

Information management needs of hydrologic observatory test beds. Survey at the CUAHSI HIS – test bed projects joint meeting, Austin, TX, November 15-17, 2006. Preliminary Analysis.

One of the goals of the Austin meeting was to understand information management needs of the test bed projects so that the tasks and priorities in the Phase II CUAHSI Hydrologic Information System project could be adjusted to meet these needs. The questionnaires were distributed in electronic form to test bed PIs, and completed within a two week period before and after the workshop. The original questionnaires are available at the CUAHSI HIS team system.

The questionnaires were filled out mostly by project PIs (9 out of 11). The range of funding for the two-year projects is 200-400K. The general conclusion is that there is a significant diversity among projects, models and variables they use, and their perception of interaction with CUAHSI HIS. There are several test bed projects where CUAHSI HIS technologies can be immediately useful, and projects where CUAHSI HIS approaches would satisfy a fraction of the requirements.

I. Common research and data management practices and obstacles

The first group of questions focused on the common research and management practices in each project, and the most important obstacles the projects are facing. These are preliminary estimates made at the time when the projects are commencing, and thus reflect the general set of obstacles the research groups are dealing with.

On question *What are your most time consuming research activities?* , the responses ranked as following:

collection of data	9 out of 11 respondents marked this as a time-consuming activity
data reduction	8 out of 11
sensor development	7 out of 11
writing papers and proposals	6 out of 11
semantic data integration	6 out of 11
coding numerical models or simulations	6 out of 11
running numerical models or simulations	5 out of 11
data visualization	5 out of 11
integration of different data formats	4 out of 11
theoretical work	4 out of 11
training and education	4 out of 11
spatial and temporal adjustments	3 out of 11
locating data sources	3 out of 11
system administration	3 out of 11

The results are generally consistent with the WATERS survey where a similar question was used, and with the survey of CUAHSI users. As before, data issues came out on top as the most time-consuming. Emphasis on sensor development and numerical modeling are specific for this group of respondents.

The obstacles (“*In your experience, what are the most important obstacles to using data from different sources?*”) are ranked as follows.

Inconsistent data formats	8 out of 11 respondents marked this as an important obstacle
Lack of metadata	6/11
Inconsistency of metadata	6/11
Inconsistent spatial extent	6/11
Unknown or inconsistent units	6/11
Irregular or different time steps	6/11
Large size of data	5/11
Lack of software for data integration	5/11
Lack of linking/workflow middleware	5/11
Streaming data	4/11
Lack of software to scale up with data	3/11
Data contributor is unknown	4/11

Again, consistent with the previous question and with earlier surveys, data acquisition and interpretation issues top the agenda.

The majority of respondents rely on common computing resources for research tasks (“*What computational resources do you use most frequently for analysis?*”). Only 2 of the test beds indicated they use computationally powerful servers, while 16 responses pointed to desktop (10) and laptop (6) use. These resources are used to compute a wide variety of models. Responses to the question “*What environmental or hydrologic models do you use most in your test bed research?*” are distributed as in the table below (this question was answered by 9 test beds out of 11):

MODFLOW	Mentioned 2 times
PARFLOW	1
SWMM	1
HSPF	1
RHESsys	2
SWAT	1
HEC-RAS	2
QUALZK	1
Code developed in house	2
SPARROW	1
CH3D	1
UnTRIM	1

PIHM	1
PRISM	1
SEQUAL, TBD	1
Penn State Integrated Model, Hydrology Lab Rainfall-Runoff Modeling System, Chesapeake Bay HSPF model	1
We have been developing large-eddy simulation (LES) models	1
Bayesian network models	1

Supporting such a variety of model requirements will certainly be a challenging task.

