartile: 210.00
Unit: micrograms per liter
Data Points: 29.00
Abort
Done

Fig. 11 D2K workflow and box plot and time series plot of Water Quality (WQ) data for Minor and Trace Inorganic for USGS station 02087701. In the figure tab for Aluminum, water, unfiltered, recoverable WQ is selected. Plot shows that from May 1989 to July 1995, 29 measurements were taken and unit of measurement is microgram per liter. Statistics associated with the measurement are: minimum: 10 µg/l, maximum: 1600 µg/l, first quartile: 70 µg/l, median: 100 µg/l and third quartile: 210 µg/l.

We have developed an application in the D2K platform which connects to CUAHSI NWIS web services to get data and displays time series and box plot for stream flow and water quality data for a given station. Fig. 11 shows the D2K workflow and box plot and time series plot of Water Quality (WQ) data for Minor and Trace Inorganic for USGS station 02087701. In this application, a user can specify his region of interest by specifying a latitude-longitude box and application connects to the NWIS web services and retrieves a list of stations which have WQ data record and which lie inside the given latitude-longitude box. User can then select a particular station and retrieve plot for all classes of WQ data or specify a specific class of WQ data e.g. Minor and Trace Inorganic WQ data and view the plot. The integration of remote sensing analysis capability along with web service access to point (in situ) observation provides a unique opportunity to pursue issues related to the joint analysis and the trade off between space-time resolutions associated with these datasets.

V. Data mining algorithms for time-series observational data

HIS data repository will include time series data, remote sensing data, and 3D hydrovolume data. The data in the repository can be conceptualized as a data cube (Fig 12). A particular observed data value is located as a function of where it was observed, its time of observation, and what kind of variable it is. According to the distribution of measurement points, space can be sliced into four categories – 1) Zero Dimensional Space 2) One Dimensional Space 3) Two Dimensional Space 4) Three Dimensional Space. Below we present preliminary concepts for development of a framework for the systematic analyses of data in the data cube.

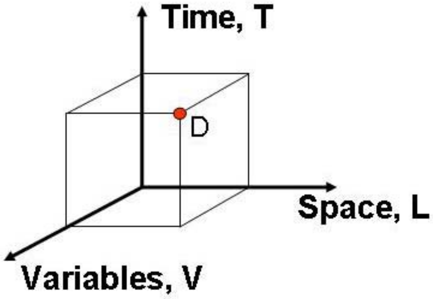
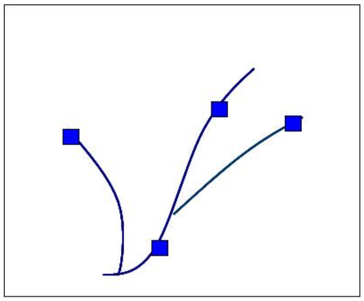


Figure 12: Data cube

Zero Dimensional Space: Measurements at zero dimensional space are essentially point measurement. These are time series data. An example includes streamflow data at a gaging station. Time series data can be plotted on a Cartesian coordinate system or a box plot of the variables can be drawn to explore maximum, minimum, median, upper quartile, lower quartile and outliers in the dataset. If there are multiple time series data, we can do a multiple box plot of the variables or draw a 2D or 3D scatter plot between any two or three variables respectively. To visualize more than three variables in one representation, a parallel coordinate plot can be drawn in which each observation in a data set is represented as an unbroken series of line segments which intersect vertical axes, each scaled to a different variable. We can also do a correlation study among different variables. Leveraging time series data mining techniques, similar patterns can be searched for in a time series or given a time series database, two time series which behave 'almost similarly' can be discovered. Discovery of such patterns in huge datasets by standard statistical method may not be a trivial task.

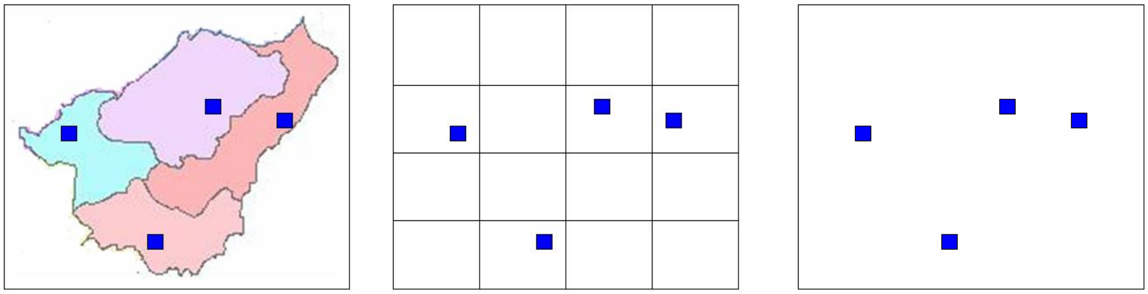
One Dimensional Space: Measurements along a line are measurements in one dimensional space. An example includes measurement of streamflow at multiple gaging stations along a river channel.



Measurement points along a line

In zero dimensional space analysis, time series data was studied in isolation. In one dimensional space analysis, the emphasis is on understanding the inherent relationship among the events along the channel. Streamflow data along the gaging stations can be used to determine travel-time and residence-time. Using data for concentrations of sediments or nutrients like total phosphorus and ammonia nitrogen, and BOD, we can study the pattern of variation in concentrations from upstream to downstream.

Two Dimensional Space: Measurement points scattered over a space constitute a two dimensional measurement space. One dimensional space is a special case of two dimensional space where all the measurement points are along a line. A two dimensional space analysis of data will depend on the spatial representation of measurement points. The points may be an area average, or may be lying on a grid, or just discrete points in space.



a) Measurement points as area average b) Measurement points lying on grid c) Measurement points in discrete space

An example of gridded data in two dimensional space is remote sensing data. Tools for the analysis of such data are described in the previous section.

Three Dimensional Space: Measurement points are scattered over a volume. Data analysis of such a space has been dealt in the Modelshed framework presented above.

VI. Summary and Conclusions

With the recent emergence of field of data mining, there is a need for a system that can handle large data sets and data assembly, preprocessing, and integration tasks. We are developing I2K as a common interface which can load HDF data from NASA data sources, supports visualization as well as data preprocessing and integration tasks (see Figure 13). Further, it can use and extend functionalities present in ArcGIS using ArcEngine. Data mining algorithms present in D2K are applied on integrated data sets to find various patterns and relations between different variables. Further, combining all these capabilities with ingesting NWIS data using CUAHSI web data services, provides a conceptual framework to take a watershed, derive data from remotely sensed images ingest stream flow or water quality information, will help hydrologist to answer fundamental science questions. The understanding developed through our analyses will enable us to better parameterize the various natural processes for weather and climate models and thereby improving their predictability.

