Hydrologic Information System
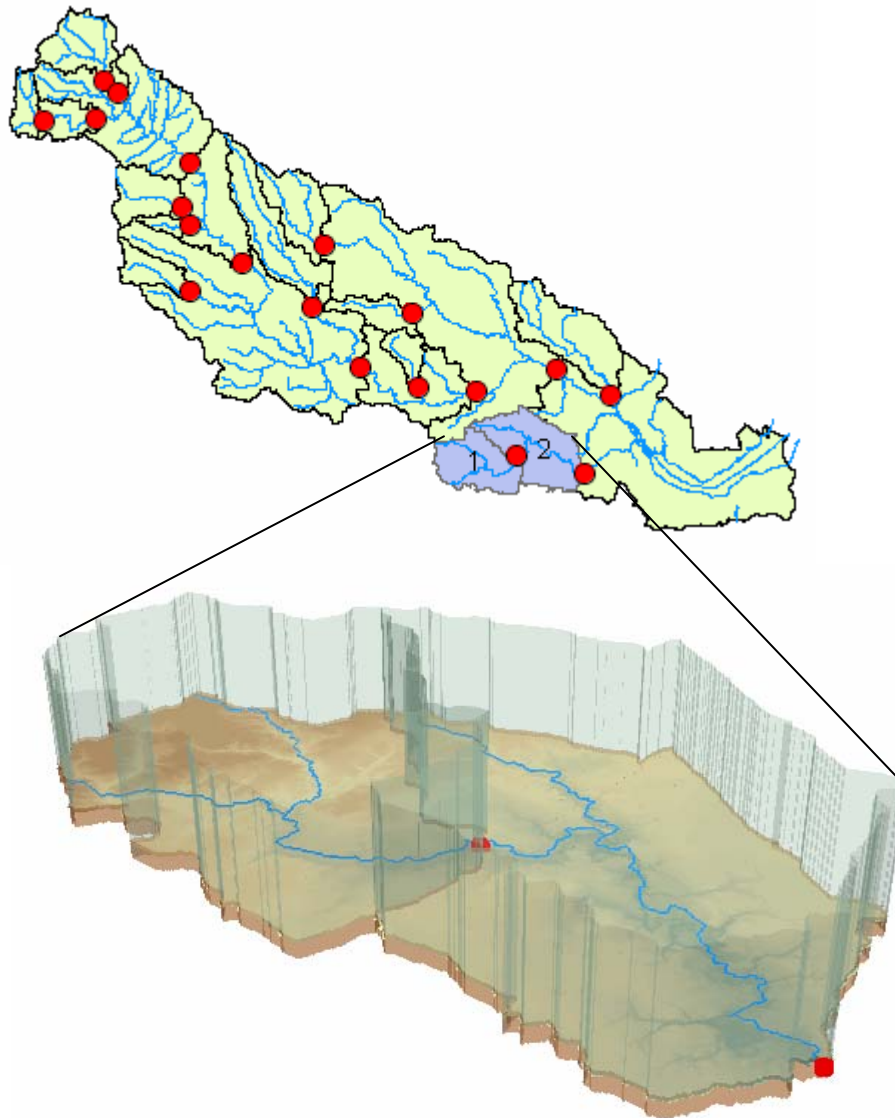# Status Report

Version 1: September 15, 2005
Edited by David R. Maidment
Full report is available at:
http://www.cuahsi.org/docs/HISStatusSept15.pdf

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.