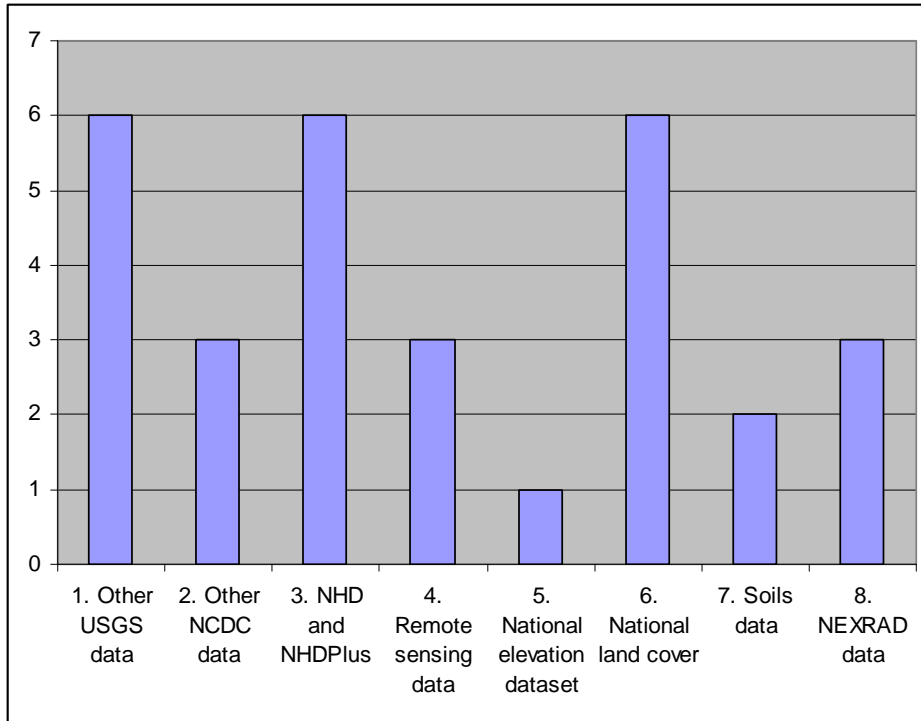
II. Interaction with CUAHSI HIS, and related data needs.

Most of the representatives of the test bed project stated that they are familiar with CUAHSI HIS, with 2 stating that they are “*Very Familiar (I am one of developers or researchers in the project)*” and 6 selecting “*Familiar (I have tried CUAHSI HIS website, experimented with web services, talked to developers)*”. Three test bed projects marked “*Somewhat familiar (heard about it)*” as the answer.

With respect to additional data sources that CUAHSI HIS may provide access to in the coming year, test bed preferences are ranked as follows:

Other USGS data (beyond NWIS already provided via CUAHSI HIS)	6
NHD and NHDPlus	6
National land cover	6
Other NCDC data	3
Remote sensing data	3
NEXRAD data	3
Soils data	2
National elevation dataset	1

Consistency of this result with earlier surveys suggests that interest in NHDPlus and NHD, as well as the National land cover and other USGS datasets, goes beyond specific test bed requirements and reflects general community needs.



Responses from two test beds included additional pointers to data sets, namely “MODIS snow cover product; SNOTEL Data; NRCS Snow Course Data; Paleoclimatology data (sediment, tree rings, geochemical, etc. proxies); Landuse, reservoirs, irrigation and land use change; water wells and habitations and change”, and “A variety of state & local data, including CDEC”

There is also considerable diversity across projects on how these data will be disseminated. The following table lists responses to question "*What are the ways you would like to share test bed data?*" which allowed one or two selections.

Simple web pages with file download	7
Programmatic access to data (e.g. web services)	6
CUAHSI HIS Workgroup server	5
Publish, subscribe mechanisms, for streaming	1
Other - digital library	1

These results show that the majority of test beds would benefit from technologies being developed within CUAHSI HIS, specifically the web services and the test-bed level (workgroup) observation data servers. Alternately, it appears that test beds would prefer to avoid the dissemination hurdle by simply publishing the data on the Web as downloadable files. For some of the research groups this option, or a digital library, are preferred because the data they typically collect are different than common point observation network data which CUAHSI HIS has focused on to date.