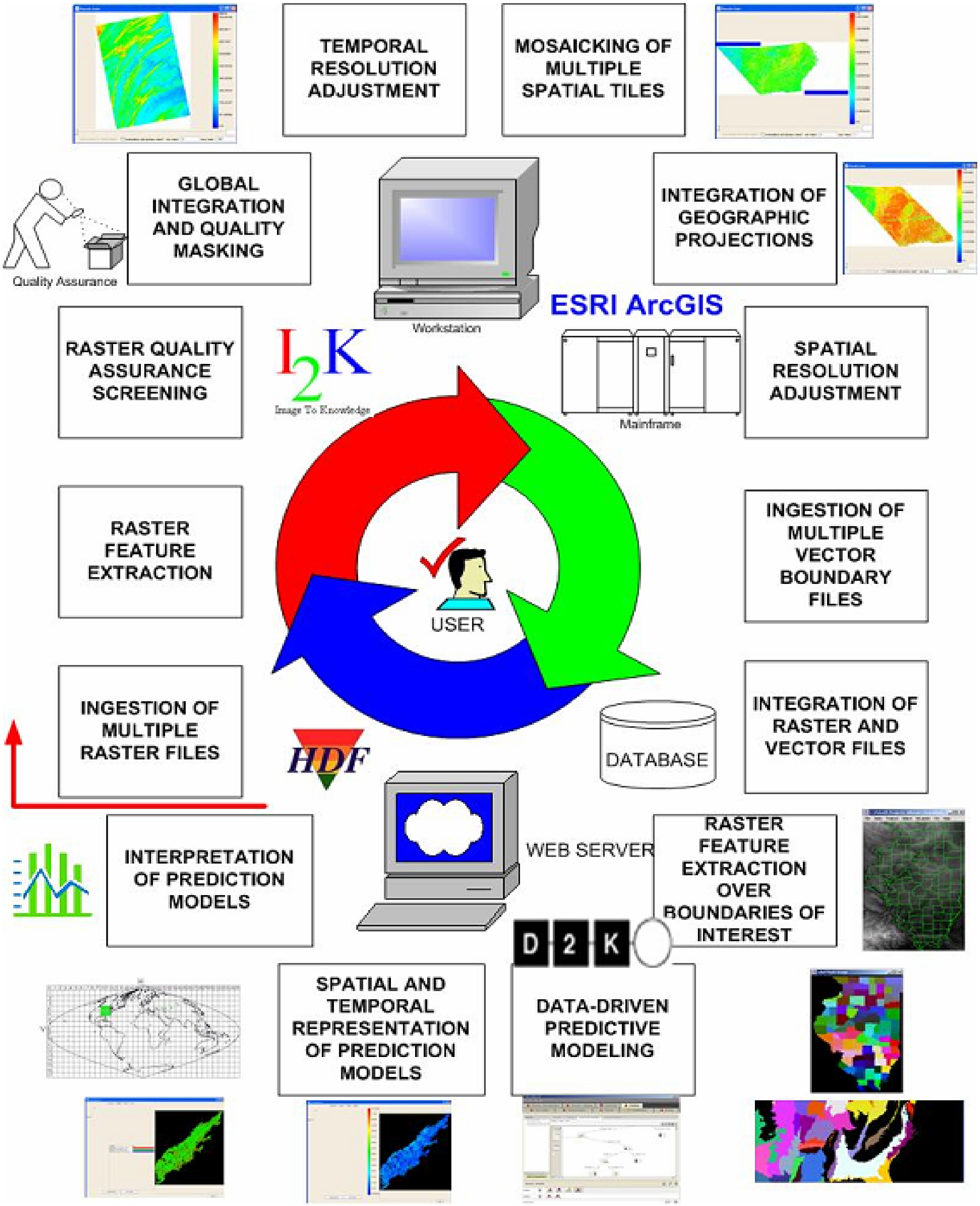

TEMPORAL RESOLUTION ADJUSTMENT
MOSAICKING OF MULTIPLE SPATIAL TILES
GLOBAL INTEGRATION AND QUALITY MASKING
Quality Assurance
Workstation
INTEGRATION OF GEOGRAPHIC PROJECTIONS
ESRI ArcGIS
RASTER QUALITY ASSURANCE SCREENING
I2K
Image To Knowledge
Mainframe
SPATIAL RESOLUTION ADJUSTMENT
RASTER FEATURE EXTRACTION
USER
INGESTION OF MULTIPLE VECTOR BOUNDARY FILES
INGESTION OF MULTIPLE RASTER FILES
HDF
DATABASE
INTEGRATION OF RASTER AND VECTOR FILES
INTERPRETATION OF PREDICTION MODELS
WEB SERVER
RASTER FEATURE EXTRACTION OVER BOUNDARIES OF INTEREST
D2K
SPATIAL AND TEMPORAL REPRESENTATION OF PREDICTION MODELS
DATA-DRIVEN PREDICTIVE MODELING

Figure 13: System functionalities currently vailable for the analysis of MODIS data products.

ACKNOWLEDGMENT

The support of this work is provided by the National Science Foundation (NSF) (Grant number EAR 04-12859) and National Aeronautics and Space Administration (NASA) (Grant numbers NNG04GP78G). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflects the views of NASA or NSF.

REFERENCES

[1] P. Bajcy, P. Groves, S. Saha, T.J. Alumbaugh, and D. Tcheng, " A system for territorial partitioning based on GIS raster and vector data," Technical Report NCSA-ALG-03-0002, February 2003.

[2] K. Koperski, "A progressive refinement approach to spatial data mining,", Ph.D thesis, Simon Fraser University, pp. 175, April 1999.

[3] E. Mesrobian et al., "Mining geophysical data for knowledge," IEEE EXPERT, pp. 34-44, 1996.

[4] A.B. White, P. Kumar, and D. Tcheng, "A data mining approach for understanding topographic control on climate-induced inter-annual vegetation variability over the United States," Remote Sensing of Environment, 98, pp. 1-20, 2005.

[5] Ruddell, B.L. and P. Kumar, Modelshed Data Model, in *Hydroinformatics: Data Integrative Methods in Computation, Analysis and Modeling*, by Kumar et al., to appear in 2005.

[6] Kumar, P., J. Alameda, P. Bajcsy, M. Folk and M. Markus, *Hydroinformatics: Data Integrative Methods in Computation, Analysis and Modeling*, Authored Handbook, CRC Press, Manuscript Submitted to the Publisher, expected publication date November 2005.

