CUAHSI Hydrologic Information Systems

Richard P. Hooper, David R. Maidment, John Helly,

Praveen Kumar, and Michael Piasecki

1. Introduction

The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) is in the process of developing an information infrastructure for the support of the hydrologic science throughout the hydrologic community.This system will be based on a hydrologic data model integrating data from many water disciplines, as well as many related disciplines such as ecology and solid earth science, tracing the flow of water and transport of constituents throughout the landscape, and integrating data across a range of scales of space and time. The resulting system will provide an innovative platform for asking an entirely new class of hydrologic science questions concerning the holistic functioning of the hydrologic cycle.

This proposal is part of a larger effort at NSF to advance cyberinfrastructure in science and engineering research. Cyberinfrastructure is built on a connected system of computers and information linked through communications technology. 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.“ This report recommends that NSF establish a new initiative for an Advanced Cyberinfrastructure Program. 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. A current NSF Information Technology Research project called GEON, the Geosciences Network, is a collaboration among information technology and earth science researchers to build cyberinfrastructure for the Geosciences. GEON is centered at the San Diego Supercomputer Center.

Hydrologic Science is part of the NSF Geosciences Directorate, and this proposal, like GEON, envisages collaboration between information technology professionals and CUAHSI hydrologic scientists to develop cyberinfrastructure for the hydrologic sciences, using the San Diego Supercomputer Center as a technology partner. By sharing technology development costs with other geoscience communities, CUAHSI will be help to advance the overall goal of environmental cyberinfrastructure while also building its own hydrologic information system. Because water is a common linking element among the atmosphere, oceans and the solid earth, hydrologic science is a good vehicle for advancing interdisciplinary understanding among the geosciences (National Research Council, 2001). Hydrology also has wider links to biological and ecological science and thus is a good vehicle for advancing understanding of environmental systems.

2. Hydrologic Information Systems

User Needs:

It is evident that there is a push throughout science and engineering research to make better use of information technology (NSF

Advisory Panel on Cyberinfrastructure, 2003). The Hydrologic Information Systems (HIS) component of CUAHSI is a logical response to meet the specific information technology needs of the hydrology community. A hydrologic information system consists of a hydrologic information database coupled with tools for acquiring data to fill the database and tools for analyzing, visualizing and modeling the data contained within it. No such comprehensive hydrologic information system presently exists, and the central task of the CUAHSI HIS component is to design and develop this system, implement it in universities across the United States, and support hydrologic scientists in its use. To provide coherence within the overall CUAHSI program, the first applications of the CUAHSI HIS will be the regions selected for evaluation as CUAHSI hydrologic observatories. By building a hydrologic data infrastructure for each of these regions by using pre-existing data as well as data that will be collected in the future, a series of —Digital Watersheds will be created with a capacity to compare conditions in different hydrologic regimes in the nation.

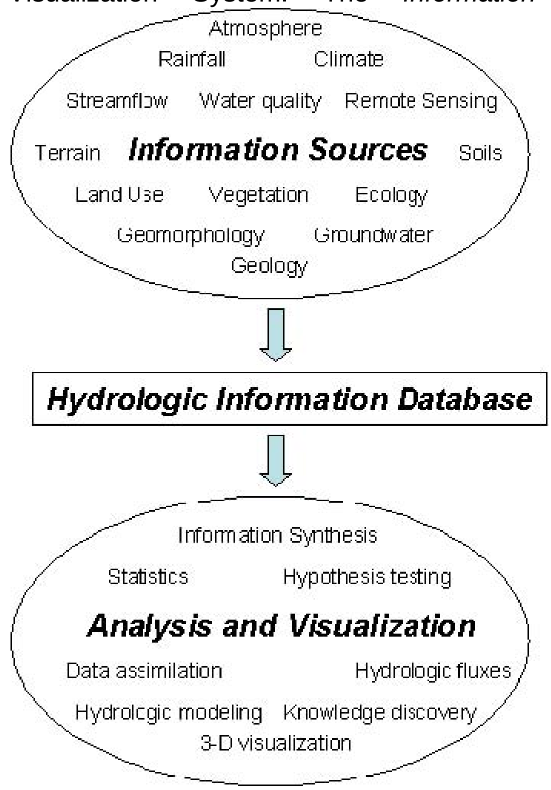
To help decide how a hydrologic information system could best support the hydrologic science community, a survey of the need for hydrologic data, information systems infrastructure, and services was conducted among the participants at the CUAHSI regional workshops in the spring of 2002. The results of this survey (complete responses to which can be viewed on the CUAHSI HIS website, http://www.iihr.uiowa.edu/~cuahsi/his/) reveal a great diversity among hydrologic scientists in the manner in which they use hydrologic information in their research. Nonetheless, some clear common needs are apparent. Among these are:

• better access to a large volume and variety of high quality data by means of the internet;
• better access to visualization tools and data analysis software to inspect and assess data; and
• access to standardized datasets for hydrologic fluxes and state variables across the United

The CUAHSI HIS will support these needs articulated by the community.

Hydrologic Information Database

The CUAHSI HIS will consist of three components as shown in the figure: Information Sources, a Hydrologic Information Database, and an Analysis and Visualization System. The *Information*



Sources component comprises all the information that a hydrologist needs to use to describe a Digital Watershed, including atmospheric conditions, rainfall, climate, remote sensing, terrain, land use, geomorphology, streamflow, water quality, ecology, vegetation, soils, geology, and groundwater conditions. The *Hydrologic Information Database* is a repository within which data from the information sources is loaded to describe a particular watershed, and a framework within which benchmark data sets describing particular hydrologic fluxes and state variables are stored. The *Analysis and Visualization* system is used by hydrologic scientists to reduce this vast volume of information to meaningful summaries, to perform statistical analysis and hypothesis testing, to create hydrologic models and assimilate data into them on current and past hydrologic conditions, to quantify hydrologic fluxes, and to support knowledge discovery, data mining and other innovative analysis techniques. The visualization component of this system will produce maps and graphs and other visual products that help hydrologic scientists interpret and understand the results of analysis.

Hydrologic information is routinely collected and provided by federal government agencies, such as the USGS, EPA, NASA, USDA and NWS, and by an array of state and local agencies. Individual hydrologic scientists and hydrologic field campaigns have created data sets for particular subjects and regions of the nation. This collection of hydrologic information sources is too vast to list individually in this proposal.

One of the critical challenges in building the CUAHSI HIS is that the information sources are geographically dispersed, the data comes in many formats, some unfamiliar to hydrologists, and the data files may be huge so the task of assembling, converting and synthesizing this information is so large that perhaps 80% of the effort in a study may be devoted to assembling the data and 20% to analyzing them. Compromises in the quality of the data and analysis frequently are made because the task of doing a better job is daunting to a scientist, and overwhelming to a graduate student. The intent in building the CUAHSI HIS is to simplify and streamline this process so that more emphasis can be placed on analyzing data and testing hypotheses and less on the sheer drudgery of data assembly and organization.

Hydrologic Data Model

For the hydrologic information database to function as contemplated, it must be carefully designed. A Hydrologic data model is an abstract database design developed by using a computer aided software tool which is then physically implemented in whatever database system is being locally employed (Jacobson, et al., 1999). Such systems include relational databases such as MS Access, SQL/Server or Oracle, in Extended Markup Language (XML) format as text files, or in binary files in a UNIX or Linux based file server. All tools developed for data acquisition, analysis, and visualization depend on the way the data is structured in the database. Special interfaces to the database are needed for each information source and analysis task that defines how data is input or acquired from the database. One of the critical early tasks in this project is the design of this data model.

Given the needs of integrative hydrologic science and the nature of the hydrologic data sources listed, a hydrologic data model needs to be able to meet the following needs for integrative hydrologic science:

- Represent hydrologic information in each phase of the hydrologic cycle (atmospheric water, soil water, surface water, groundwater);
- Be capable of tracking the movement of water and the transport of constituents across interfaces between phases of the hydrologic cycle, and accounting for the storage of water and transported constituents in each phase (McKay, 1991)
- Incorporate biological data over various temporal and spatial scales (Eagleson, 2002)
- Allow for the construction of a vertical water balance between atmospheric water, soil and surface water, and groundwater (Wilson and Gallant, 2002; Pinder, 2002)
- Allow for describing the horizontal flow of water through elements of the landscape, such as watershed to stream, to river or lake, to estuary, to bay, and finally to the ocean (Abbott and Minns, 1998; Naiman and Bilby, 1998)
- Be capable of integrating geospatial and temporal information;
- Be able to deal with both continuous and discrete spatial representations of data and of the transformations of data from one spatial framework to another;
- Be able to link data from different spatial scales, such as detailed descriptions of channel morphology on small stream reaches referenced to their correct location within the stream network of the basin (Gupta et al., 1986; Quattrochi and Goodchild, 1997)
- Be able to deal with data of any desired time scale and manipulations to transform data from one time interval to another (e.g. accumulation of daily data to

monthly data, monthly data to annual data, etc);
- Support the statistical description of error or uncertainty in the information; and
- Have a metadata component so that the lineage of how the data were produced can be traced.