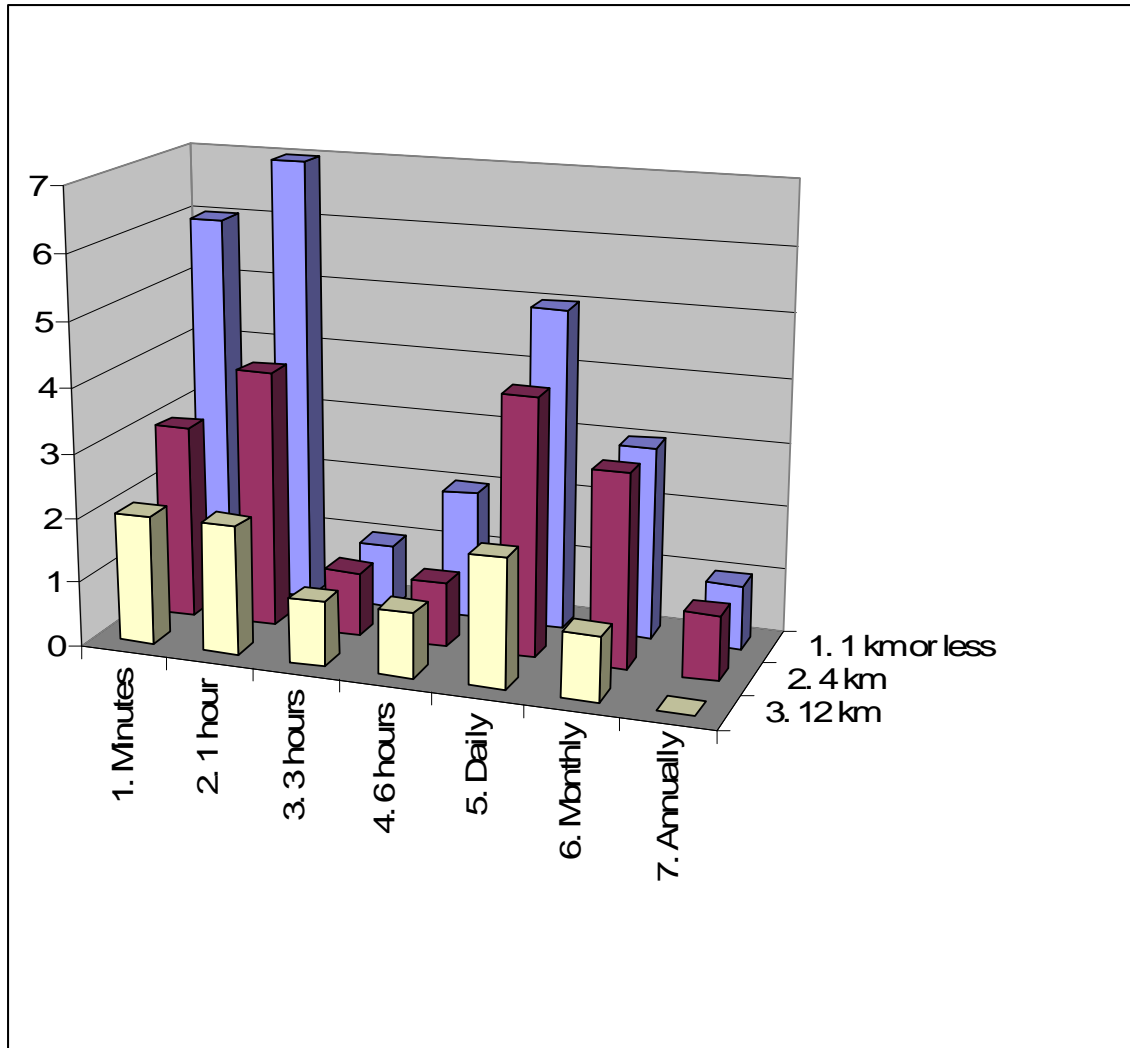
III. Use of atmospheric science data.