Chapter 11

Linkages between the National Center for Hydrologic Synthesis (NCHS) and HIS

Norman L. Miller, Susan S. Hubbard, Deborah A. Agarwal
Lawrence Berkeley National Laboratory
Earth Science Division
Berkeley, CA

Abstract

The CUAHSI HydroView framework for tackling 21st Century water problems consists of several key facilities (Figure 1), one of which is the National Center for Hydrologic Synthesis (NCHS). Berkeley was recently chosen as the location for the NCHS, and plans are underway to create a vibrant 'hub' of national and international water-related activity at the Synthesis Center. Central to the success of the NCHS is to ensure that scientists working at the NCHS can access the data, models, analysis tools, compute resources, and storage facilities needed to address hydrological questions and to perform hydrological synthesis. Additionally critical to the success of NCHS will be the use of collaboratory tools which permit dynamic and interactive collaborations between the scientists located at NCHS and the other members of the community. The NCHS Computational Hydrology, data Assimilation and Infrastructure (CHAI) Working Group has begun to address these needs and to develop basic infrastructure and collaboratory tools for use by NCHS affiliates. Indeed, the NCHS working groups are actively using NCHS collaboratory tools to share ideas and documents and to track developments. The NCHS is actively working with private sector IT partners to develop more advanced IT capabilities at the NCHS, including advanced architecture needed to link the NCHS synthesis tools, compute resources and storage facilities together. As the Synthesis Center will bring together scientists involved in advancing hydrologic research with NCHS IT/computational partners, it offers an ideal forum to explore how these researchers can benefit from advanced technologies and effectively use collaboratory and other tools/datasets and modeling approaches for research and dissemination purposes.

In this chapter, we briefly review the concepts of the NCHS at Berkeley. We then describe the importance of computational hydrology and data assimilation to the center's mission, and discuss a few key components of that infrastructure, including suggestions toward the implementation of collaboratory and analysis tools at the NCHS.

1. Background: The National Center for Hydrology Synthesis

Three factors suggest that water resource management can no longer be based solely on engineering solutions as it has primarily been in the past. The first is the emergence of continental- and global-scale problems. The second is the recognition of the interconnectedness of nature and the changes being wrought by humans. And the third is the need for balancing between economical and political values, ecosystem requirements for water, and equitable sharing of water resources. To effectively manage our water resources, a new paradigm is necessary to integrate hydrologic science with mathematical, engineering, physical, life, information, and social sciences. Against this backdrop of scientific change and the needs of professionals, the NCHS will:

- Develop a new vision of hydrology and water resources management.
- Translate this vision into a focused but evolving research program.
- Encourage synthesis activities that integrate science, technology, and societal needs.
- Support collaborative research across disciplines, institutions, and sectors.
- Support the computational, modeling, and data needs of the synthesis activities.
- Design and demonstrate new methods and tools for hydrologic research through the integration of mathematical, engineering, physical, life, information and social sciences.
- Disseminate research results broadly to scientific and professional communities as well as educational institutions and public media outlets.

To meet these challenges, as demonstrated above, hydrologists will need to form unconventional and nontraditional research coalitions. The success of this Center will depend on the involvement of institutional partners and individual researchers from different methodological, technical, and organizational backgrounds. In addition to proactively seeking partners and researchers from across the different fields in the mathematical, engineering, physical, life, information, and social sciences, the Center will purposefully recruit partners and researchers from academic, governmental, non-governmental, and industrial   organizations alike. Only with such a diversified body of participants, representing the broadest spectrum of hydrologic interests - from basic to applied and from regulatory to public interest - will the Center have the potential of transforming the science of hydrology and its practical applications.

Bringing all these components under the same roof is an important step, but it alone is insufficient. What is additionally needed is access to data that can support research into the hypotheses of the new hydrology paradigm, and the modern tools needed to analyze them. The need to develop new comprehensive data bases was discussed in great detail by the National Research Council (NRC 2004, 1991) and U.S. Climate Change Research Program’s Water Cycle Study Group (USGCRP 2001): it requires unprecedented efforts due to the need to collect data over large spatial scales, over long periods of times, and over diverse environments. Such an effort requires careful planning, and is currently being undertaken by CUAHSI, under its HydroView Plan (Figure 1). This effort includes massive and coordinated data acquisition efforts through Hydrologic Observatories (HO) and Digital Watersheds (DW), development of the information technology needed to access, structure, store, display and disseminate this data through Hydrologic Information Systems (HIS), and development of a new generation of measurement technology through the Hydrologic Measurement Technology (HMT) program. With access to HO, HIS and HMT, the components needed for modern hydrologic synthesis are in place.

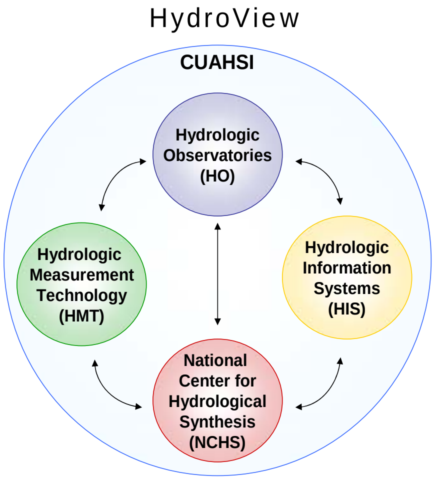
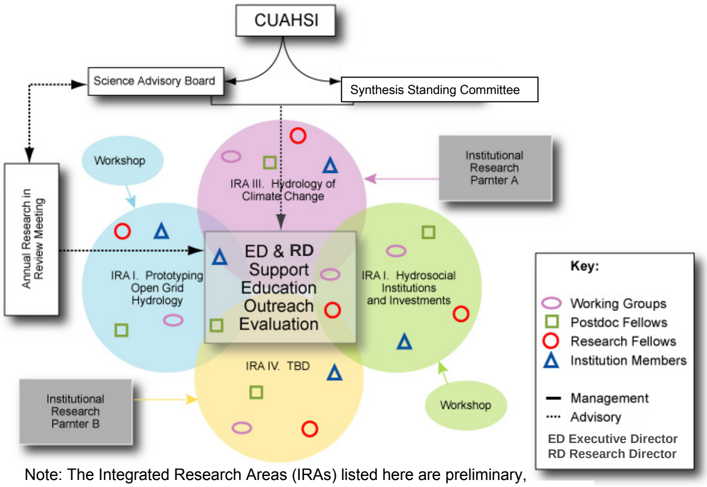

HydroView
CUAHSI
Hydrologic Observatories (HO)
Hydrologic Measurement Technology (HMT)
Hydrologic Information Systems (HIS)
National Center for Hydrological Synthesis (NCHS)

Figure 5 CUAHSI HydroView Research Framework

Organization and activities at the NCHS will be centered around Integrated Research Areas (IRAs). IRAs represent arenas of inquiry where cross-cutting collaboration and cooperation are likely to produce significant advances in hydrologic synthesis. Activities in each IRA will include three components: basic and applied research, development of enabling technology or tools, and the application of tools. The IRAs will be carried out by a combination of multi-stakeholder Working Groups, Visiting Research Fellows, Postdoctoral and Doctoral Fellows and Senior Research Personnel. Examples of IRA topics are given in Figure 2. The NCHS is currently soliciting input from the community about IRA topics – such input can be given via http://nchs.berkeley.edu/community_input.html.