Although the task of defining a hydrologic data model seems formidable, significant progress has been made on some components of this model (Babovic and Larsen, 1998; Wright and Bartlett, 2000). The many kinds of data stored in a hydrologic information database can be classified into four categories:
- *Monitoring data*, such as that coming from the USGS National Water Information System;
- *Geospatial data*, such as river and stream networks, watershed and aquifer boundaries, soil and land use maps, borehole and gage locations œ a data model for this information is presented by Maidment (2002);
- *Gridded space-time data*, including Nexrad radar rainfall maps, grids of weather and climate information, remote sensing images, and terrain maps œ these data are normally stored in specially formatted binary file formats such as netCDF (Vieux, 2001).
- *Statistical data*, such as the statistical parameters describing a data set, or the percentile values specifying its cumulative distribution function œ these are normally developed using an analysis program such as SAS or S+, but there is utility in having these statistical descriptors attached to the primary data so that the uncertainty and expected range of variability of the data can be quantified (Cressie, 1991; Hosking and Wallis, 1997).

Production of data in other file formats such as spreadsheet, text, Mathematica, and MathCAD needs also to be considered in the data model design. It is evident that there is no single file format or data structure that will satisfy all needs, and mechanisms for converting data between file formats will have to be built or acquired from other data systems.

The Hydrologic data model and the verbal description of its various components are critical elements of the intellectual infrastructure of CUAHSI HIS because they provide the language for describing hydrologic systems, and higher level concepts can be constructed for analyzing such systems once the language for describing the foundation has been established. In particular, it is necessary in data model design to be concerned not just with file formats, data conversions, and the like, but also with understanding how these data are going to fit into an analysis framework and how mathematical models using equations describing physical processes can be constructed to use the data. In other words, a conceptual framework is needed that connects the data describing the hydrologic environment with the physical laws that govern how water moves and its properties change in time and space (Gray et al., 1993). This is a challenging task, but a critical one since it forms the foundation for advancements in hydrologic knowledge built using hydrologic information systems.

Since all the tools that the CUAHSI HIS develops are dependant on the structure of the data that underlies them, the design and implementation of the Hydrologic data model is an important task for CUAHSI HIS that needs to be undertaken in the first year of the proposed project, and will be refined and extended later as needs dictate.

Hydrologic Meta-Data

CUAHSI HIS will contain an automated data harvesting system for data derived from diverse geographic sources. For this system to function, each data sources must be sufficiently well described with metadata that a user can identify a data set of interest. Hence, metadata descriptions must contain information that can be queried to answer when the data was collected, by whom, where, spanning what period in time, and a unique descriptor of the data variable. Because data may be edited and changed by the entity that collected it, a versioning system is required to identify such changes to the analyst. Secondly, metadata elements must contain information about how the data is stored, for example in ASCII or binary format (digital), or in paper form, or on tape only, and what storage format has been used (for example HDF, or GRIB, netCDF, etc). Third, metadata must contain information about the purpose for which this data set has been collected to give the user an idea in what context this data set exists. In addition, metadata should provide information about issues such as data quality assurance, calibration values for the measurement device, the sampling interval, zero or reference levels for the data, and pointers that direct the user to additional information about the data set (for example, a report associated with a data collection campaign) and the physical location of the data itself.

Several Metadata standards have been developed during the past few years of which the most important ones are: International Standards Organization Technical Committee 211, 19115 norm, the Dublin Core Metadata Initiative, the Federal Geographic Data Committee, the Global Change Master Directory, and the EOSDIS Core System. Currently, the various metadata standards are converging towards compliance with the ISO 19115 norm. Defining and implementing appropriate meta-data standards is important for CUAHSI HIS development.