Several questions were contributed by NCAR, in light of collaboration between CUAHSI HIS and NCAR on supporting “CUAHSI-style” access to atmospheric data for hydrologists. The questions focused on meteorological variables and spatial and temporal scales that test bed groups are interested in, and allowed for multiple selections. On question "What atmospheric variables are you interested in accessing from CUAHSI HIS?" the responses distributed as follows.

Total Precipitation	9
Precipitation rate	9
Temperature	9
Evaporation	9
Humidity	6
Soil moisture	7
Other: Snow	1
Other: Net radiation	1
Other: Solar radiation at several frequencies, cloud cover, dust, wind speed, barometric pressure	1

The table shows that most test bed projects need access to the several common meteorological variables (at the exception of the UC Merced project, which indicated that NCAR data are not a priority at this time). The most requested spatial resolutions are 1km and finer (9 responses), with temporal resolution of 1 hour (8 responses) or less (7 responses), and a secondary peak at daily values (6 responses). The joint distribution on the two variables is shown below.

In response to questions about test bed interest in meteorological observation data versus output from weather forecast models, the majority indicated they need both (9 out of 11 need observation data; 8 out of 11 are interested in model output)



IV. General CI development preferences.

Respondents were asked to express their degree of agreement with several statements related to general features of CI for hydrologic sciences. The results are summarized in the following table:

	Agree or strongly agree	Neutral	Disagree or strongly disagree
1. A dataset collector/contributor should decide the scope of sharing the data	8	2	1
2. HIS shall let you assemble data sources for a watershed, and convert them to a database	8	2	-
3. HIS shall let you formally describe a digital watershed as an integrated view over distributed	8	3	-

resources			
4. HIS shall provide access to high-performance computing	7	3	1
5. HIS shall maintain data provenance information	7	3	-
6. HIS shall let users query using a common vocabulary (ontology)	9	2	-
7. HIS shall be able to export the discovered data into a format of your choice	8	2	-
8. Metadata is important and you are willing to invest the time necessary to annotate your data	10	1	-

Interpreting these results, we note strong agreement across test beds on the needs for adequate metadata description and convenient data handling. This is consistent with the data and metadata focus of the HIS effort.

The third question in the table was accompanied by a request to define a digital watershed. Several insightful definitions were given:

A digital watershed is composed of any aspect of a watershed that is represented electronically or modeled by a computer. This can include both biophysical and socioeconomic data sets or model output.

The CI component that enables the fusion of data/observations with numerical simulations, the dissemination of the data/information/knowledge, and collaborative work

An electronic representation of the spatial characteristics of a watershed and time series data relating to hydrologic information. Data includes elevation, water features, land use, land cover, point observation data and gridded data (remote sensing, processed climate products etc) and is all put into a system where the data is related and usable for investigative purposes.

Three possibilities: 1) all the raw data (every number that has been collected; 2) processed data (gridded, time-sampled, etc.); 3) simulations of the watershed.

The data necessary to provide the topography, geology, soils, vegetation, land-use, forcing, and observed states for a delineated watershed