Working Groups are an important component of the Synthesis center research and will consist of 5-15 individuals from different disciplines, institutions and sectors who have convened to work on a specific and mutually-defined problem, project, or product pertinent to the hydrology community within the broad constraints of an IRA. The Working Groups will primarily be selected via a semi-annual competitive proposal process, using the criteria of intellectual merit, scientific relevance, organizational fit, and methodological innovation. Initial Working Groups include those associated with Hydrologic Observatories, Instrumentation, Computational Hydrology, Computational Hydrology data, Assimilation and Infrastructure (CHAI), Global Water Science, Education and Outreach, Hydrologically Compatible Institutions, and Hydromorphology. A description of these initial, proposed Working Groups is given at http://nchs.berkeley.edu/working_groups.html.


CUAHSI
Science Advisory Board
Synthesis Standing Committee
Annual Research in Review Meeting
Workshop
IRA III. Hydrology of Climate Change
Institutional Research Parnter A
IRA I. Prototyping Open Grid Hydrology
ED & RD Support Education Outreach Evaluation
IRA I. Hydrosocial Institutions and Investments
IRA IV. TBD
Institutional Research Parnter B
Workshop
Key:
⬭ Working Groups
□ Postdoc Fellows
○ Research Fellows
△ Institution Members
— Management
···· Advisory
ED Executive Director
RD Research Director

Note: The Integrated Research Areas (IRAs) listed here are preliminary, and the distribution of personnel within the IRAs is schematic.

Figure 6 The organizational structure of the NCHS is keyed to Integrated Research Areas (IRAs).

Because multi-sectoral and multi-disciplinary input is critical to tackling large-scale synthesis in a holistic manner, the NCHS has partnered widely with various (non-academic) institutions. Our partnerships fall into four categories: institutional members, institutional research partners, computational/IT partners, and education & outreach partners. IT Partners will participate in development of and support the Center's IT platform with technical expertise, computational resources, software, and equipment. Examples of the NCHS IT partners include: The National Energy Research Supercomputer Center (NERSC) at LBNL, the San Diego Supercomputer Center at UCSD, Hewlett Packard (HP), Microsoft, IBM Almaden, and NASA. We are also planning to partner with NSF IT projects such as TeraGrid and GEON. Regardless of the category of affiliation, our partners are committed to participation in this national effort. Our partners will offer the NCHS perspective, expertise and varied types of support. We firmly believe that this support will be a crucial component for developing the infrastructure needed at the NCHS and for ensuring involvement and knowledge transfer between the NCHS, the hydrological community, outside organizations, policymakers, and stakeholders.

The NCHS has developed an extensive education and outreach program (including the Hydrology Leadership Summer Course for doctoral students and young investigators, the Undergraduate Hydrology Summer Camps, a web site www.H202U, and more), and a knowledge transfer program. The education and outreach program is intended to educate the public and train the leadership needed to sustain hydrology research momentum well into the future. More information about the National Center for Hydrology Synthesis is available at http://nchs.berkeley.edu/index.html.

2. NCHS Computational Infrastructure

Synthesis activities will require an advanced software and hardware infrastructure and software to transform hydrology into a field where various participants can: assimilate complex, multi-scale datasets; employ computationally intensive modeling; utilize knowledge discovery tools that permit hypothesis testing; combine different model and analytical tools, across heterogeneous computational platforms to tackle large-scale computational problems; perform workflow tracking, and archiving of results; perform visualizations; and effectively collaborate and disseminate ideas, results, and resources internationally. The tools and resources that form the synthesis computational infrastructure must allow (and even encourage) global interaction between water researchers and professionals, social scientists, policy makers and end-users throughout the world in an efficient and distributed fashion.

The NCHS will provide computing support to a wide range of users including: residents at the center such as postdocs, visiting scientists, and staff; university collaborators who are working with the center locally and remotely; collaborators in private industry; and members of IRAs and the associated Working Groups. In order to provide this support, a wide range of computing infrastructure will need to be provided at the NCHS. In this section we describe the NCHS plans for providing the basic underlying hardware and computational resources needed at the NCHS. We then describe the more advanced infrastructure needed at the NCHS and how we intend to achieve this advanced infrastructure through leveraging with contributions provided by our IT partners and in parallel with developments in other HydroView elements, such as HIS.

2.1 Basic Computational Infrastructure

The basic NCHS computational infrastructure required to support users will provide networking, desktop machines, web servers, videoconferencing facilities, collaboration tools, and computational resources. The desktops provided at the NCHS will be high-end PCs and Macs so that users can perform code development, routine work, and run small-scale computations. A highspeed networking link to the NCHS and between machines at the NCHS will be installed to ensure that users can gain access to data and results easily over the network. A moderate size computational server and file server will also be provided at the center to support smaller scale modeling and analysis efforts undertaken by collaborators involved in the NCHS. Commercial videoconferencing, collaboration tools, and web services, discussed further below, will also be provided at the NCHS.

Recognizing the importance of hydrologic synthesis, high-performance computing facilities such as the San Diego Supercomputer Center (SDSC) and the National Energy Research Scientific

Computing Center (NERSC), have graciously donated computational and storage resources to the NCHS. Through a partnership with Hewlett Packard Corporation, the NCHS has also secured access to the UC Berkeley Millennium Cluster. As the center evolves, the NCHS will also possibly purchase a moderate size cluster and visualization facility for use at the NCHS facility. These purchase decisions will be made based on needs assessed as the center's usage matures.