3. Summary

CUAHSI HIS will advance integrative hydrologic science through the development of a hydrologic information system that can be implemented at universities throughout the United States. It involves collaboration between hydrologic scientists from the Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) and computer scientists from the San Diego Supercomputer Center, and supports a larger strategy at NSF to develop cyberinfrastructure to advance interdisciplinary study of environmental systems. The CUAHSI Hydrologic Information System will be built around a hydrologic data model that synthesizes data from diverse sources describing the water environment, and will support knowledge discovery through its own analysis and visualization tools and links to external modeling and software systems. The Hydrologic Information System will be built as a network so that data sources from many geographic locations can be automatically harvested and imbedded within a common data framework at the user's location. In this project, a prototype system will be developed with the San Diego Supercomputer Center as the hub, and the experience thus gained used to define a request for proposals for a CUAHSI Center for Hydrologic Information which will be responsible for the long-term maintenance and support of this system.

Integrative hydrologic science will be addressed through a series of CUAHSI science driver questions posed first for a hydrologic observatory planning study of the Neuse River Basin, whose data will be synthesized into the CUAHSI hydrologic data model as a prototype Digital Watershed. Similar Digital Watersheds will be constructed using later CUAHSI planning studies for hydrologic observatories elsewhere in the nation, thus providing a comprehensive information base for comparative studies across different hydrologic regimes. These datasets will be made freely available through an internet interface, and the hydrologic data model will also be provided to hydrologic scientists who want to develop Digital Watersheds at other locations.

References

Abbott, M.B., and A.W. Minns, 1998, Computational hydraulics (2nd Ed.), Ashgate Press, Aldershot, 557p.

Babovic, V., and L.C. Larsen, (ed), 1998, Hydroinformatics 98, (2 Vols), Balkema, Rotterdam, 1500p.

Cressie, N., 1991, Statistics for spatial data, Wiley, New York, 900p.

Eagleson, P.S., Ecohydrology, 2002, Darwinian expression of vegetation form and function, Cambridge University Press, 443p.

Gray, W.G., A. Leijnse, R.L. Kollar, C.A. Blain, 1993, Mathematical tools for changing spatial scales in the analysis of physical systems, CRC Press, Boca Raton, FL, 232p.

Gupta, V.K., I. Rodriguez-Iturbe, and E.F. Wood, 1986, Scale problems in hydrology, Reidel Publishing Co., Dordrecht, 246p.

Hosking, J.R.M, and J.R. Wallis, 1997, Regional frequency analysis, Cambridge University Press, 224p.

Jacobson, I., G. Booch and J. Rumbaugh, 1999, The unified software development process, Addison Wesley, 463p.

Maidment, D.R. (ed.), 2002, —Arc Hydro: GIS for Water Resources", ESRI Press, Redlands CA.

McKay, D., 1991, Multimedia Environmental Models, the Fugacity Approach, Lewis Publishers, Boca Raton, Fl, 257p.

Naiman, R.J. and R.E. Bilby, 1998, (ed.) River Ecology and Management, Springer, New York, 705p.

National Research Council, 2001, Basic research opportunities in the earth sciences, Committee on Basic Research Opportunities in the Earth Sciences, National Academy Press, Washington DC, 154p.

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.

Pinder, G.F., 2002, Groundwater modeling using geographical information systems, Wiley, New York, 231p.

Quattrochi, D.A., and M.F. Goodchild, 1997, Scale and remote sensing in GIS, Lewis Publishers, Boca Raton, Fl,

Vieux, B.E., 2001, Distributed hydrologic modeling using GIS, Kluwer Academic Publishers, Dordrecht, 293p.

Wilson, J.P., and J.C. Gallant, 2000, Terrain analysis, principles and applications, Wiley, New York, 479p.

Wright, D., and D. Bartlett, 2000, Marine and coastal geographical information systems, Taylor and Francis, London, 320p.

Richard P. Hooper, CUAHSI, 2000 Florida Avenue N.W., Washington D.C. 2009.

David Maidment, Department of Civil Engineering, The University of Texas at Austin, Austin, Texas, 78712

John Helly, San Diego Supercomputer Center, University of California San Diego, La Jolla, California, 92093

Praveen Kumar, Department of Civil and Environmental Engineering, University of Illinois, Urbana, Illinois, 61801

Michael Piasecki, Department of Civil and Architectural Engineering, Drexel University, Philadelphia, Pennsylvania, 19104.