Relevant measurement & model data, including characterization

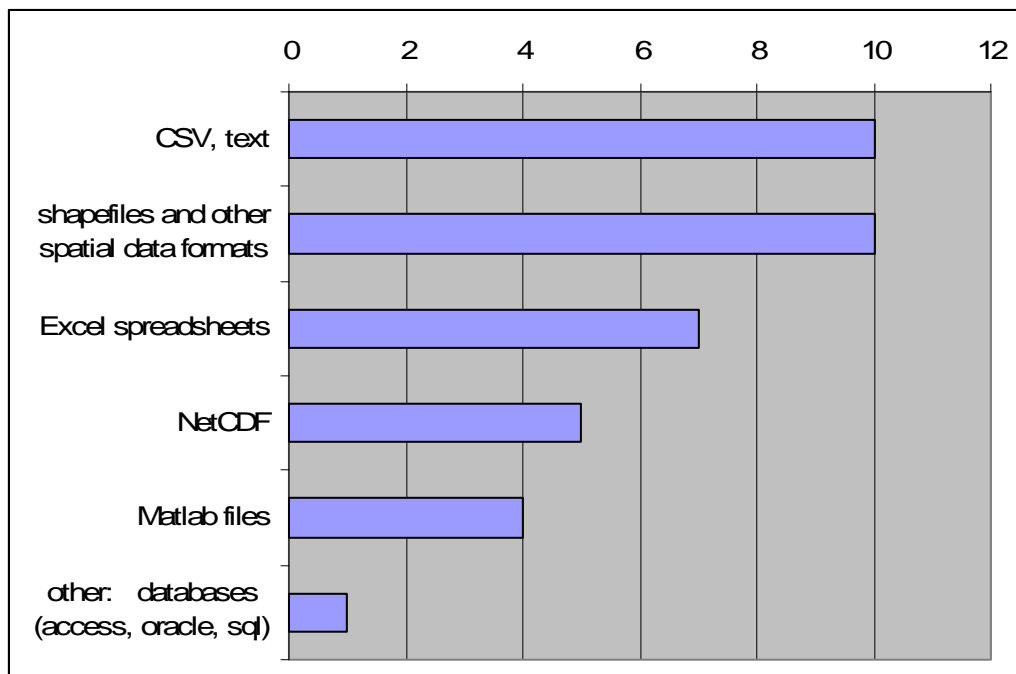
Data integrator/generator

While there are different perspectives on digital watershed, several main themes appear common. These include understanding digital watershed as an integration of different

data resources, as a comprehensive digital repository of both data and model outputs, and as a database supporting watershed-level modeling.

The seventh question in the table was continued by a request to list data formats in common use. The ranked responses are as follows:

CSV, text	10
shapefiles and other spatial data formats	10
Excel spreadsheets	7
NetCDF	5
Matlab files	4
other: databases (access, oracle, sql)	1



Among other large CI projects or data centers that HIS shall connect with, the following were mentioned: NEON and LTER (twice each), and singularly: GEON, CLEANER, SEEK, ORION, EPA NEIEN effort, and the National snow and ice data center, with MODIS snow cover product.

V. Interaction between CUAHSI HIS and test bed projects.

In this section, test bed representatives were asked to address CUAHSI HIS software tools and related support arrangements, and assess test bed needs in terms of number of supported users and datasets.

Question: “Software tools necessary to make deployment of the HIS user friendly”. Respondents could select several entries, and also add tool descriptions. The most universally needed tool is a data loader (marked as necessary by all test beds), followed by metadata and QA-QC tools.

Data loader	11
Metadata tool	9
QA-QC scripts for incoming data	9
Data fusion tools	8
Data viewing and analysis tools	7
Workflow tool to stream data from source to ODM	5

Question: “Your expectations towards HIS support.” Various technical support and software delivery options have been considered. Email support is a favorite, however, some form of face-to-face interaction (whether training events or regular committee meetings or workshops) is also needed: at least one of face-to-face support options were selected by 7 test beds out of 11.

Email support	10
Monthly update downloads	6
Info managers training events	4
Regular committee (info managers and HIS team) meetings	3
Support hotline (phone)	3
Info managers workshops every 3 months	3
Other: Maybe a wiki	1
Other: conf calls; other options as needed	1

Several questions addressed potential contributions of test bed projects to the cyberinfrastructure being developed by CUAHSI HIS, in terms of both data and services. 9 out of 11 test beds indicated that they have observations datasets to share with the larger community and are willing to do this, while 2 test beds qualified this intention by saying that they would be interested in sharing the data at a later time, and if there is interest in data for the particular watershed.