2.1.1 Data to Support Synthesis Activities

Synthesis activities at the NCHS will often require large amounts and different types of datasets. Datasets that will be used at the NCHS will include both hydrological datasets, as well as less conventional datasets, such as those associated with ecological, sociological, and economic analysis. A significant challenge researchers face will be locating all the data required, transferring it to the location where it will be used, and converting the data to a common format, reference framework (e.g. GIS), and time scale. The data will also need to be annotated with metadata if it is not already annotated by its source and converted to common units to allow use in models, comparisons, and analyses. Capturing and retaining data quality and data uncertainty indications will also be extremely important. CUAHSI affiliates and several of the NCHS's Working Groups will require easy and rapid access to data storage, and to data analysis and mining tools, which may be stored locally or available via high speed networks from HIS and other locations. The NCHS is implementing a web services-based approach to facilitate easy use of NCHS infrastructure and to access other datasets and tools. This will allow datasets which traditionally have not had any relationship in the past will be combined in new and innovative ways by synthesis researchers.

2.1.2 Collaborative Tools

Synthesis research at the NCHS will be carried out by Working Groups composed of cross-discipline teams of scientists, as was described in Section 1. These teams will require support for dynamic and interactive participation and collaboration capabilities between researchers and with stakeholders in a manner unprecedented in hydrological sciences. For example, a hydrologist who has performed a large, multi-terabyte simulation might want colleagues from other CUASHI centers and around the world to visualize the results in the same way and at the same time so that the group can discuss the results in real time using shared visualization walls provided by grid technologies. Similarly, it may be most efficient for NCHS Working Groups to meet 'virtually'. Thus, the collaborative tools will play a significant role in the development of a *virtual center*, which will allow for broad participation in the NCHS projects, and will facilitate the Working Group activities. For example, a critical success factor for the Working Groups will be their ability to work together to solve the difficult problems they will be tackling while meeting physically at the center, and to meet virtually when they are away from the center. As an example, consider researchers located at different institutions who are interested in using the HIS Neuse digital watershed to address hydrologic synthesis questions. Collaborative tools could facilitate the interaction of the researchers with each other, with the HIS, with site managers at the Neuse watershed, and potentially with decision makers and students. As such, we view collaborative tools as a critical capability to support the synthesis process.

Collaborative tools can be used to support the continuum of interaction from synchronous to asynchronous interactions. Synchronous interactions such as meetings can be supported by videoconferencing capabilities. There are several technologies available for videoconferencing such as H.323 (commercial), Virtual Rooms Videoconferencing System (free download), Conference XP (commercial - freeware), Webex (commercial), and Access Grid (open source) immersive conferencing facilities. If only audio is needed, tools such as voice over IP can provide audio conferencing among small groups of people. Shared whiteboards, shared presentations, and shared desktops are also very effective in supporting synchronous interactions. Semi-synchronous communication tools support both real time and time delayed communication. A capability classically used to support semi-synchronous communication is presence, chat, and instant messaging. Jabber (XMPP standards-based messaging tool) has open source and commercial tools that can provide secure instant messaging, chat, and presence information. Another technology typically used to support semi-synchronous interaction is seminar capture and broadcast. Some other key asynchronous collaboration tools include wiki workspaces, blogs, document sharing systems, and shared file spaces. The foundation of the shared space for such collaboration will be a wiki workspace for each Working Group provided by the center and a shared file space. The NCHS is committed to exploring the utility of collaboratory tools for use to support the NCHS Working Groups and where feasible and potentially useful, between the NCHS and partners. The NCHS has already implemented a NCHS wiki and a NCHS blog for use by the Working Groups and planning committees as they form. These working group wikis and blogs are already being actively used by the NCHS affiliates; examples of these are given at: http://nchs.berkeley.edu/blog/. These capabilities were configured by modifying widely available open source wiki and blog distributions. A Microsoft Sharepoint server will soon also be configured to support file sharing and shared calendaring needs of the Working Groups. The semi-synchronous communication between Working Group members and presence information will be supported by Jabber (XMPP standards-based messaging tool). In addition, videoconferencing capabilities, such as Access Grid technologies, that will allow remote participation in Working Group meetings, are available for NCHS researchers.

Although there are many products available to support collaboration between members of a distributed group, they are not all equally effective at supporting any *particular* scientific collaboration. Each collaborating group has its own needs and usage scenarios and an essential part of building a successful collaborative environment is understanding what activities need to be supported and what tools would best support these activities. In addition, the available equipment, operating systems, and support options need to be considered when choosing tools. The first step in designing a collaborative environment is to visit a representative set of the institutions and types of collaborators and develop an understanding of the usage scenarios for collaboration and the local constraints at these institutions. For example, a site that only uses linux would have difficulty using a collaborative tool that only runs on windows. The usage scenarios help recognize key activities and make sure that they are shared. The second step involves tool installation and participant training. This phase focuses on deploying the identified technologies, and on training the project participants to effectively use the tools. A final stage in the collaborative tool component involves ongoing evaluation and community training. This final phase consists of evaluating the usage of the deployed technologies and in evaluating new technologies as they evolve with the objective of ensuring that the optimal collaboratory tools are in place within the projects. An integral part of the collaboration environment will be measurement and assessment mechanisms, usage statistics can help measure effectiveness of the tools and suggest changes that might be needed in the tool set.

Another task will be performing more general training of the CUASHI community to familiarize them with the collaboratory aspects of the various projects.

2.2 Advanced NCHS CyberInfrastructure

The advanced NCHS computational infrastructure will consist of numerous tools that will enable researchers to easily access data from a variety of disciplines, perform model simulations and other types of analysis, and to collaborate. Our vision is that the Center's CyberInfrastructure will enable groups of researchers/scientists to meet together in a resource-rich environment and address cutting edge science questions that can only be solved by teams of interdisciplinary scientists meeting together through *face-to-face* discussions and *what if* computations. This vision will be realized by combining advanced grid computing, data management, visualization, and collaboration systems, with large cluster computing and remote storage capacity. The CHAI group, which is composed of hydrologists, computer science researchers, private computer science partner representatives, and representatives of key NSF projects building cyberinfrastructure, will help to identify and define methodologies to leverage existing technologies already available to help build the advanced computational infrastructure needed for synthesis. An extremely important component of the advanced infrastructure development at the NCHS is the involvement of the private sector NCHS IT partners, who will assist the NCHS by providing both expertise and financial leveraging for this endeavor. As an example, the NCHS has just been declared to be a Microsoft Technical Computing Imitative, and NCHS will work with NCHS to develop the architecture at the NCHS needed to bring in datasets from remote locations, perform synthesis at the NCHS using advanced tools and computational resources, and send the results elsewhere for archiving. Demonstrating the use of such distributed datasets and resources will be a first step in demonstrating the resources that the NCHS can provide for its researchers.

To tackle synthesis questions using the vast amounts of hydrological, socioeconomic, ecological, and other datasets, analysis and numerical modeling tools are needed that can simulate components of the hydrologic cycle and eventually couple the hydrological simulations with other (policy, socioeconomic) analysis. There are many numerical models where hydrological processes are described. The NCHS is currently considering what level of hydrological modeling support to provide to NCHS researchers. Should the NCHS install a basic suite of analysis tools and hydrological models as an initial IT buildup effort, or should these tools be added 'on demand' from various postdocs and working groups? If an initial suite of models is to be incorporated, what models should be included and what style of framework should be developed? What level of computational infrastructure and information technology (IT) staff are needed to support a modeling environment? The solutions to these questions will evolve over time as the NCHS matures and the needs of the NCHS researchers are crystallized. This philosophy of common framework capabilities and interfaces will quickly advance the NCHS CHAI mission to enable researchers to apply new dataflow and modeling techniques to address scientific hypothesis and form new synergies and synthesis. A discussion about various modeling approaches that could be considered at the NCHS over time is given in the Appendix.

3. Summary

The University of California at Berkeley was recently chosen to host the National Hydrology Synthesis Center (NCHS). The Center is expected to become the focal point for creating and disseminating the new hydrologic paradigm needed to overcome the modern scientific and societal challenges facing hydrologists. Both basic and advanced computational infrastructure will be necessary at the NCHS to enable access to data, conduct modeling, perform analyses, and enable results dissemination. The basic IT, including hardware, software, some plumbing, and collaboratory tools are currently being developed at the NCHS. The NCHS working groups are actively using NCHS wiki and blog tools to share ideas and documents, and to track work progress. The advanced IT infrastructure, including the architecture needed to access disseminated datasets, tools, and computational resources in a secure environment and the installation of advanced analysis tools, will be developed over time at the NCHS in partnership with the NCHS private sector IT partners. NCHS computational, collaboratory, and analysis tools should be integrated properly with the CLEANER, CLEO, and the CyberDashboard, once the dashboard is developed. The NCHS is interested in working with other elements of HydroView to effectively integrate their IT platform. We solicit input from the community about these concepts via an online questionnaire at: http://nchs.berkeley.edu/community_input.html.

Appendix

Background on Modeling Frameworks

In hydrology and related disciplines (e.g. climatology), modeling frameworks vary widely, from tightly coupled single community models to loosely coupled, multiple model frameworks. There are benefits and limitations to these disparate structures. Single modeling system approaches rely on a suite of tightly coupled models, such as the Community Climate System Model (Blackmon et al. 2000, 2001). The Community Climate System Model (CCSM) components include an Atmospheric General Circulation Model, Ocean General Circulation Model, Sea Ice Model, and a Land Surface Model. These model components are tightly coupled, with flux transfers from one model directly into another using matching time steps and spatial resolutions. This single system modeling approach has been successful at the NSF-supported National Center for Atmospheric Research (NCAR). Here the user community gives consensus on process level advancements through Working Groups that implement, test, and evaluate component models for the broader research community. Each Working Group is co-lead by a researcher internal to NCAR and a researcher from the external academic community. This approach includes an annual system modeling workshop that is open to all researchers and includes detailed presentations on model advances with time for feedback from individual researchers. The benefit of this approach is that a single modeling system may entrain users via the momentum behind such a community-based activity. In a sense, single community models are easier to use than multiple models, and modeling maintenance is relatively manageable. An easy-to-use and visible community model may bolster the number of researchers interested in hypothesis testing and advancing research questions through numerical modeling. A key limitation of this approach is that all modeling simulations depend on a given set of concepts and assumptions, which may or may not be appropriate to the specific problem and/or scales of interest.

At the other end of the spectrum are modeling frameworks or platforms that permit the incorporation of many different models, and facilitate the interaction of these different models through common standards and pre- and post-processors. Unlike the single community model approach, *loosely coupled* models within a framework can run independently, using different time steps and different spatial domains and resolutions. This approach allows for the output from one model (e.g. climate) to be used as input to another model (e.g. land surface hydrology) in a quasi or fully off-line mode. The advantage of this modeling framework is the flexibility of numerical investigations through the use of a wide variety of model types, and the ability to test and implement new models as they are developed. One of the challenges of this approach is the maintenance of multiple models, a more labor intensive activity than a single community model approach. Examples of multiple model frameworks include the Modular Modeling System (Leavesley et al. 1997) and the framework for intercomparing single-column atmospheric-radiation models (Sommerville et al. 1999). The conceptual framework for the Modular Modeling System (MMS) includes three major components: pre-processor, model, and post-processor. A system supervisor, in the form of an X-window graphical user interface (GUI), provides users access to all the components and features of MMS. The framework has been developed for UNIX-based workstations and uses X-windows and Motif for the GUI. The GUI provides an interactive environment for users to access model-component features, apply selected options, and graphically display simulation and analysis results. MMS was designed for the USGS Precipitation Runoff Modeling System (PRMS), and has been running TOPMODEL and a version of the Sacramento Soil Moisture Accounting (SAC-SMA) model. At present, it is not clear if it is sufficiently robust for handling a wide range of model types since its original design was and primarily still is for the application of PRMS. That is, MMS was originally written as a stand-alone rainfall-runoff code and later a GUI wrapper was added. Further refinements resulted in the combined MMS and its GUI interface becoming the USGS PRMS. Recent modifications to PRMS have made it possible to import other rainfall-runoff codes, but it is cumbersome and requires considerable support from the PRMS developers.

An example of a hybrid modeling framework is the Regional Climate System Model (Miller and Kim 1997a, b). This hybrid system model has model components that are tightly coupled (mesoscale atmosphere model and land surface soil-plant-snow model) as well as loosely coupled (catchment-scale distributed hydrology model, groundwater model, sediment and water quality model, agro-economic model, crop models). The tightly coupled model components share the same spatial resolution and domain, iterate on a common set of time steps, and pass updated variables and flux information bi-directionally between the two models per time step. The loosely coupled model components do not pass variable and flux information, and typically run as stand-alone with input forcing provided by the tightly coupled models, and other input data from the pre-processors.. At the end of these simulations, output files are automatically transferred via script files to the post-processors for analysis. Post-processing, in this example, includes climate, weather, and streamflow predictions, impacts assessments, and statistical analyses.