The attitude to software sharing (“*We have software modules to contribute to CUAHSI HIS*”) is varied across test beds. 6 of the 11 testbeds mentioned that they don’t have software modules to contribute at this time. The other half of the projects answered positively, in particular mentioning real-time data integrators, Matlab codes, simulation and model evaluation code. Difficulties of sharing software (the need for documentation, comments, support – and the required time investment) are clearly understood as a deterrent.

The willingness to develop and maintain web services is stated more clearly (6 out of 11 answered “yes”), though lack of specific funding and absence of web service developers on test bed teams are usually given as reasons for not doing this under the current grant.

9 out of 11 test beds responded to the question about the volume of data expected to be managed and served within the test bed project. The range of estimates is quite dramatic reflecting different research foci of the projects, while many responses indicated that the estimates are preliminary (or too early to make). Two of the testbeds indicated the need to store over 1Tb of data, while others seem to require storage in tens or hundreds of Gigabytes.

Similarly, the projects are uncertain as to the number of users that are expected to access test bed servers. For 6 test bed projects that gave a numeric response to this question, the results vary from below 10 (2 projects), to low tens (2 projects), to up to 200.

VI. Finally, words of wisdom.

There were several open ended questions in the survey, asking to formulate research focus of the test bed projects, and describe grand challenges faced by HIS infrastructure. The responses indicate great diversity among the projects, and in how supporting infrastructure is envisioned.

The foci of the test bed projects are:

Quantifying flowpaths, fluxes and stores of groundwater in the urban environment (UMBC)

Investigation of water cycle processes... Integrating wired and wireless sensors, low power wireless communication, embedded microcontrollers, commodity cellular networks, internet, unattended QA, metadata, relational DB, interfaces to models, etc. (Univ of Iowa)

Using CI to add the study of hypoxia (Corpus Christi Bay)

This project develops new engineering approaches to address science questions about loading of nitrate to watersheds: how geologic setting, land use and terrain influence loading; the relationship between flow mixtures and nitrate loads; and the feasibility of using surrogate measures for estimating basin scale nitrate dynamics in space and time. The engineering approaches are two-pronged. (1) Commercially available nitrate, conductivity, temperature, and pressure sensors will be installed in an array that will sample at sub-daily frequencies and communicate via cell phone technology in real time to a central computer. (2) Existing data will be used to develop a statistical algorithm

based on information theory to identify minimum spatial and temporal spacing of sampling necessary to reduce uncertainty to a prescribed level. This algorithm will be validated with the newly collected data (University of Florida)

Fluxes and stores of nutrients from lower Susquehanna River into upper portion of the Chesapeake Bay. Tasks, deployment of WorkGroup HIS, population of ODM using HIS tools, further development of HIS tools and feedback, deployment of SPARROW and predictions of flux data, data gap analysis using SPARROW, web-services to link SPARROW into ODM for direct feed of data. (Drexel, JHU, PennState)

We will address the following questions related to both research design and hydrologic observatory design:

- 1) How can we overcome multiple data formats and sources to build near- and long-term data sets for hydrologic analyses of snow melt runoff?*
- 2) What is the distribution of snowmelt runoff timing and how has it changed historically and over the past few centuries and millennia?*
- 3) How has human disturbance of watersheds affected timing and amount of runoff compared to climatic controls?*
- 4) How do glaciers influence runoff through time?*
- 5) What are the first-order processes controlling Spring snow melt amount and timing?*
- 6) How can remote sensing and ground data be combined to effectively model snow melt hydrographs?*

These questions will be explored through a series of parallel research efforts that will combine numerical simulation with analysis of existing ground and remotely sensed data for climate, snow, land cover, and river runoff. Modeling will elucidate first principals governing the time and space release of snow melt water from large basins. This research

will cross a range of scales and will consider time frames from paleo-megadroughts (1-7 kyr ago) through the instrumental era to better understand snow runoff in large watersheds of the Northern Rockies. The long-term goal is to build a complete digital watershed with climatologic and hydrologic data meshed with water and land use data in the basinwide physiographic and ecologic context (University of Montana)