A more advanced and generalized framework approach for model components and stand-alone model is the Earth System Model Framework (Hill et al. 2004). The ESMF is designed to run on high performance computing platforms, it can handle a large complex system model, or it can be run in a mode with stand-alone models. The ESMF software infrastructure provides ease of use, performance portability, interoperability, and reuse in climate, numerical weather prediction, data assimilation, and other Earth science applications. It defines architecture for composing multi-component applications and includes data structures and utilities for developing model components. The ESMF allows diverse scientific groups to leverage common software to solve routine computational problems such as efficient data communication, model component coupling and sequencing, time management, and parameter specification. This community-developed (NSF, NASA, DOE supported) framework is intended for Earth science research and has extensible capabilities for implementation of hydrologic models. This framework is used not only for Global Climate System Modeling, but for the Land Information System, a NASA-led consortium that is investigating a range of land surface models and generating daily simulations.

To maintain flexibility in investigator-based computational hydrology during the early stage of development, it will be important to keep the conceptual modeling framework of the NCHS simple and open, permitting ease of use to a wide hydrological community. For compatibility and maintenance, a multi-model framework will require some common standards, including computer platform specifications and data input/output formats. Of key importance is that researchers can implement their own models or access existing models at NCHS with minimal effort, and that data input/output, formats, and linkages between models is manageable. Figure 3 provides a schematic of a simple configuration for a possible NCHS framework, which may be developed over time at the NCHS, as indicated by NCHS working group, postdoc, and visiting scientist needs needs.

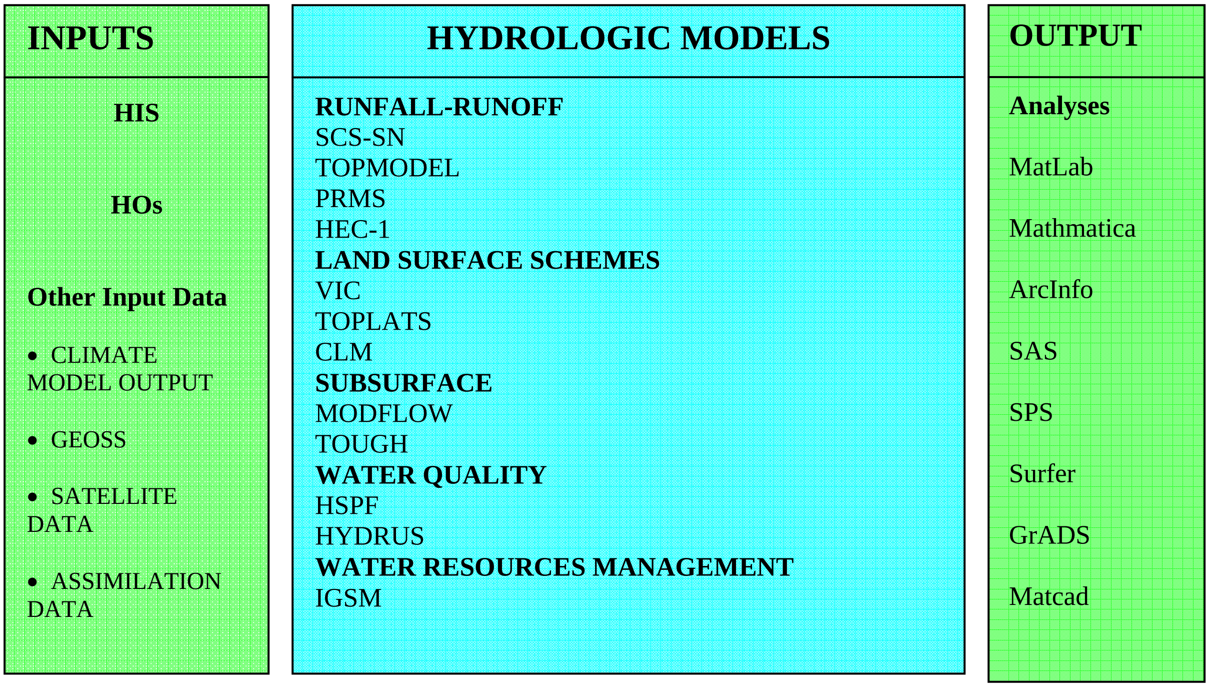


Figure 3. Schematic representation of an example modeling framework shows example inputs, models, and output components. The input includes HIS, HOs, and other data sources. An archive of hydrologic models may be available through simple links with each other, and output with analysis tools. The models shown are only examples of the different types of models that could be included in such a framework.

It is important that researchers have early access to modeling and related tools that are readily available and easy to use with minimal assistance from NCHS staff. It is also important that these models can be modified or used to perform more complicated tasks. Examples of model categories with some hypothetical models that could be included are provided in figure 3. A community poll, conducted by HIS, has indicated that some of the key observatory modeling service requests include hydrologic rainfall-runoff modeling; models for the main rivers and streams; a water quality model for the main water quality constituents in streams and rivers; groundwater modeling for main aquifers, a regional climate model for weather and climate simulations for observatories. A list of models, operating systems, and analysis packages that are being considered for inclusion is at http://esd.lbl.gov/NCHS/comp_infrastructure/questionnaire.html. A brief discussion on the different categories of models, as well as examples of the models, is given below.

Rainfall Runoff Models. A range of rainfall-runoff model complexity can be captured using three different model types: empirical fits, low parameter, and high parameter. Examples of such models include the Soil Conservation Service Curve Number (SCS-CN) model (Soil Conservation Service 1972), TOPMODEL (Beven et al. 1994) and the U.S. Army Corps of Engineering's Hydrologic Engineering Center - 1 (HEC-1) model (USACE 1981). The SCS-CN consists of a nonlinear relationship between rainfall and the curve number (CN), which indicates the potential maximum retention after rainfall begins. Additional methods, such as the probabilistic approach, also fall under this category. TOPMODEL is a low parameter, rainfall-runoff model that includes physical processes describing infiltration, excess overland flow, and streamflow routing. Lateral flow is a function of a topographic index that can be described as a distribution or explicitly. HEC-1 is a spatially lumped network model that based on sub-basin inputs (precipitation, temperature), generates a direct surface runoff hydrograph and different options for modeling rainfall, losses, unit hydrographs, and stream routing. One of the advantages of HEC-1 is its compatibility with the other HEC codes. Fully distributed rainfall-runoff models are in the next (land surface) model category.

Land Surface Models. To simulate the surface water and energy budgets at the HO watershed scale (i.e. 10,000 km^2), semi- and fully-distributed land surface models will be needed. Three land surface models, the Variable Infiltration Capacity (VIC) model, the Community Land Model (CLM), and the TOPMODEL-based Land Atmosphere Transport Scheme (TOPLATS), represent three such models, each with different features. VIC is a macroscale hydrologic model that solves full water and energy balances, (Liang et al. 1994). It is most appropriate at spatial scales of 10 km and coarser, and includes a statistical formulation for infiltration capacity. TOPLATS (Famiglitti 1992; Famiglitti and Wood 1994) incorporates a TOPMODEL framework to account for lateral redistribution of subsurface water based on the local topography and soil transmissivity. It is appropriate at scales ranging from 30m to greater than 10 km. The statistical version uses the TOPMODEL index to describe the spatial variability of topography and land cover types. The distributed version solves the water and energy balance for each pixel. The NCAR Community Land Model (CLM) includes ecohydrology and biogeochemistry processes, and is more comprehensive than VIC and TOPLATS. CLM has vegetation dynamics, making it a unique and data intensive model. It includes TOPMODEL assumptions and has an advanced snow scheme. CLM was originally developed for global models, but has been successfully used at scales ranging from 30m to 10 km (Jin and Miller 2005).

Subsurface Models. Several vadose zone/groundwater flow models exist that could be incorporated into the modeling framework. Perhaps the most well-known subsurface code is MODFLOW. MODFLOW (USGS 1984) is a three-dimensional finite-difference ground-water model that was developed by the USGS and is currently one of the most widely used groundwater modeling packages. It has a friendly GUI and a modular structure that permits easy updating of new versions and can be applied at scales ranging from less than 1m to greater than 1km horizontally. *The NSF SAHRA has developed several packages to MODFLOW; SEEPAGE, DRAIN, and a riparian evapotranspiration package (Madock, 1996)*. The TOUGH2 family of codes (Pruess et al. 1997) represents a general-purpose numerical program for multi-phase fluid and heat flow in porous and fractured media. It can simulate a wide range of complex phenomena, including inverse modeling, multi-component, multi-phase chemical reactions, and coupled biogeochemical-hydrological processes. TOUGH2 codes typically are applied at very fine scale horizontal resolutions (i.e. cm to m). It does not have a GUI interface, but offers a wide range of flexibility and capability.

Water Quality Models.The Hydrologic Simulation Code-Fortran (HSPF), and its more recent C version, simulates land and soil contaminant runoff processes and is a widely used code, developed initially for the EPA. HSPF can be applied at spatial resolutions ranging from 10m to about 1km. It computes the hydrograph, upper and lower zone infiltration, and sediment transport. HSPF is designed to run on PC systems and has built in calibration procedures. The NSF SAHRA HYDRUS program is a Microsoft Windows based, finite element model used for simulating one-dimensional movement of water, heat, and multiple solutes in variably saturated media. The solute transport equations consider advective-dispersive transport in the liquid phase, and diffusion in the gaseous phase.

Water Resources Management Models. IGSM-II (Kadir et al. 2004) is a water resources management and planning model that simulates groundwater, surface water, groundwater-surface water interaction, as well as other components of the hydrologic system. It simulates groundwater elevations of a multi-layer aquifer with conservation, it simulates water demand as a function of different land use and crop types, and can compare it to historical or projected water supply. IGSM-II can allow the user to specify stream diversion and pumping locations for water supply, along with irrigation and urban water withdrawals.

We emphasize that we are not proposing to embed the models described above nor endorsing any particular model. The discussion given above is presented to show the range of different types of models that might be necessary at the NCHS to permit NCHS researchers to perform hydrological synthesis, and to illustrate the range of choices available within each category.

References

Blackmon, M. B., B. Boville, F. Bryan, R. Dickinson, P. Gent, J. Kiehl, R. Moritz, D. Randall, J. Shukla, S. Solomon, G. Bonan, S. Doney, I. Fung, J. Hack, E. Hunke, J. Hurrell, et al., 2001: The Community Climate System Model. *BAMS*, 82, 2357-2376.

Blackmon, M., B. Boville, F. Bryan, R. Dickinson, P. Gent, J. Kiehl, R. Moritz, D. Randall, J. Shukla, S. Solomon, J. Fein, and CCSM Working Group Co-Chairs, cited 2000: Community Climate System Model Plan 2000-2005. [Available on-line from http://www.ccsm.ucar.edu/management/plan2000/index.html.]

Chahine, M.T. 1992. The hydrological cycle and its influence on climate. *Nature*, 359: 373-380.

CUAHSI, 2004, a National Center for Hydrologuc Synthesis: Scientific Objectives, Structures, and Implementation, a report of a CUAHSI workshop held in Santa Barbara, California, July 10-12, 2003.

DeMenocal, P.B., Lamont-Doherty Earth Observatory, in Carey, J., 2004, Global Warming. A Special Report, BusinessWeek, August 16.

James, L.D., 1995, NSF research in hydrologic sciences, J. of Hydology, 172:3-14

NRC, 1991, Opportunities in the hydrologic sciences, Washington, D.C., National Academy Press. Washington DC. 348 pp.

Leavesley, G.H., 1997, The Modular Modeling System (MMS) - A modeling framework for multidisciplinary research and operational applications, Chap 8 of Freeman, G.E., and Frazier, A.G., eds., Proceedings of the Scientific Assessment and Strategy Team workshop on hydrology, ecology, and hydraulics, v. 5 of Kelmelis, J.A., ed., Science for floodplain management into the 21st century: Washington, D.C., U.S. Government Printing Office.

Miller, N.L. and J. Kim, J. Duan, 1997: The Regional Climate System Model: Southwestern United States and Eastern Asia studies. WMO/CAS/JSC *World Climate Res Prog,*. **25**, WMO/TD-No. 792.

Miller, N.L. and J, Kim 1997: The Regional Climate System Model. In *Mission Earth: Modeling* *and Simulation for a Sustainable Global System*, Ed. M.G. Clymer and C. Mechoso, *Soc. Comp. Sim. Inter*. 55-60.

NRC, 1991, Opportunities in the hydrologic sciences, Washington, D.C., National Academy Press. Washington DC. 348 pp.

NRC, 2004, Confronting the Nation's Water Problems: The Role of Research, Washington, D.C., National Academy Press.

Preuss, K., S. Finsterle, G. Moridis, C. Oldenburg, Curt. Y-S Wu, 1997: General-purpose reservoir simulators: The TOUGH2 family, Geothermal Resources Council Bulletin, 26, 53-57.

Somerville, R. C. J., and S. F. Iacobellis, 1999: Single-column models, ARM observations, and GCM cloud-radiation schemes. *Phys. Chem. Earth (B)*, **24**, 733-740.

USGCRP 2001: *Plan for a New Science Initiative on the Global Water Cycle*. Water Cycle Study Group. Chair: G. Hornberger. U.S. Global Change Research Program, Washington, DC.