We are enhancing physical and biological sensing/characterization capabilities, and to develop a real-time web-based informational and data management system. Specifically, we are requesting an AOA and Automated Vertical Profiling (AVP) module to improve Ferry-based rapid detection and quantification of harmful bloom-forming phytoplankton taxa and characterization of physical-chemical conditions promoting and controlling bloom conditions in the Pamlico Sound System, NC. These advancements will be adaptable to other ferry- and ship-based unattended water quality and environmental sensing/assessment programs for estuarine, coastal and large lake ecosystems (NC Chapel Hill)

Blending predictive modeling, geodatabase, and experimentation to advance the open source Penn State Integrated Modeling framework and advancing our understanding of process coupling over 3 scales (hillslope, watershed, basin). (Penn State)

Establish a “virtual” hydrologic observatory, and provide direction for building new infrastructure in an actual observatory (UC Merced)

The overall objective of our research is to establish a wireless network with embedded sensing capable of monitoring fundamental water quality parameters. The ability of these fundamental water quality parameters to be used for predicting the presence of emerging chemical contaminants in urban streams will also be determined. It is hypothesized that the water quality in streams draining similar impervious urban areas is controlled by the mean and variance of effective stormwater residence time (University of Minnesota)

The overarching purpose of this research is to improve understanding of the processes involved in transport of waterborne nutrients to streams and delivery of nutrients to water bodies related to land use and land management practices through examining the interaction between low and high frequency observations. The goal is to improve upon the capability for quantifying sediment and nutrient loading by combining high frequency measurements of surrogate variables (e.g. Turbidity) with less frequent wet chemistry samples that quantify the constituent of interest (e.g. Phosphorous) using a seamless sensor to archive hydrologic information system and Bayesian Network to trigger the collection of wet chemistry samples based on hydrologic conditions in the watershed. This overarching goal is organized into the following specific objectives and hypotheses (Utah State University)

Finally, three of the questionnaires contained responses to the question about grand challenge research questions:

Natural variability of hydrologic systems compared to long-term human forcings (e.g., climate, landuse, etc.). (University of Montana)

How are humans and climate change impacting the environmental systems within river basins? (Penn State)

Again it's a little early to say but our biggest challenges are related to how to convert the swarms of sensor signals coming in into consolidated data that can be used to test hypotheses and populate models. For those involved in more 'management-related' research an important is data archiving for future used, sort of like the fossils at a paleontology museum, where researchers keep returning to learn new things from data (Utah State)

VII. Conclusions

- There is a great variety in information management needs among test bed projects: in terms of data sources and formats used, models, expected number of users and volumes of data to be managed, etc. It would be hard to create a system answering all their needs; instead it would make sense to focus on cross-test bed needs and create a flexible and configurable system that can be further customized for each project.
- Data management, data discovery and dissemination issues is where most test beds need assistance from CUAHSI HIS (specifically, data reduction, resolving format and semantic inconsistencies, and metadata)
- NHDPlus and NHD, National Land Cover Dataset, and other USGS datasets are the additional top datasets of interest for the test beds (beyond those already provided by CUAHSI HIS)
- While generally consistent with earlier CUAHSI and WATERS surveys, the survey results provide additional details about conceptual frameworks, models used, and expectations of each test bed project.
- The test beds are willing to share observation data they collect, but will require technical guidance and support. Technical communication via email and web site are preferred, as well as periodic face-to-face meetings with the HIS team.
- Test bed teams recognize the need to interoperate with other earth sciences observations projects, including NEON, LTER, GEON, CLEANER, SEEK, ORION, etc. In particular, there is a strong need for meteorological fields, both observation data and model outputs.